

Computer-guided ureteroscopy and laser lithotripsy



Soumya Gupta
St. Hugh's College
University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2023

Acknowledgements

First of all, I would like to express my sincere appreciation and gratitude to my primary supervisor Prof. Jens Rittscher for his invaluable support, constant guidance and encouragement throughout my PhD. I am also extremely grateful to my co-supervisor Dr Ben Turney for his invaluable input from the clinical side that helped me discover the clinical challenges faced in the field of Ureteroscopy. I am also thankful to Dr Ben for providing me with the opportunity to witness some clinical ureteroscopy procedures and acquire clinical data for my project. I would also like to sincerely acknowledge my co-supervisor Dr Sharib Ali for his invaluable suggestions, immense support and guidance during my PhD. In addition, Dr Sharib's publications helped me gain a good understanding of deep learning and computer vision algorithms for medical datasets.

I am also extremely grateful to Dr Louise Goldsmith, who has been extremely helpful in finding out the time to help me perform experiments with the Lithovue and acquire *in vitro* data that really helped me gain insights into ureteroscopy and laser lithotripsy. I thank profusely all the members of Prof Rittscher's group, for their stimulating discussions, support and post-lunch fun conversations. I would also like to acknowledge that the computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC.

I highly appreciate the financial support from Boston Scientific for my DPhil project. I would especially like to thank my industrial supervisors Dr Aditi Ray and Niraj Rauniyar for providing me with the opportunity to pursue two internships with Boston Scientific and the opportunity to meet the Boston Scientific team based in Marlborough, USA. These opportunities not only helped me understand the cutting-edge work being pursued at BSC but also helped me organise my project goals. I would also like to thank Longquan Chen and Rob Riker from Boston Scientific for their valuable suggestions and stimulating discussions during the meetings.

I would like to express my utmost gratitude to my family for their immense love, support, and sacrifices during my PhD journey. I would also like to thank my friends for their companionship and encouragement that helped me in completing this project. The COVID-19 pandemic was a difficult time and I am immensely grateful to God and my loved ones for giving me the strength to overcome it and keep going. Thank you very much to everyone I have met along the way. Hope this thesis is just the start of my long research journey.

Abstract

Ureteroscopy with laser lithotripsy has evolved into a minimally invasive routine treatment for stones in the urinary tract. Automated segmentation can assist the urologists during the procedure and is also the primary step towards the estimation of stone size, one of the most important parameters to decide if the stone requires further fragmentation. Continuous irrigation is maintained during the procedure for optimal visualisation and to reduce the excess heat generated from the laser. The irrigation flow is increased when the visualisation is impaired by stone debris. However, this increases the intrarenal pressure which can potentially lead to several complications. This highlights the need for computer vision algorithms that can virtually flush out the debris whilst keeping the irrigation flow low. The ureteroscopy and laser lithotripsy datasets suffer from challenges such as small field of view, occlusions from stone debris, intra-operative bleeding, motion blur, specular highlights, dynamic background, varying illumination conditions, stone heterogeneity and artefacts from the turbid fluid. In addition, there is both inter-patient and intra-patient variability in terms of tissue appearance, illumination conditions, stone variability, and different viewpoints with respect to the camera. This makes it even more challenging to develop generalisable machine-learning models for ureteroscopy and laser lithotripsy. To this date, there are no comprehensive computer vision studies that explore or solve the real-world challenges of ureteroscopy and laser lithotripsy.

In this thesis, novel datasets have been acquired and annotated to facilitate the development of computer vision algorithms for ureteroscopy and laser lithotripsy. This thesis presents novel algorithms that leverage motion information between frames for improved segmentation of stone fragments and laser fibre. This thesis also presents frameworks for unsupervised domain adaptation and improved generalisation of segmentation models for various endoscopy datasets. Further, this thesis investigates turbidity mitigation in lithotripsy datasets for an improved understanding of the challenging problem and proposes a framework to effectively enhance image quality.

The original contributions are as follows: (1) Classical approach to stone segmentation in ureteroscopy and laser lithotripsy datasets. (2) Deep learning-based segmentation of stone and laser fibre in ureteroscopy and laser lithotripsy datasets. (3) Improving domain adaptation and generalisation of segmentation models using Variational Autoencoders (VAEs) in various endoscopy datasets. (4) Turbidity mitigation for improved visualisation in laser lithotripsy. (5) New synthetic, animal and clinical ureteroscopic laser lithotripsy datasets acquired and annotated to facilitate the development of computer vision algorithms.

Contents

List of Figures	viii
List of Tables	xiii
List of Abbreviations	xvi
List of Symbols	xviii
1 Introduction	1
1.1 Motivation	1
1.2 Datasets	5
1.2.1 <i>In vitro</i> dataset	5
1.2.2 <i>In vivo</i> dataset	6
1.3 Thesis structure and contributions	7
1.4 Publications	9
2 Background	11
2.1 Clinical background	12
2.1.1 Urolithiasis	12
2.1.2 Ureterscopy	14
2.1.3 Laser lithotripsy	15
2.2 Technical background	16
2.2.1 Image segmentation	16
2.2.1.1 Segmentation of abnormalities in endoscopy	21
2.2.1.2 Segmentation of kidney stones	22
2.2.1.3 Optical flow-based approaches	23
2.2.1.4 Encoder-decoder based approaches	27
2.2.1.4.1 Residual and recurrent connections based networks	28
2.2.1.4.2 Dilated convolutions and ASPP-based net- works	29
2.2.1.4.3 Attention mechanism-based networks	30
2.2.1.5 Auxiliary tasks to improve segmentation	32

2.2.1.5.1	Motion-based image registration	32
2.2.1.5.2	VAE-based image reconstruction	35
2.2.1.6	Metrics for evaluating image segmentation	37
2.2.2	Image denoising and quality enhancement	38
2.2.2.1	Traditional methods	38
2.2.2.2	Learning-based methods	39
3	Classical computer vision-based image segmentation	44
3.1	Introduction	45
3.2	Method	48
3.2.1	Self-similarity descriptors	48
3.2.1.1	Modality independent neighbourhood descriptor . . .	48
3.2.1.2	Normalised correlation transform descriptor	49
3.2.2	Data-term	50
3.2.3	Regularisation	51
3.2.4	Energy minimisation	51
3.2.5	Minimum variance quantisation	53
3.3	Results	54
3.4	Discussion	57
3.5	Conclusion	59
4	Motion-based semantic segmentation using deep learning	60
4.1	Introduction	62
4.2	Materials and Method	64
4.2.1	Materials	64
4.2.2	Method	67
4.2.2.1	HybResUNet	68
4.2.2.2	DVFNet	69
4.2.2.3	Loss function	69
4.2.2.4	Proposed framework	72
4.3	Results	72
4.3.1	Experimental setup	74
4.3.1.1	Network implementation	74
4.3.1.2	Independent runs for <i>in vitro</i> and <i>in vivo</i>	74
4.3.1.3	Single vs sequence samples in the test set	75
4.3.2	Data augmentation	76
4.3.3	Loss function	77
4.3.4	Quantitative results	78
4.3.5	Qualitative results	80
4.3.6	Extended out-of-sample test	81

4.3.7	Inference time analysis	83
4.3.8	Ablation study of network design	86
4.4	Discussion	86
4.5	Conclusion	92
5	Use of VAEs beyond simple image generation	93
5.1	Introduction	95
5.2	Improved generalisation using VAE-embodied learning (UNet-eVAE) .	97
5.2.1	Method	97
5.2.1.1	UNet module with VAE	97
5.2.1.2	Joint loss minimisation	99
5.2.2	Experiments and results	100
5.2.2.1	Implementation Details and setup.	100
5.2.2.2	Datasets.	100
5.2.2.3	Quantitative results	101
5.2.2.4	Qualitative results	103
5.2.3	Discussion	104
5.3	Unsupervised domain adaptation in endoscopy (EndoUDA)	106
5.3.1	Method	108
5.3.1.1	VAE training on source domain	108
5.3.1.2	Segmentation module	109
5.3.1.3	VAE latent search optimisation via new joint Loss . .	109
5.3.2	Experiments and results	110
5.3.2.1	Implementation Details	110
5.3.2.2	Datasets	111
5.3.2.3	Benchmark comparison and ablation study	111
5.3.2.4	Comparison with SOTA method	112
5.3.2.5	Effect of out-of-distribution data in supervised training	113
5.4	Conclusion	113
6	Turbidity mitigation for improved visualisation	115
6.1	Introduction	116
6.2	Implementation of state-of-the-art approaches for image denoising . . .	119
6.2.1	Averaging	120
6.2.2	Traditional dehazing approaches	120
6.2.3	Noise2Noise	122
6.2.4	Deep image prior	122
6.2.5	CycleGAN	124
6.3	Method	124
6.3.1	Network modules	125

6.3.2	Proposed framework	126
6.4	Experiments and Results	127
6.4.1	Datasets	128
6.4.2	Network implementation	128
6.4.3	Evaluation metrics	129
6.4.4	Quantitative results	129
6.4.5	Qualitative results	131
6.4.6	Quality score prediction using a neural network	132
6.5	Discussion	134
6.6	Conclusion	137
7	Conclusion and Future work	138
7.1	Summary of contributions	138
7.2	Directions for future work	140
7.3	Conclusion	141
	References	143

List of Figures

1.1	Stones formed in various parts of the urinary tract. Figure courtesy of [10]	2
1.2	Ureteroscopy and laser lithotripsy frames from animal studies and clinical procedures illustrating the challenges of this dataset such as scene variability, turbidity, and occlusions from stone debris and blood. Some occlusions due to debris and blood have been highlighted using yellow and red rectangles, respectively. The region highlighted using a blue rectangle shows turbid fluid obscuring the field of view.	3
1.3	Project overview showing how the research presented in this thesis fits into the feedback loop system of the ureteroscopy procedure	7
2.1	Illustration of various types of kidney stones. Figure courtesy of [16] . .	13
2.2	Insertion of ureteroscope for urinary tract examination. Figure courtesy of [19]	14
2.3	Lithovue TM and its workstation. Figure courtesy of [20]	14
2.4	Illustration of laser lithotripsy and basketing procedures during ureteroscopy. Figure courtesy of [24]	15
2.5	Illustration of semantic and instance segmentation. Figure courtesy of [28]	17
2.6	Hard and soft computing image segmentation techniques	18
2.7	Examples of various labelled classes in the HyperKvasir dataset illustrating the wide range of region-of-interests and abnormalities in endoscopy. Figure courtesy of [56]	20
2.8	Illustration of the boundary of kidney stones in CT [69, 70], Ultrasound [71] and Ureteroscopy images	21
2.9	Plot of optical flow equation illustrating the normal and parallel components of optical flow	23
2.10	Illustration of the UNet architecture. Figure courtesy of [40]	28
2.11	Illustration of different variants of the convolutional, residual and recurrent units	29
2.12	Illustration of Atrous spatial pyramid pooling that consists of parallel dilated convolutions with different rates	30
2.13	Schematic of the attention gate (AG) mechanism. Figure courtesy of [115]	32

2.14	Illustration of an autoencoder and variational autoencoder model architecture	35
2.15	Illustration of Deep Image Prior results for diverse image reconstruction problems. Figure courtesy of [187]	39
2.16	Illustration of denoising of a Monte Carlo rendered image using N2N a) input, b) trained using noisy targets, c) trained using clean targets, d) reference image. Figure courtesy of [188]	40
2.17	Illustration of the GAN architecture consisting of a generator and discriminator that are trained synchronously. Figure courtesy of [189] . . .	40
2.18	VAE-based framework that uses a pretrained segmentation network $\mathbf{S}$ to partition the image into two subspaces for object-wise disentanglement. Figure courtesy of [198]	42
3.1	a) $\mathbf{x}_i$ represented by black dots ($i \in [1,8]$) are the 8 connected neighbours of the centre pixel $\mathbf{x}$ (represented by red dot) in window W . b) A patch P_i centred at third element $\mathbf{x}_3$ of window W with green dots representing the neighbouring pixels $\mathbf{p}_{\mathbf{x}_3}^i$. c) Illustration of translation between corresponding pixels $\mathbf{x}_i$ and $\mathbf{p}_{\mathbf{x}_3}^i$ ($i \in [1, 8]$) of window W and patch P_3 respectively.	49
3.2	Optical flow images and the corresponding results after application of minimum variance quantisation. The saturation and hue in the optical flow images represent the magnitude of displacement and orientation of the fragment respectively. The colour encoding is shown on the right with the horizontal and vertical components of displacement represented along the x-axis and y-axis respectively.	55
3.3	Qualitative evaluation of optical flow estimation (before minimum variance quantisation) with NCoT and MIND descriptors using both two-frame and three-frame strategy. The saturation and hue in the optical flow images represent the magnitude of displacement and orientation of the fragment respectively. The colour encoding is shown on the right with the horizontal and vertical components of displacement represented along the x-axis and y-axis respectively.	56
3.4	Optical flow results (after minimum variance quantisation) obtained using different combinations of values for parameters such as warps per level, iterations per warp, antialiasing start level, and pyramid factor	58
4.1	Example images for <i>in vitro</i> , <i>in vivo</i> (animal) and <i>in vivo</i> (human) clip sequences with corresponding ground truth masks showing stone and laser fibre. Stone debris and blood are highlighted in yellow and red rectangles (on the left), respectively.	63

4.2	Box-plots showing the size distribution of stone and laser class across training, validation, and test sets in <i>in vitro</i> and <i>in vivo</i> datasets. It can be seen that the size distribution of kidney stones is different across train, validation, and test sets, indicating that they come from different videos. The large variability between <i>in vitro</i> and <i>in vivo</i> datasets is also clearly evident, explaining why experiments were run independently for the <i>in vitro</i> and <i>in vivo</i> datasets.	65
4.3	Pie charts showing the division of <i>in vitro</i> and <i>in vivo</i> datasets. The proportion of both human and animal samples has also been illustrated for the <i>in vivo</i> dataset	65
4.4	Proposed framework: a) Overall proposed semantic segmentation framework illustrating both training and inference paths presented in this study. Here, DVFNet utilising image pairs is only used for training shown by solid path while the learned weights of HybResUNet are used during frame-wise inference indicated by the dotted line. b) Network representation of the HybResUNet used in this work. It is a 9-level deep UNet architecture with residual connections in the encoder path and c) an encoder-decoder DVFNet architecture used to compute deformation vector fields (DVF) between image pairs that are used to prune the segmentation results obtained from the HybResUNet.	73
4.5	Qualitative comparison of independent runs against joint training of <i>in vitro</i> and <i>in vivo</i> datasets	75
4.6	t-SNE plot of features extracted from <i>in vitro</i> (shown in blue) and <i>in vivo</i> (shown in red) samples. The clear separation in their distribution indicates the variability between the <i>in vitro</i> and <i>in vivo</i> datasets	75
4.7	Qualitative analysis of the proposed method (Att-HybResUNet+DVFNet (with DVF) for the <i>in vitro</i> and HybResUNet+DVFNet (with warped image) for the <i>in vivo</i>) against existing SOTA methods on the ureteroscopy and laser lithotripsy test sets (Test-I). Each row shows a test image, followed by its ground truth segmentation mask (showing laser fibre and stone), followed by SOTA approaches: UNet, HybResUNet, DeepResUnet, R2-UNet, and DeepLabv3+, and finally the proposed model which is Att-HybResUNet+DVFNet (with DVF) for the <i>in vitro</i> and HybResUNet+DVFNet (with warped image) for the <i>in vivo</i>	82
4.8	Random frames from different patients illustrating the intra and inter-patient variability across data. Here, patient 3 represents the out-of-sample test dataset used for generalisability assessment.	84
4.9	Qualitative comparison of proposed network architecture for <i>in vivo</i> data against baseline methods on 18 samples from an unseen patient data (example images shown in row 3 of Figure 4.8)	84

4.10	Qualitative analysis of the proposed method (HybResUNet+DVFNet (with warped image)) for <i>in vivo</i> against existing SOTA methods on the unseen <i>in vivo</i> test dataset (Test-II).	85
5.1	Proposed UNet2VAE framework: Here, four connections from the VAE-decoder are fused with the segmentation-decoder. Since the initial decoder layers have small spatial details ($1^2, 4^2, 8^2$), latter layers are chosen ($16^2, 32^2, 64^2, 128^2$).	98
5.2	Qualitative evaluation of segmentation results obtained from UNet and UNet2VAE trained on Kvasir-SEG dataset	103
5.3	Qualitative evaluation of segmentation results obtained from UNet and UNet2VAE trained on <i>in vitro</i> and <i>in vivo</i> ureteroscopic laser lithotripsy datasets.	104
5.4	Example images and masks of the two GI datasets for which source and target domains are separated based on their modality.	107
5.5	Complete architecture of the proposed EndoUDA: a target-independent unsupervised domain adaptation wherein VAE module allows learning the source domain representations while the UNet module leverages the learned encoded features for semantic segmentation. [234]	107
5.6	Qualitative study: Baseline UDA experiments on BE and polyp images. [234]	112
6.1	Example images for <i>in vitro</i> and <i>in vivo</i> (animal) ureteroscopic laser lithotripsy images showing turbidity and other artefacts such as blood, saturation, and specularities	117
6.2	Results for temporal averaging of pixel intensities (across five adjacent frames) shown for three <i>in vitro</i> and three animal dataset samples . . .	120
6.3	Dehazing results using approach (a) presented in [242] and approach (b) presented in [243] implemented on two natural scene images followed by three <i>in vitro</i> and three animal dataset ureteroscopy images	121
6.4	Qualitative results of Noise2Noise(N2N) implementation on three <i>in vitro</i> and three animal dataset ureteroscopy images	123
6.5	Image restoration and denoising using Deep Image Prior on three <i>in vitro</i> and three animal dataset ureteroscopy images showing there is no visible change in the image quality of resulting images	123
6.6	Implementation of CycleGAN on our dataset to convert Domain-A (debris) images to Domain-B (clean) images. Failure cases where stone debris or saturated pixels are getting reconstructed as a stone have been highlighted using red rectangles	124

6.7	Proposed end-to-end framework for enhancing the quality of ureteroscopy images. The framework consists of three sub-networks- UNet, AE and a β -VAE. The predicted foreground and background image obtained with the help of UNet is fed into AE and β -VAE, respectively for explicit disentanglement of the latent space	126
6.8	Qualitative evaluation of five <i>in vitro</i> images (first row) against the resulting images obtained from state-of-the-art β -VAE and our frameworks (shown in subsequent rows). The proposed network can be seen to preserve the foreground structure whilst reducing turbidity (see debris in the regions highlighted using red rectangles (first row) getting reduced using the proposed network (last row)).	130
6.9	Qualitative evaluation of five <i>in vivo</i> images (first row) against the resulting images obtained from state-of-the-art β -VAE and our frameworks (shown in subsequent rows).The proposed network can be seen to preserve the foreground structure whilst reducing turbidity (see debris in the regions highlighted using yellow rectangles (first row) getting reduced using the proposed network (last row)).	131
6.10	Reconstructed images obtained by interpolating the latent space (only shown first five of the ten dimensions) from -3 to 3 for various networks. AE- β VAE-Seg can be seen to preserve foreground information across all latent dimensions with the second latent dimension (highlighted in red) disentangling debris. The right-most image across this second dimension (highlighted in blue) is taken to be the enhanced output image.	132
6.11	Figure illustrating <i>in vitro</i> images classified into <i>good</i> (green), <i>medium</i> (blue), and <i>bad</i> (red) categories	133
6.12	t-SNE plot illustrating the distribution of manually annotated debris scores that were used to train the autoencoder-based classification network on <i>in vitro</i> dataset	133

List of Tables

3.1	A list of algorithm parameters and their corresponding empirically established values	54
3.2	Quantitative evaluation of the motion induced segmentation after minimum variance quantisation	54
4.1	Overview of the train, validation, and test splits for the <i>in vitro</i> and <i>in vivo</i> experiments. Mean (μ) and standard deviation (σ) for each stone and laser annotations in each split are also provided. Here, Test-I refers to the split samples from the curated dataset as 20% split and Test-II refers to the unseen animal samples.	67
4.2	Number of trainable parameters for different CNN networks explored in this work	73
4.3	Quantitative comparison of independent runs against joint training of <i>in vitro</i> and <i>in vivo</i> datasets	75
4.4	Comparison of dice score and computation time (for 10 samples on NVIDIA Quadro RTX 6000) for HybResUNet+DVFNNet applied on sequence data vs single sample data in the test dataset	76
4.5	Augmentation settings used in this study for both <i>in vitro</i> and <i>in vivo</i> datasets	77
4.6	Comparison of augmentation techniques (applied on HybResUNet) for both <i>in vitro</i> and <i>in vivo</i> datasets	77
4.7	Comparison of loss functions showing accuracy improvement with compound loss as opposed to state-of-the-art focal loss	78

4.8	Quantitative comparison of proposed network architectures against existing approaches on the ureteroscopy and laser lithotripsy <i>in vitro</i> test set (Test-I) where listed values represent average performances. Here, 'I' in the table represents the performance of various baseline models, 'II' represents the combination of the best-performing baseline model (HybResUNet) and DVFNet (with DVF) under different configurations, and 'III' represents the combination of the best-performing baseline model (HybResUNet) and DVFNet(with warped image) under different configurations. The SOTA methods have been accordingly referenced and the experimental methods have been indicated by a †superscript. <i>p</i> -values that represent statistical significance between the proposed method and other implementations with <i>p</i> -value < 0.05 are computed using paired t-test (highlighted by a * superscript)	80
4.9	Quantitative comparison of proposed network architectures against existing approaches on the ureteroscopy and laser lithotripsy <i>in vivo</i> test set (Test-I) where listed values represent average performances. Here, 'I' in the table represents the performance of various baseline models, 'II' represents the combination of the best performing baseline model (HybResUNet) and DVFNet (with DVF) under different configurations, and 'III' represents the combination of the best performing baseline model (HybResUNet) and DVFNet(with warped image) under different configurations. The SOTA methods have been accordingly referenced and the experimental methods have been indicated by a †superscript. <i>p</i> -values that represent statistical significance between the proposed method and other implementations with <i>p</i> -value < 0.05 are computed using paired t-test (highlighted by a * superscript)	81
4.10	Quantitative comparison of proposed network architecture for <i>in vivo</i> data against existing approaches on an unseen <i>in vivo</i> test dataset (Test-II) where listed values represent average performances. <i>p</i> -values represent statistical significance between proposed method and other implementations with <i>p</i> -value < 0.05 (highlighted by a * superscript)	83
4.11	Quantitative comparison of proposed network architecture for <i>in vivo</i> data against baseline methods on 18 samples from unseen patient data where listed values represent average performances.	84
4.12	Performance comparison of 4-fold patient-wise cross-fold validation performed on <i>in vivo</i> data where each fold represents validation results on different subjects (Fold1, Fold2 and Fold4 are human patient data while Fold3 represents animal data).	85

4.13	Inference time of base network for 10 images of the test set on NVIDIA Quadro RTX 6000. DVNet is only used during training and not in test inference	86
4.14	Quantitative comparison of proposed network architectures against existing approaches on the ureteroscopy and laser lithotripsy test sets (Test-I). The published networks have been accordingly cited and the experiment networks have been indicated by a †superscript.	87
5.1	Ablation study for hyper-parameter β	99
5.2	Quantitative evaluation of segmentation results obtained from UNet and proposed UNet-eVAE (ours) trained on Kvasir-SEG, <i>in vitro</i> and <i>in vivo</i> ureteroscopic laser lithotripsy datasets	102
5.3	Quantitative comparison of baseline DeepResUNet against DeepResUNet in the proposed configuration(DeepResUNet-eVAE) when trained with 900 samples of Kvasir-SEG	105
5.4	Supervised Segmentation analysis: Benchmarking BE and polyp datasets on standard segmentation architectures on 80-20 train-test split. [234] .	111
5.5	Empirical results of EndoUDA along with SOTA UDA methods with mean μ and standard deviation σ are provided. All comparisons are provided for source-only trained models and tested on target data. [234]	112
5.6	IoU and Dice comparisons for out-of-distribution target samples (NBI) in the source (WLI) GI endoscopy datasets [234]	113
6.1	VIF and RECO metric values (mean $\pm$ std) for quantitative assessment of image quality obtained from state-of-the-art β -VAE and our frameworks on <i>in vitro</i> and <i>in vivo</i> test datasets.	130
6.2	Percentage of images that showed improvement or degradation out of the 88 <i>in vitro</i> images as observed with state-of-the-art β -VAE and our frameworks. Improvement or degradation is calculated based on whether the debris score predicted by the network was decreased or increased, respectively after application of the networks	134

List of Abbreviations

CV	Computer vision
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
DNN	Deep Neural Network
FCN	Fully Convolutional Network
AE	Autoencoder
VAE	Variational Autoencoder
ReLU	Rectified Linear Unit
CLAHE	Contrast Limited Adaptive Histogram Equalisation
ROI	Region-of-interest
SOTA	State-of-the-art
PCNL	Percutaneous Nephrolithotomy
ESWL	Extra Corporeal Shock Wave Lithotripsy
URS	Ureteroscopy
US	Ultrasound
MR	Magnetic Resonance
CT	Computed Tomography
LDOF	Large Displacement Optical Flow
SIFT	Scale Invariant Feature Transform
t-SNE	t-distributed Stochastic Neighbourhood Embedding
DSC	Dice Similarity Coefficient
JI	Jaccard Index
HD	Haussdorf Distance

IoU		Intersection over Union
PPV		Positive Predictive Value
CRF		Conditional Random Field
MRF		Markov Random Field
ASPP		Atrous Spatial Pyramid Pooling
BCA		Brightness Constancy assumption
GCA		Gradient Constancy assumption
HCA		Higher-order Constancy assumption
WLD		Weber Local Descriptor
CRT		Complete Rank Transform
MIND		Modality Independent Neighbourhood Descriptor
NCC		Normalised Cross-Correlation
CoT		Correlation Transform
NBI		Narrow-Band Imaging
WLI		White-Light Imaging
BE		Barrett's oesophagus
UDA		Unsupervised Domain Adaptation
CRC		Colorectal Cancer
GI		Gastrointestinal
BCE		Binary Cross Entropy
KL		Kulback-Leibler
DIR		Deformable Image Registration
DVF		Deformation Vector Fields
N2N		Noise2Noise
DIP		Deep Image Prior
GAN		Generative Adversarial Network

List of Symbols

β	Weight of KL divergence loss
ϵ	Random variable $\sim \mathcal{N}(0, \mathbf{I})$
η	Learning rate
γ	Parameter in Focal loss that adjusts the rate at which easy examples are down-weighted
λ	Relative weighting between data term and regularisation term
ϕ_{reg}	Regularisation term
ψ_{data}	Data term
$\mathbf{u}(u,v)$	Optical flow
$\mathbf{y}$	Predicted segmentation mask
$\mathbf{y}^b$	Boundary map of predicted segmentation map
$\mathbf{y}^{gt}$	Ground truth segmentation mask
$I(x,y,t)$	Image intensity as a function of x,y and time t
R	Radius
$\theta_{\circ}, \theta_{\Delta}$	Sliding window sizes for computation of boundary from segmentation map
$\tilde{\mathbf{X}}$	Reconstructed target image based on nearest point in source domain
$\tilde{\mathbf{X}}_t$	Target domain image
C	Clean(debris-free) image with stones and surgical instruments
K	Linear operator
M	Motion artefacts/blur in an image
n	Connectivity

P	Image patch
p	Probability
S	Similarity measure
T	Turbidity in an image
$w_{\mathbf{x}}^{\mathbf{x}'}$	Weight of neighbourhood pixel $\mathbf{x}'$ centered around pixel $\mathbf{x}$
X_i	Image where i denotes the position in the clip sequence
X_s	Source image
X_t	Target image
X_{warp}	Warped image
γ_1, γ_2	Normalisation factors for the regularisation weights
$\mathbf{y}^{gtb}$	Boundary map of ground truth segmentation map
μ	Mean
σ	Standard deviation
$\mathbf{w}$	Dual variable in closed convex set $\mathbf{S}_w$
$\mathbf{X}'$	Reconstructed image
$\mathbf{X}$	Input image
$\mathbf{x}$	Centre pixel of neighbourhood $\mathbf{W}$
$\mathbf{X}'_{FG}$	Reconstructed foreground image
$\mathbf{X}_{FG}^{gt}$	Ground truth foreground image
$\mathbf{X}_{BG}$	Predicted background image
$\mathbf{X}_{FG}$	Predicted foreground image
$\mathbf{x}'$	Neighbourhood pixels
$\mathbf{Z}$	Latent vector
DVF	Deformation vector field map
$\mathbf{K}^T$	Adjoint linear operator

1

Introduction

Contents

1.1	Motivation	1
1.2	Datasets	5
1.2.1	<i>In vitro</i> dataset	5
1.2.2	<i>In vivo</i> dataset	6
1.3	Thesis structure and contributions	7
1.4	Publications	9

1.1 Motivation

In the past decade, Artificial Intelligence (AI) has witnessed exponential growth due to the advancements in the field of deep learning, high computational power and large open-source datasets. AI-assisted diagnosis and treatment enable automated analysis of medical data that can provide several advantages such as the identification of diseases at early stages, better treatment plans and decreased healthcare costs. AI has created a significant impact in various kinds of endoscopy imaging such as colonoscopy, gastroscopy, capsule endoscopy, etc by enabling automated detection of colon polyps [1–3], detection of oesophageal cancer [4], detection of gastric cancer [5], etc. However, research on AI-based computer vision is at an early stage in the field of Ureteroscopy and laser lithotripsy, a minimally invasive routine procedure for the treatment of kidney stones.

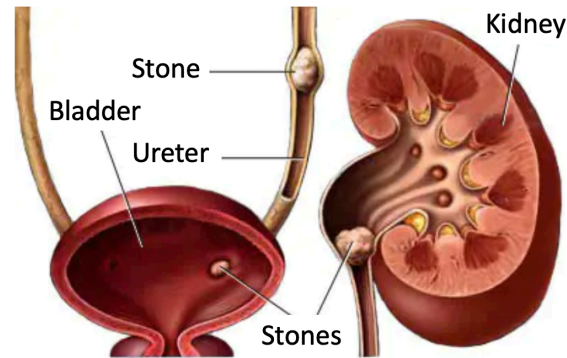


Figure 1.1: Stones formed in various parts of the urinary tract. Figure courtesy of [10]

Kidney stone is a common urological disease faced by a large number of people all over the world. It affects about 12% of the world population at some stage in their lifetime [6]. In addition, it tends to be recurrent in most kidney stone patients with a recurrence rate of 10% after one year, 50% over a period of 5-10 years and 75% over 20 years [7]. Kidney stones also called renal calculi, are formed when crystal-forming substances separate from the urine and accumulate inside the kidney or ureters (as shown in Figure 1.1). Stones typically larger than 5mm may result in a blockage in the ureter and cause severe pain in the abdomen or lower back [8]. Ureteroscopy has evolved into a minimally invasive routine treatment for a number of urological conditions such as urolithiasis (kidney stones), strictures, and hematuria [9].

Technological advancements have led to the development of low-cost single-use endoscopes with improved flexibility and image quality. Boston Scientific introduced the first disposable flexible ureteroscope - the LithovueTM, in 2015 with the ability to access the upper urinary tract. The data used in this study has been acquired using single-use LithoVueTM scope and Lithovue Elite scope (Boston Scientific). The endoscopes are typically connected to an external display during the surgery to help the clinician navigate through the body and access the target site for treatment. The procedure involves the insertion of a long flexible ureteroscope into the urinary tract passing through the bladder and into the ureter and kidney to access the kidney stones. The scope also has a working channel through which tools like laser fibre can be inserted to perform laser lithotripsy, i.e., stone fragmentation using laser energy. Based on the size, location and composition of the stone, the surgeon decides if the stone requires dusting or fragmentation and sets the

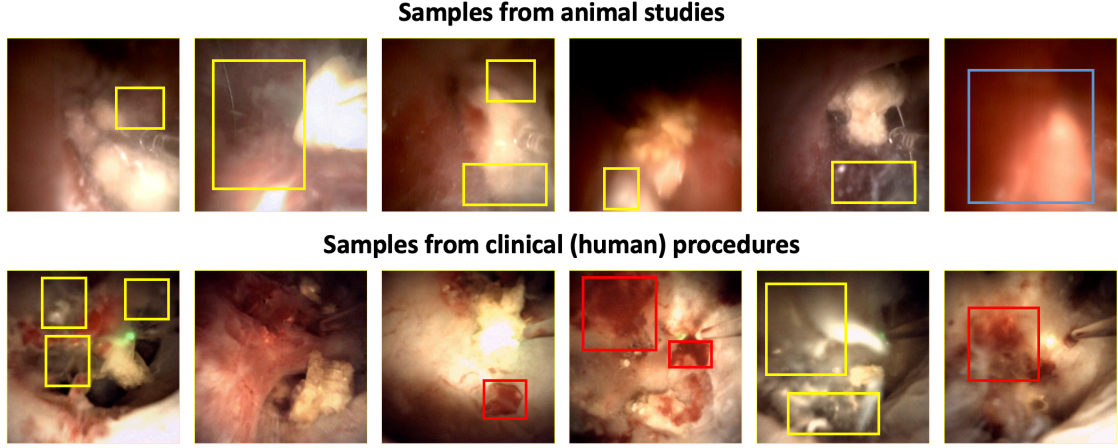


Figure 1.2: Ureteroscopy and laser lithotripsy frames from animal studies and clinical procedures illustrating the challenges of this dataset such as scene variability, turbidity, and occlusions from stone debris and blood. Some occlusions due to debris and blood have been highlighted using yellow and red rectangles, respectively. The region highlighted using a blue rectangle shows turbid fluid obscuring the field of view.

laser settings accordingly [11]. The procedure is normally performed with fluid irrigation to provide optimal visualisation and reduce the extra heat build-up from lithotripsy. Depending on their size, the fragmented stones are either left in place to clear out by themselves over time or extracted using a special wire basket. This thesis addresses some of the major challenges faced by the surgical team during the ureteroscopy procedure by proposing image analysis-based approaches to tackle them. These challenges and their corresponding approaches developed in this thesis are listed below:

- **Precise laser targeting of stones:** Power output from laser fibre results in an undesirable increase in the intrarenal temperatures that may lead to thermal ablation of tissue or result in permanent kidney damage. The surgeon, therefore, tries to carefully target the stone for effective fragmentation and to limit the excess heat generated in the confined spaces of the cavity [11]. A continuous flow of irrigation fluid is also maintained in the cavity to reduce this heat build-up.

Proposed approach: Automated segmentation of stone fragments and laser fibre can help the surgeon in accurate laser targeting, thereby reducing the operative time and excess heat generation from laser lithotripsy. This thesis presents an end-to-end deep learning-based framework that leverages temporal motion information to provide accurate segmentation of stones and laser fibre.

- **Real-time estimation of stone size:** The surgeon is required to make a real-time estimation of stone size to decide if the stone needs to be fragmented further or if the procedure can be terminated. The procedure often involves the insertion of a stent in the ureter for easy passage of the residual stone debris and fragments. The surgeon then either leaves the fragmented stone debris in place to naturally flush out in urine or uses a tiny wire basket to retrieve the residual fragments. However, stone retrieval fails if the stone is larger than the diameter of the ureter and a re-surgery is required.

Proposed approach: Automated segmentation is also the primary step towards the estimation of stone size which is the most important parameter to decide if the fragmentation procedure can be terminated. This will also reduce the chances of any residual fragments and ensure high stone-free rates. This thesis presents a classical approach and a deep learning-based approach that leverages motion information between frames to provide automated segmentation of stone fragments.

- **Optimal vision:** Vision is the key to performing stone fragmentation in an efficient and effective manner. However, the ureteroscopy video is affected by stone debris originating from stone fragmentation that obscures the vision in the kidney. Intra-operative bleeding can also occur during ureteroscopy due to the continuous application of laser energy, intra-renal pressure and the trauma caused to the walls of the ureter [12]. The visualisation is often improved by increasing the irrigation flow. However, this increases intrarenal pressure and can potentially lead to backflow or septic complications.

Proposed approach: To help avoid septic complications, a robust algorithm that can mitigate turbidity and provide optimal vision whilst keeping the irrigation flow low is required. This thesis contributes towards this goal by investigating the problem of turbidity mitigation and presenting a generative model-based framework to disentangle turbidity from lithotripsy datasets.

This thesis presents image analysis methods that can tackle the aforementioned surgical challenges and assist the surgical team by providing supplementary information that can

help them deliver treatment with improved efficiency. Some ureteroscopy frames from animal studies and clinical procedures have been illustrated in Figure 1.2. As can be seen from Figure 1.2, the ureteroscopy and laser lithotripsy data is significantly different and challenging as compared to the other endoscopy datasets in numerous aspects such as - small field-of-view, intra-operative bleeding, motion blur arising from unavoidable kidney movements and camera motion, specular highlights and dynamic background, varying illumination conditions, refractory errors from the irrigation fluid, turbid fluid inside the target cavity, transparent fibre tip, and heterogeneity in the size, shape and composition of the stone. The core technical contributions of this thesis are summarised below:

- Traditional motion-based method for segmentation of stone fragments in ureteroscopy
- Deep learning-based data-driven approach for segmentation of both stone and laser fibre in ureteroscopy
- Generative model-based frameworks for unsupervised domain adaptation and improved generalisation of segmentation models on various endoscopy datasets
- An end-to-end framework for turbidity mitigation to improve visualisation during ureteroscopic laser lithotripsy

1.2 Datasets

The ureteroscopy and laser lithotripsy data used in this thesis has been acquired using single-use LithoVue™ scope and Lithovue Elite scope from Boston Scientific. A summary of the *in vitro* and *in vivo* datasets used in this thesis has been mentioned below (refer to details in individual chapters):

1.2.1 *In vitro* dataset

The *in vitro* dataset was acquired under controlled settings wherein laser lithotripsy of four different human kidney stones was individually performed inside a container with irrigation fluid flowing through it. This dataset was used in various chapters in the following ways:

Chapter 3- Four sequences, each containing three frames (details in Section 3.3)

Chapter 4- 88 sequences, each containing five frames, split into 52 train, 18 validation and 18 test sequences(details in Section 4.2.1)

Chapter 5 - 70 training, nine validation and nine test samples(details in Section 5.2.2.2)

Chapter 6 - 70 training, nine validation and nine test samples(details in Section 6.4.1)

1.2.2 *In vivo* dataset

The *in vivo* dataset was provided by the Oxford University Hospitals (OUH) and Boston Scientific. This dataset consists of samples from clinical procedures and animal studies. The clinical dataset consists of 136 sequences (with five frames each) of clinical samples obtained from four patients. The animal study dataset consists of 18 sequences from one set of animal studies (setting 1), and 20 sequences from another set of animal studies (setting 2) acquired with different hardware settings and software configurations. 456 samples from four new videos (setting 3) were acquired as part of another set of animal studies at Boston Scientific. The *in vivo* dataset was used in various chapters in the following ways:

Chapter 3- One sequence containing three frames (details in Section 3.3)

Chapter 4- 136 clinical sequences and 38 animal study sequences (each containing five frames). The training set consists of 81 clinical and 11 animal sequences (setting 1), the validation set consists of 32 clinical sequences, Test-I (setting 1)consists of 23 clinical and 7 animal sequences, and the Test-II(setting 2) dataset consists of 20 animal sequences(details in Section 4.2.1)

Chapter 5 - 136 clinical samples split into 108 training, 14 validation and 14 test samples. 18 samples from animal study(setting 1) used as out-of-distribution dataset to assess generalisability(details in Section 5.2.2.2)

Chapter 6 - 456 samples (setting 3) from animal studies split into 364 training, 46 validation and 46 test samples(details in Section 6.4.1)

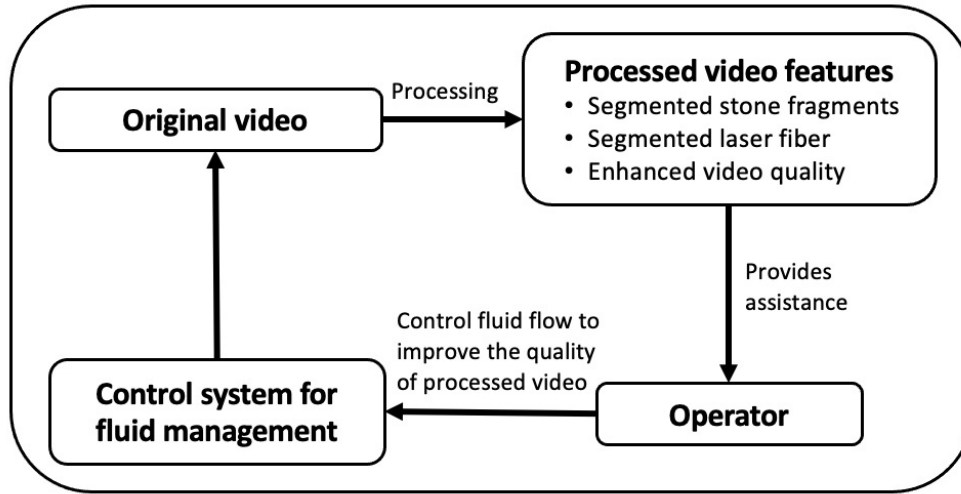


Figure 1.3: Project overview showing how the research presented in this thesis fits into the feedback loop system of the ureteroscopy procedure

1.3 Thesis structure and contributions

The work presented in this thesis was done in collaboration with Boston Scientific and the research plan was influenced by multiple discussions with Boston Scientific. This thesis presents image analysis-based methods for segmentation and image enhancement in ureteroscopy and laser lithotripsy. Figure 1.3 illustrates the vision of how this research fits into the feedback loop system of the ureteroscopy procedure. The enhanced ureteroscopy video can provide supplementary information to assist the surgeon and automated feedback can be sent to the fluid management system to control the fluid flow depending on the quality of the processed video. The structure of this thesis is as follows:

- **Chapter 1** is the introductory chapter that starts with a discussion on the need for image analysis algorithms in ureteroscopy procedures. This is followed by a description of the ureteroscopy and the associated challenges faced by the surgical team during the procedure as they form the clinical motivation behind this thesis. This chapter also provides an overview of the datasets used, thesis structure and related contributions.
- **Chapter 2** starts with a presentation of the clinical background to provide a deeper understanding of the clinical problem at hand. This is followed by a

technical background that consists of a comprehensive review of the existing image segmentation and image quality enhancement methods that are particularly relevant to the work presented in this thesis.

- **Chapter 3** presents a classical approach towards the segmentation of stone fragments in ureteroscopy datasets. The proposed method is a multi-frame-based approach that estimates dense optical flow using a robust data term in a primal-dual optimisation framework to provide segmentation of stone fragments. The proposed method leverages motion fields between multiple frames to tackle problems related to illumination variability, blur, occlusions and deformations.
- **Chapter 4** presents the first work ever reported on multi-class segmentation of ureteroscopy and laser lithotripsy images and is supported by a comprehensive experimental analysis. The proposed framework effectively makes use of residual connections and motion information between adjacent frames to produce robust and reliable segmentation of renal stones and laser fibre.
- **Chapter 5** presents two use cases of VAE-based generative models for tasks beyond simple image generation- generalisation and domain adaptation. The first framework is an end-to-end learning framework called UNet-eVAE that provides iterative refinement of segmentation using VAE-embodied learning. In this proposed framework, VAE-based reconstruction is used as an auxiliary task to enhance region-of-interest (ROI) specific feature representations that act as attention to the segmentation network, thereby providing improved segmentation and generalisation on various ureteroscopy and colonoscopy datasets. The other framework is a novel VAE-based framework called EndoUDA, for target-independent unsupervised domain adaptation for both Barrett’s oesophagus (BE) and polyp datasets.
- **Chapter 6** presents an investigation of the challenging task of turbidity mitigation followed by an introduction of an end-to-end learning framework for the mitigation of stone debris and other factors that contribute to turbidity during laser lithotripsy. The proposed framework is based on the usage of segmentation as an inductive prior for effective disentanglement of latent space to remove turbidity from images.

- **Chapter 7** summarises the major contributions of this thesis and discusses the limitations of the current work whilst also proposing directions for future research in this field.

1.4 Publications

- **Chapter 3- Gupta, Soumya**, Sharib Ali, Louise Goldsmith, Ben Turney, and Jens Rittscher. "Motion-induced segmentation of stone fragments in ureteroscopy video." In *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modelling*, vol. 11315, pp. 290-298. SPIE, 2020.
Award: Runner-up in the Image-Guided Procedures, Robotic Interventions, and Modelling Young Scientist Award at SPIE 2020.
- **Chapter 4- Gupta, Soumya**, Sharib Ali, Louise Goldsmith, Ben Turney, and Jens Rittscher. "MI-UNet: Improved segmentation in ureteroscopy." In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 212-216. IEEE, 2020.
- **Chapter 4- Gupta, Soumya**, Sharib Ali, Louise Goldsmith, Ben Turney, and Jens Rittscher. "Multi-class motion-based semantic segmentation for ureteroscopy and laser lithotripsy." *Computerized Medical Imaging and Graphics* 101 (2022): 102112.
- **Chapter 5- Gupta, Soumya**, Sharib Ali, Ziang Xu, Binod Bhattarai, Ben Turney, and Jens Rittscher, "UNet-eVAE: Iterative refinement using VAE embodied learning for endoscopic image segmentation"- In *Machine Learning in Medical Imaging*, pp. 161–170. Springer, Cham, 2022.
- Filed Invention disclosure forms (IDFs) with Boston Scientific.

Co-authored publications

- Celik, Numan, Sharib Ali, Soumya Gupta, Barbara Braden, and Jens Rittscher. "EndoUDA: a modality independent segmentation approach for endoscopy imaging." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 303-312. Springer, Cham, 2021.

- Xu, Ziang, Sharib Ali, Soumya Gupta, Numan Celik, and Jens Rittscher. "Improved Artefact Detection in Endoscopy Imaging Through Profile Pruning." In *Annual Conference on Medical Image Understanding and Analysis*, pp. 87-97. Springer, Cham, 2021.
- Xu, Ziang, Sharib Ali, Soumya Gupta, Simon Leedham, James E. East, and Jens Rittscher. "Patch-level instance-group discrimination with pretext-invariant learning for colitis scoring." In *Machine Learning in Medical Imaging*, 101–110. Springer, Cham, 2022.

2

Background

Contents

2.1 Clinical background	12
2.1.1 Urolithiasis	12
2.1.2 Ureteroscopy	14
2.1.3 Laser lithotripsy	15
2.2 Technical background	16
2.2.1 Image segmentation	16
2.2.1.1 Segmentation of abnormalities in endoscopy	21
2.2.1.2 Segmentation of kidney stones	22
2.2.1.3 Optical flow-based approaches	23
2.2.1.4 Encoder-decoder based approaches	27
2.2.1.4.1 Residual and recurrent connections based networks	28
2.2.1.4.2 Dilated convolutions and ASPP-based networks	29
2.2.1.4.3 Attention mechanism-based networks	30
2.2.1.5 Auxiliary tasks to improve segmentation	32
2.2.1.5.1 Motion-based image registration	32
2.2.1.5.2 VAE-based image reconstruction	35
2.2.1.6 Metrics for evaluating image segmentation	37
2.2.2 Image denoising and quality enhancement	38
2.2.2.1 Traditional methods	38
2.2.2.2 Learning-based methods	39

This chapter presents a comprehensive review of the literature relevant to this thesis. It starts with a presentation of the clinical context behind Urolithiasis, Ureteroscopy and

Laser lithotripsy. Computer vision research in this field is still in its infancy stage. The clinical background presented in this chapter is followed by the relevant technical literature that covers two specific research questions addressed in this thesis: 1) Image segmentation can help the surgeon in accurate laser targeting and is also the primary step towards the real-time quantitative assessment of stone fragments; 2) Image enhancement by mitigating the turbidity and stone debris can help improve vision during the procedure whilst keeping the irrigation flow low to avoid septic complications. The image segmentation section under technical background covers both traditional and deep learning-based segmentation approaches with an emphasis on medical datasets. The challenging task of image enhancement in ureteroscopy and laser lithotripsy has not been addressed before in the literature. The technical background in this chapter also outlines various state-of-the-art image enhancement approaches that have provided impressive results on other datasets and are relevant to the development of the work presented in this thesis.

2.1 Clinical background

This section presents a description of Urolithiasis followed by an overview of Ureteroscopy and laser lithotripsy procedures.

2.1.1 Urolithiasis

Urolithiasis (or kidney stone disease) affects about 5-15% of the population and is becoming an ever-increasing primary care problem with an ever-increasing demand on healthcare resources [8]. The rising prevalence of kidney stone disease can be majorly attributed to rising obesity, poor dietary habits and inadequate fluid intake [13]. When certain minerals such as calcium oxalate, struvite (ammonium magnesium phosphate), uric acid and cystine supersaturate the urine, they begin to collect and crystallise. These crystals keep compounding together and continue to enlarge with the potential to form stones that cause an obstruction in the urinary tract. These urinary tract stones vary considerably in appearance and size ranging from small gravel-like stones to large calculi stones [14]. The main types of kidney stones include calcium, uric acid, struvite and cystine stones, with calcium stones being the most common ones. Figure 2.1 illustrates



Figure 2.1: Illustration of various types of kidney stones. Figure courtesy of [16]

the variability in the appearance of the various types of kidney stones. Calcium stones contain calcium oxalate, calcium phosphate or both. Calcium oxalate stones have jagged surfaces when made up of calcium oxalate dihydrate (COD) crystals and have a smooth surface in the case of calcium oxalate monohydrate (COM) crystals. Brushite stone, on the other hand, forms small rosettes of thin and sharp crystals. Uric acid stones have a pebble-like appearance and are usually softer on the inside. Struvite stones are typically larger than others and typically form due to urinary tract infections. Finally, cystine stones are formed when an amino acid called cystine leaks into the urine to form compact and partially-opaque stones [15].

Most calculi (stones) < 5mm in diameter pass spontaneously while the ones larger than 5 mm cause obstruction and pain in the abdomen or lower back [8]. Imaging of kidney stones is an important diagnostic tool to decide the most suitable treatment for the management of kidney stones. The available imaging modalities broadly include CT, ultrasound, and MRI [17]. Once characteristics such as the size, number, location, and composition of the stone are determined using imaging modalities, the most suitable treatment option is decided. The treatment options available are PCNL (Percutaneous nephrolithotomy), ESWL (Extra corporeal shock wave lithotripsy) and URS (Ureteroscopy). PCNL is the most invasive technique and is used for the treatment of very large calculi or after the failure of previous surgeries; ESWL uses high-energy shock waves to break up the stone into smaller fragments and is effective for small stones, and URS is usually recommended for calculi in the range 10-20 mm [18].

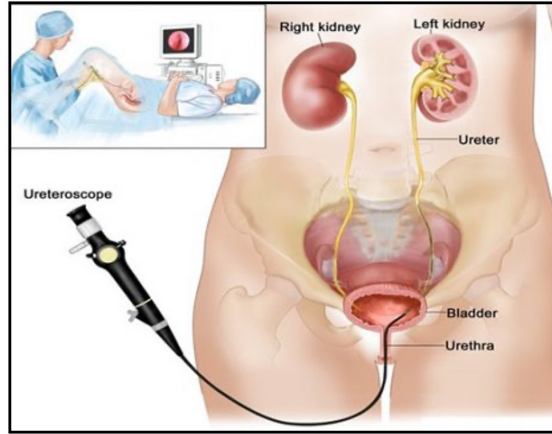


Figure 2.2: Insertion of ureterscope for urinary tract examination. Figure courtesy of [19]

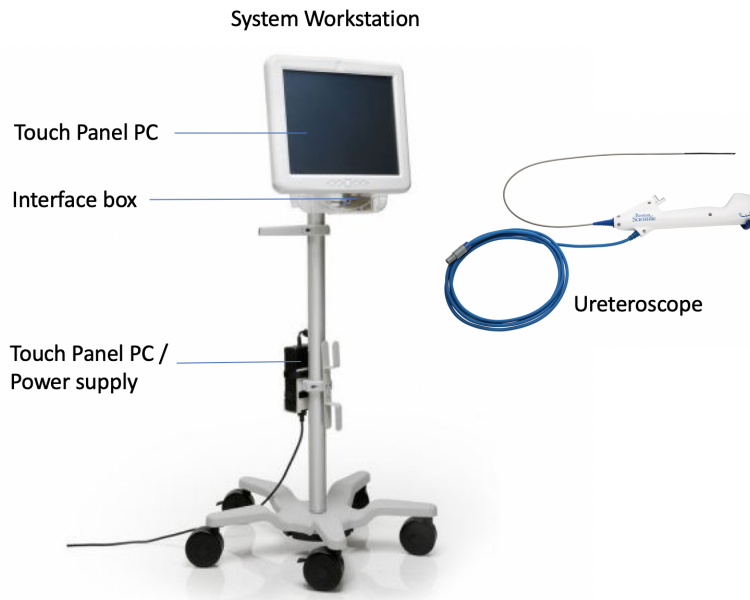


Figure 2.3: LithovueTM and its workstation. Figure courtesy of [20]

2.1.2 Ureteroscopy

Ureteroscopy is performed using a rigid, semi-rigid, or flexible endoscope that consists of a small light and a camera at its tip for examining the urinary tract. The scope is inserted through the urethra, into the bladder, and then up into the ureter and kidney to access the target location as shown in Figure 2.2. The first ureteroscopes were introduced in the 1970s and were rigid, shorter and thicker in nature, thus limiting their use to only the distal end of the ureter. The design was refined with time and semi-rigid ureteroscopes were developed that could reach the proximal ureter [21]. In the 1980s, the first flexible

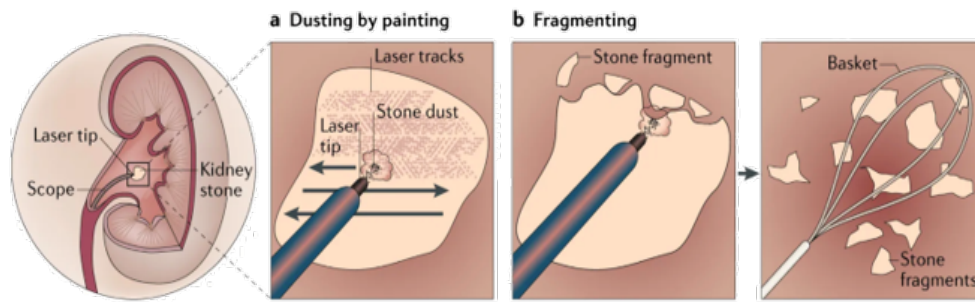


Figure 2.4: Illustration of laser lithotripsy and basketing procedures during ureteroscopy. Figure courtesy of [24]

ureteroscope was developed with the ability to reach the upper urinary tract [22]. However, their use was limited due to the high cost, fragility and poor visibility as compared to semi-rigid scopes. Disposable flexible ureteroscopes have been developed to reduce the overall cost associated with ureteroscopy via the usage of inexpensive multi-lumen catheters, avoiding hassle related to reusable scopes such as instrument-related repairs, sterilisation of the scope shaft, high chances of infection to the patient [21]. Technological advances in the design and size of ureteroscopes have made it possible to access nearly all areas of the urinary tract. In 2015, Boston Scientific introduced the first disposable flexible ureteroscope - LithovueTM [20, 23]. The ureteroscopes are typically connected to an external monitor to display real-time video signal that is used by the surgical team to guide their actions throughout the procedure. Figure 2.3 illustrates the Lithovue scope and its workstation that displays the endoscopy video to the surgical team.

2.1.3 Laser lithotripsy

Ureteroscopy with laser lithotripsy has evolved into a minimally invasive routine procedure for the treatment of kidney stones. Laser lithotripsy induces stone fragmentation when the laser energy is directly transferred to the stone surface during activation of the laser [25]. The ureteroscopes typically have a working channel through which a variety of tools can be inserted such as the laser fibre to perform laser lithotripsy and stone baskets to retrieve the residual stone fragments as shown in Figure 2.4. Holmium lasers are most commonly used for stone fragmentation due to their ability to effectively fragment various types of stones [21]. Based on the size, location and composition of

the stone, the surgeon decides if the stone requires dusting or fragmentation and sets the laser settings accordingly [21]. Dusting is achieved by using Holmium laser at a high frequency with low energy and is used to fragment the stone into dust-size particles that can pass spontaneously. In the case of dusting, the surgeon is required to carefully work tangentially from the edge of the stone towards the centre without breaking off large fragments from the main stone. Fragmentation, on the other hand, is achieved using higher pulse energy with lower frequency wherein the mid-portion of the stone is targeted first to effectively break the stone into halves. Targeting the mid-portion of the stone also maximises the margin of safety between laser fibre and the organ walls [25]. A continuous irrigation flow is maintained during the procedure to dilate the surgical cavity, improve visualisation and bring down the temperature buildup from laser. This irrigation flow is often increased to improve visualisation when it is significantly impaired by stone dust and debris during active lithotripsy [26]. However, increasing irrigation flow increases the intra-renal pressure, potentially resulting in backflow and septic complications [27]. Intrarenal pressure during laser lithotripsy is noted to be as high as 50-350 mmHg while the baseline intrarenal pressure is approximately 10 mmHg. Fragmented stones are either left in place to clear out by themselves over time or extracted using a special wire basket. In order to prevent any obstruction from ureteral oedema or residual stone fragments, a ureteral stent needs to be introduced into the patient to allow for easy passage of the residual stone debris after the ureteroscopy is performed.

2.2 Technical background

This section presents relevant background literature that covers both traditional and deep learning-based approaches for the computer vision tasks focussed in this thesis: image segmentation and image quality enhancement.

2.2.1 Image segmentation

Image segmentation plays an important role in numerous applications such as medical image analysis, traffic control systems, video surveillance, autonomous vehicles, etc. Segmentation can be defined as partitioning an image into segments and assigning a

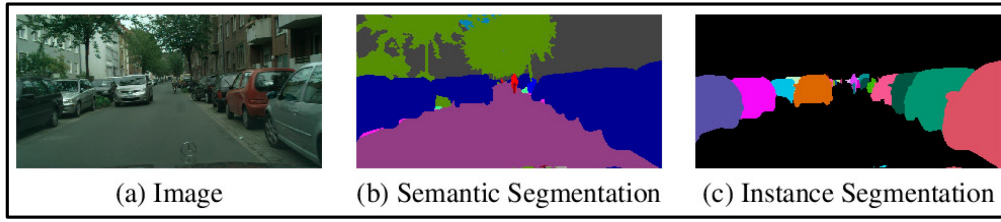


Figure 2.5: Illustration of semantic and instance segmentation. Figure courtesy of [28]

label to each of these segments. Image segmentation can be seen as a classification of pixels into semantic labels (semantic segmentation) or partition of individual objects (instance segmentation) as shown in Figure 2.5. All the image segmentation methods can be broadly classified into two groups - hard computing methods and soft computing methods. Sometimes, hard and soft computing techniques are employed together to obtain cost-effective and high-performance results [29]. Figure 2.6 shows the major types of hard computing and soft computing segmentation techniques.

Hard computing segmentation techniques: Hard computing, also known as the traditional form of computing is based on the analytical model that follows binary logic and numerical analysis. Hard computing methods include techniques such as thresholding [30], region growing [31], edge detection [32], watershed [33], and clustering [34]. Thresholding is a frequently used technique for segmentation when the intensity of pixels to be segmented is completely different to the background while edge-based methods rely on the changes of pixel intensities in the image to identify the edges of the target object. Region-based methods rely on recursive clustering of adjacent pixels that represent similar properties such as intensity. It essentially performs connected-component analysis on the pixel values by comparing neighbouring pixels to the seed pixel value for each pixel in the image. Watershed segmentation is in essence a region-based method that relies on image morphology. It requires the selection of at least one seed point inside each object of the image followed by region growing using morphological watershed transformation. Clustering methods such as k-means is implemented using a function that minimises the distance between data points and the centroid of the cluster to create clusters of similar data points.

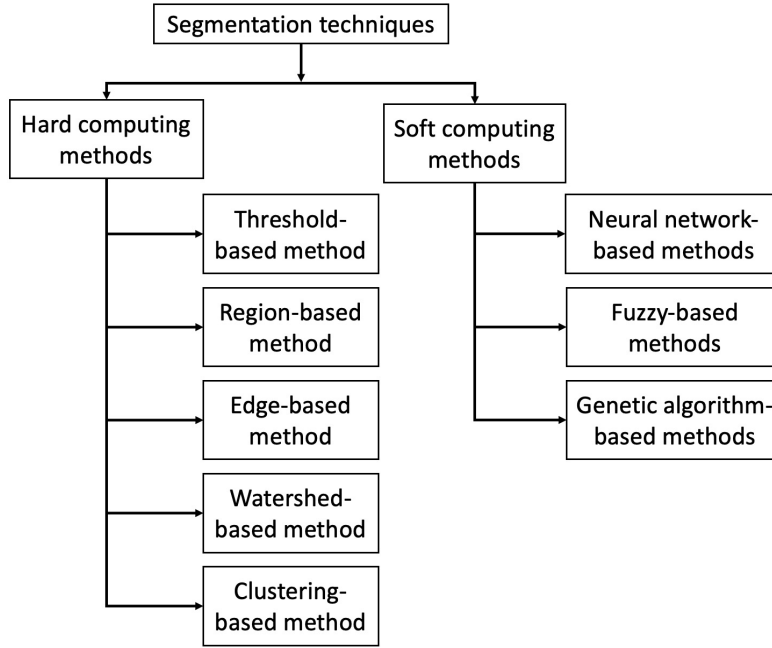


Figure 2.6: Hard and soft computing image segmentation techniques

Soft computing segmentation techniques: Soft computing techniques can be understood as a combination of AI and natural selection that produces quick and robust solutions to complex problems. Soft computing techniques are gaining more attention due to their high adaptability, immense flexibility, and high accuracy. Soft computing techniques include Neural network-based, Fuzzy based, and Genetic algorithm-based approaches [29]. Neural network-based methods are trainable models that consist of one input layer, one or more hidden layers and one output layer. Each node or neuron has a weight and a bias associated with it. The model tries to adjust these weights and biases by minimising the cost function during training until model convergence. Fuzzy-based methods are probabilistic in nature and based on a system of fuzzy logic. It uses membership functions to determine if a pixel belongs to an edge or a homogeneous region. Genetic algorithms allow stochastic exploration of the solution space using strings of integers that represent the parameters to be optimised.

With the advancements of high computational power and large open source datasets in the past decade, deep learning-based segmentation has proven to be the most accurate approach for image segmentation in recent years [35]. A variety of deep learning networks have been proposed for segmentation such as Fully convolutional networks that consist

of convolutional layers only [36, 37], encoder-decoder based networks that consist of an encoder to compress the high dimensional image into a low dimensional feature vector which is then upsampled using a decoder to obtain the output mask [38–40], dilated convolutional models to enlarge the receptive field with no increase in computational cost, and DeepLab family [41, 42], pyramid based models that rely on feature pyramids to merge low and high resolution features [43, 44], recurrent neural networks that rely on sequential processing of images to model short/long term dependencies [45, 46], attention based models wherein attention mechanisms help to weight multi-scale features [47] and capture contextual dependencies [48], generative models that rely on adversarial training [49, 50], and convolutional networks with probabilistic graphical models such as CRFs [51] and MRFs [52].

Challenges in automated segmentation of medical data: Although image segmentation is an essential task in many real-world applications, it is the most crucial component in medical image analysis. Manual interpretation of medical images for diagnosis and prognosis is laborious and prone to error. Therefore, a computer-assisted system is required that can automatically analyse the medical images and provide a more objective and quantitative analysis. Automated segmentation forms an essential part of such computer-assisted systems as it can significantly reduce the manual workload involved in tasks such as quantification of tissue volumes, pathology localisation, examination of anatomical structures, diagnosis, treatment planning, and computer-assisted surgeries. However, medical datasets present several technical challenges to segmentation such as a limited amount of annotated data, heterogeneous pixel intensities (such as high variation in the appearance of renal stones across different patients in the context of ureteroscopy), low resolution, ill-defined or fuzzy boundaries, geometric deformations, imaging artefacts, and high variability between the anatomy of different patients. Endoscopy is one of the medical image analysis applications that has evolved as the key technology for minimally-invasive examinations and minimally-invasive surgeries of different regions of the body. During the endoscopy procedure, a video signal produced by the endoscope is visualised by the surgical team to guide their actions. Automated analysis of this video signal can assist the physician in decision-making. Although the field of various types of endoscopy

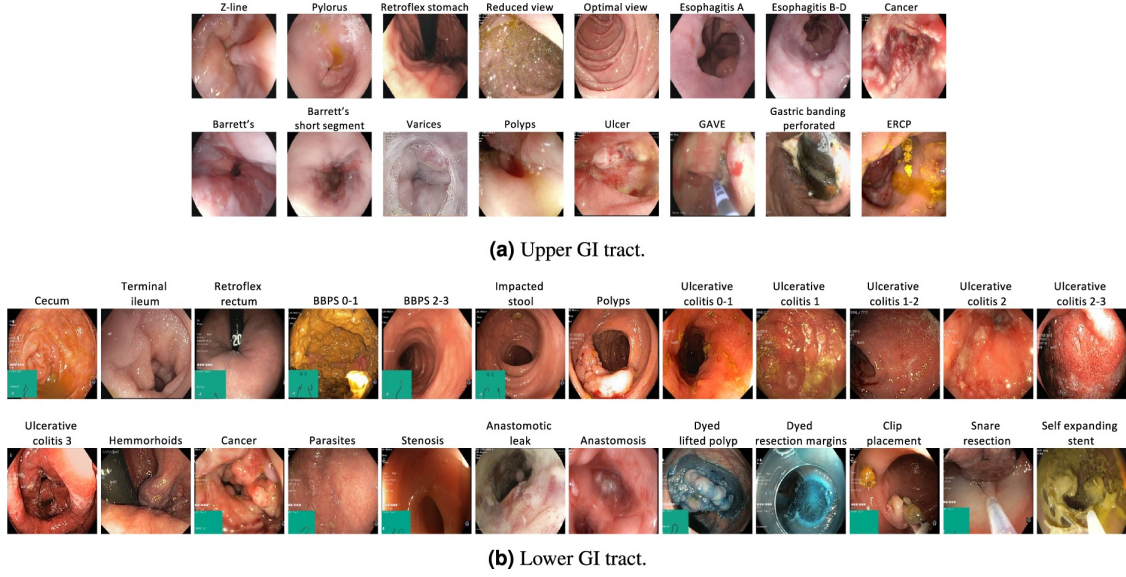


Figure 2.7: Examples of various labelled classes in the HyperKvasir dataset illustrating the wide range of region-of-interests and abnormalities in endoscopy. Figure courtesy of [56]

imaging has been well explored in the field of computer vision [53, 54], work in the field of ureteroscopy and laser lithotripsy is still in its infancy. The ureteroscopy and laser lithotripsy data is significantly different and challenging as compared to the other endoscopy datasets in numerous aspects such as the significant amount of stone debris that obscures the target; dynamic background; high variation in the appearance, size, shape and composition of stone; specular highlights; high motion blur; and additional image artefacts from turbid fluid inside the target cavity [18, 55]. Intra-operative bleeding can also occur during ureteroscopy due to the continuous application of laser energy, intra-renal pressure and the trauma caused to the walls of the ureter [12]. These factors make it difficult for surgeons to perform stone-treatment procedures efficiently, thereby adding to the treatment time.

The following sub-sections present a review of approaches to segmentation of various abnormalities in endoscopy followed by a discussion of the existing approaches to kidney segmentation. This is followed by a comprehensive literature of approaches that inspired the contributions of this thesis - traditional mathematically derived optical flow-related approaches, encoder-decoder-based machine learning approaches and auxiliary tasks incorporated in current deep learning methods to improve segmentation.

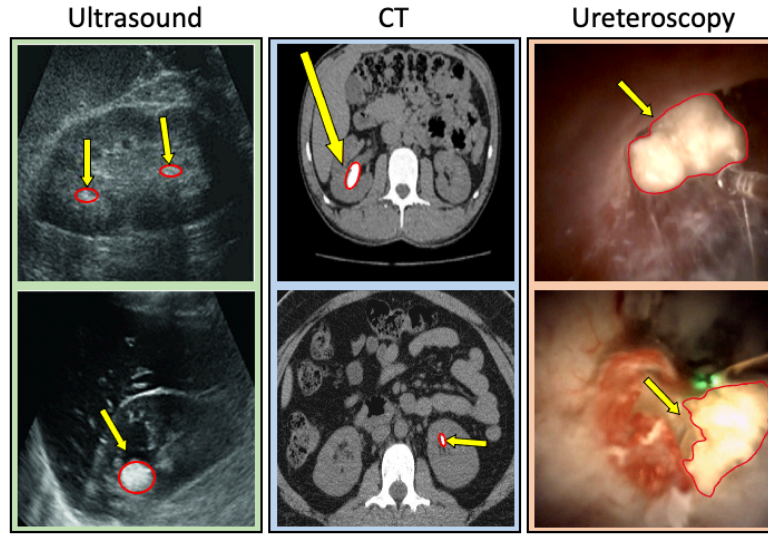


Figure 2.8: Illustration of the boundary of kidney stones in CT [69, 70], Ultrasound [71] and Ureteroscopy images

2.2.1.1 Segmentation of abnormalities in endoscopy

Handcrafted features have been applied for the segmentation and detection of various abnormalities such as bleeding [57], polyps [58], ulcers [59] and tumour regions [60] in endoscopy videos. Various deep learning-based strategies on automated segmentation of bleeding zones in wireless capsule endoscopy (WCE) have been proposed [61, 62]. Mask R-CNN has been employed for polyp segmentation in several works [63, 64]. Ali *et al.* [65] presented a comprehensive analysis of various approaches that were submitted to the Endoscopy Artefact Detection (EAD2020) challenge and Endoscopic Disease Detection 2020 (EDD2020) challenge. A multi-scale context-guided deep network based on FCN was proposed [66] for lesion segmentation in endoscopy images of the Gastrointestinal (GI) tract. Several deep learning-based approaches for the detection and segmentation of colon polyps have been proposed [54, 67, 68]. Several groups have investigated the segmentation of various abnormalities such as bleeding [61, 62] and other lesions [66] in endoscopy. Figure 2.7 presents example images from the HyperKvasir dataset illustrating the wide range of region-of-interests and abnormalities in endoscopy [56]. Although the field of endoscopy imaging has been well explored in literature, very little work has been done towards segmentation in ureteroscopy and laser lithotripsy.

2.2.1.2 Segmentation of kidney stones

Several methods for the detection and segmentation of kidney stones in Ultrasound (US) images and Computed Tomography(CT) images have been proposed in the literature. Methods such as Region indicator with Contour segmentation (RICS) [72], modified watershed segmentation [73], and squared Euclidean distance method [74] have been implemented for the detection and segmentation of kidney stones in ultrasound (US) images. Some studies have also explored techniques such as intensity, location, and size-based thresholds [75], Fuzzy C-means clustering followed by level set [76], total variation (TV) flow and Maximally Stable Extremal Region (MSER) features [77], and CNN [78] for detection and segmentation of kidney stones in CT images. Machine learning-based active contour method has also been proposed for automated segmentation of kidney stones in CT images [79]. Rosa *et al.* [18] proposed a region-growing algorithm for renal calculi segmentation on ureteroscopy images. However, such approaches require user intervention to define seed pixel, similarity criterion, and a stopping criterion which is very challenging to determine due to the nature of variability in kidney stones in terms of colour, texture and composition (as shown in Figure 2.1). Most of these methods are based on radiology datasets and automated segmentation of kidney stones in ureteroscopy is not well-explored. This thesis proposes an optical-flow-based segmentation technique for binary segmentation of stone fragments in ureteroscopy [80]. The aforementioned methods either use traditional segmentation approaches or machine learning-based techniques that require manual interventions and cannot provide real-time inference. This thesis further proposes an end-to-end convolutional network that leveraged motion differences between adjacent frames to provide real-time segmentation of stones [81]. In addition to stones, automated segmentation of laser fibre can also help in the accurate targeting of stones. Therefore, a deep learning-based multi-class segmentation of both stones and laser fibre in ureteroscopy and laser lithotripsy has also been proposed in this thesis [82]. Figure 2.8 illustrates the boundary of kidney stones in Ultrasound, CT and Ureteroscopy datasets.

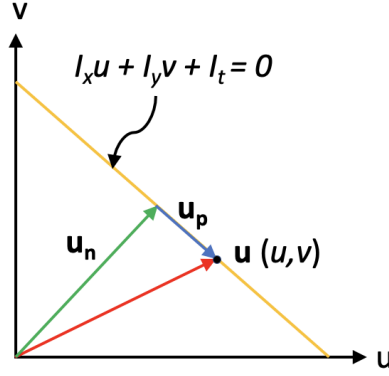


Figure 2.9: Plot of optical flow equation illustrating the normal and parallel components of optical flow

2.2.1.3 Optical flow-based approaches

This section presents background on the estimation of optical flow followed by relevant approaches for modelling of variational optical flow.

Optical flow is a well-established computer vision technique that has been successfully used in a wide variety of applications including image segmentation [83–85]. It can be defined as a pattern of motion of pixels in a frame sequence. Optical flow works by assigning a separate displacement vector to each pixel of an image that can then be used to estimate the direction or speed of each pixel of objects in each frame of the sequence.

The optical flow equation is given by :

$$I_x u + I_y v + I_t = 0 \quad (2.1)$$

where u and v represent the pixel's velocity in the x direction and y direction, respectively. The optical flow $\mathbf{u}(u, v)$ can be split into two components such that : $\mathbf{u} = \mathbf{u}_n + \mathbf{u}_p$ where $\mathbf{u}_n$ and $\mathbf{u}_p$ represent the normal and parallel flow as shown in Figure 2.9. From the optical flow plot in Figure 2.9, one can see that the normal component $\mathbf{u}_n$ can be calculated as: $\mathbf{u}_n = \frac{I_t}{\sqrt{I_x^2 + I_y^2}}$ but the parallel component cannot be computed as there are infinite possible values for u and v . This is known as the aperture problem.

To solve this ill-posed problem, differential methods of optical flow estimation have been proposed. Differential methods of optical flow based on the calculation of spatiotemporal image derivatives are classified into local and global approaches[86]:

- **Local approach:** Lucas and Kanade [87] proposed a local method that is based on the assumption that the optical flow field is constant within a small neighbourhood W . Let $\mathbf{x} = (x, y)$ represent the centre pixel of neighbourhood W and $\mathbf{x}' = (x', y')$ represent the neighbourhood pixels with a weighting function $h(\mathbf{x}', \mathbf{x})$, then the optical flow of the current pixel $\mathbf{x}$ can be estimated by minimising the squared errors in the neighbourhood $W(\mathbf{x})$ as:

$$E(\mathbf{u}) = \sum_{\mathbf{x}' \in W(\mathbf{x})} h(\mathbf{x}', \mathbf{x}) (\Delta I(\mathbf{x}', t)^T \mathbf{u} + I_t(\mathbf{x}', t))^2 \quad (2.2)$$

Equation 2.2 can be minimised and solved using the least squares estimate to obtain the solution for optical flow. It has been observed that the local methods are highly robust under noise but give sparse flow fields. Moreover, the assumption of constant motion within a neighbourhood is generally not applicable in real-world scenarios.

- **Global approach:** Global methods are based on the smoothness constraint assumption. These approaches give dense flow fields but are more sensitive to noise [88]. Horn and Schunck [89] laid the foundation of dense optical flow algorithms [89] in 1981 and has been an active area of research since then. The optical flow determination can be formulated as an optimisation problem with two terms: a data term given by the optical flow constraint and a regularity term based on flow gradients.

$$E = \min_{u,v} \left\{ \int_{\Omega} (I_x u + I_y v + I_t)^2 d\mathbf{x} + \lambda^2 \int_{\Omega} (\nabla^2 u + \nabla^2 v) d\mathbf{x} \right\} \quad (2.3)$$

In Equation 2.3, λ controls the relative weighting of the two terms. Minimising this energy cost function will give us the optimal solution for optical flow. A major drawback of this method is that the regularization term penalises the high gradients, thereby resulting in an over-smoothing effect. Also, global approaches fail to robustly handle the outliers in the data term. However, global methods provide the most accurate estimation of optical flow [90].

Optical flow estimation for medical datasets is particularly challenging due to high variability in textures and illumination conditions, local deformations, occlusions and

image artefacts such as motion blur. Various data terms and regularisation terms have been proposed in the literature to improve the original framework proposed by Horn and Schunck [89].

- **Data terms:** The classical data term in the optical flow equation 2.3 is based on the brightness constancy assumption. However, this does not hold true for real-life scenarios due to variable illuminations especially in cases of medical images due to differences in imaging protocols. Gradient constancy assumption (GCA) was proposed to tackle global illumination changes by assuming spatial brightness gradient to be constant over time. Data terms that combine brightness constancy assumption (BCA) and gradient constancy assumption (GCA) or higher order constancy assumptions (HCA) such as Hessian and Laplacian have been formulated to deal with the illumination variations [91–93]. These are still not effective enough to represent the complex illumination variations that are present in video sequences [93]. Optical flow methods based on handcrafted image features have been quite successful in handling variances in illumination. The clustering of feature space formed by image intensity values and optical flow vectors has been used to provide spatiotemporal segmentation of MR image sequences [83]. A hybrid technique employing gradient-based optical flow and shape-based matching was proposed for the detection of left ventricle motion and segmentation of a beating human heart [84]. A combination of active contour segmentation and dense deformation field from optical flow in a level-set framework was adopted to perform segmentation and atlas registration [85].

Classical coarse-to-fine optical flow formulation helps to deal with large displacements using a multi-scale pyramid strategy. This enables estimation of the flow field at coarser scales that is iteratively refined at finer scales [91, 94]. However, a multi-scale strategy may result in the disappearance of small objects or pixels with weakly contrasted textures at coarser levels. To tackle this issue, large displacement optical flow methods such as SIFT flow [95] and LDOF [96] were proposed. Such methods integrate sparse key-point correspondences obtained by descriptors and

use them as guidance in the variational framework to retrieve flow fields for large displacements. To improve sub-pixel accuracy and preserve motion discontinuities, Brox and Malik [96] proposed to combine descriptor matching with classical data terms such as brightness constancy and gradient constancy. Such approaches are still not effective for medical datasets where the textures are less or weakly contrasted and images have several artefacts such as blur and large illumination variances.

Self-similarity patterns have been very successful in handling large illumination variations, large displacements and motion blur [97, 98]. Chen *et al.* [98] proposed nearest neighbour field-based matching but this was computationally expensive due to additional steps such as outlier rejection and graph-cut. Complete rank transform (CRT) based on census transform was proposed [99] which involved descriptor build by counting the number of pixels in a local patch that has smaller intensity w.r.t. to the centre pixel. However, census transform fails to distinguish between dark and bright regions in a neighbourhood [93]. Modality-independent neighbourhood descriptor for multi-modal deformable registration was proposed [100]. Papiez *et al.* adapted the original MIND descriptors to deal with multiple image channels for optical microscopy [101]. Drulea *et al.* [97] proposed zero-mean normalised cross-correlation (ZNCC) as a matching cost to robustly overcome illumination changes. It has been shown that utilising normalised correlation transform-based similarity metric is more robust to illumination changes for computing deformation fields between histology brain slices and polarised light microscopy images [102]. A novel non-local shape descriptor that used normalised cross-correlation to enable the extraction of most descriptive geometric features in medical images while being independent of intensity relationships and noise was proposed [103]. To decrease computational cost and enable more degrees of freedom in non-rigid registration whilst reducing sensitivity to image noise, Heinrich *et al.* also proposed contextual image descriptors for the registration of 3D ultrasound and brain MRI scans [104].

- **Regularizer:** Horn and Schunck proposed to incorporate a smoothness constraint in the energy minimisation scheme to tackle the aperture problem. The spatial regularizer proposed by Horn and Schunck is modelled as a quadratic penalisation of the flow-field gradients (refer to Equation 2.3). However, this homogeneous regularisation often over-smoothes the flow field discontinuities resulting in blurry optical fields. Over the past few years, several regularizers have been proposed to obtain robust and accurate optical flow fields. Regularisation methods can be divided into flow-driven and image-driven models [105]. Image-driven models [106, 107] use image gradient-based information to penalise the regulariser in order to preserve motion discontinuities. However, image-driven models can often result in the over-segmentation of regions due to inaccuracies in preserving flow discontinuities. Flow-driven models [105, 108] use information about the flow gradients to penalise the regulariser. Several authors [97, 109] proposed weighted non-local regularisation methods to robustly preserve motion discontinuities. These methods assume that the flow is constant for all pixels belonging to the same object. Bilateral filtering has been used in various optical flow and image registration algorithms in order to reduce the over-smoothing of boundaries in medical imaging [110–112]. This is achieved by weighting the flow field of pixels such that the weight of a pixel decreases as its spatial distance or colour distance (difference in pixel intensity) from the centre increases.

2.2.1.4 Encoder-decoder based approaches

This section presents a review of deep learning architectures used for segmentation with a particular focus on advancements in encoder-decoder networks such as residual and recurrent connections [113], dilated convolutions and atrous spatial pyramid pooling [114], and attention mechanisms [115].

UNet [40] is one of the most popular encoder-decoder models used to boost the segmentation accuracy in biomedical datasets. As shown in Figure 2.10, the left side of the architecture represents the contracting path that down-samples the spatial size of feature maps while doubling the number of feature channels. On the other hand, the

right-hand side of the architecture represents the expansive path that up-samples the feature maps while reducing the number of feature maps. Skip connections in UNet help to provide spatial information to each level of the decoder so that it can recover the fine-grained spatial details of the image. The final layer consists of a 1x1 convolution to map the feature vector to the desired number of classes in the segmentation map.

2.2.1.4.1 Residual and recurrent connections based networks Figure 2.11 illustrates different variants of the convolutional, residual and recurrent units. These have been used to formulate different variants of UNet that have been shown to achieve high accuracy results in a variety of medical applications [116–120]. Convolutional layers were replaced by residual inception blocks to obtain an improved UNet for nuclei segmentation on histology data [116]. UNet has also been used for polyp detection in wireless capsule endoscopy [117] and semantic segmentation on laryngeal endoscopic images [119]. Classical 2D UNet has been extended to 3D UNet for volumetric segmentation that learns from sparse annotations [118]. A network called V-net that consists of an FCN with residual connections was proposed for 3D medical image segmentation based on volumetric images [120]. A number of past studies have suggested that increasing the hidden layers in a deep neural network improves its performance. However, it is quite

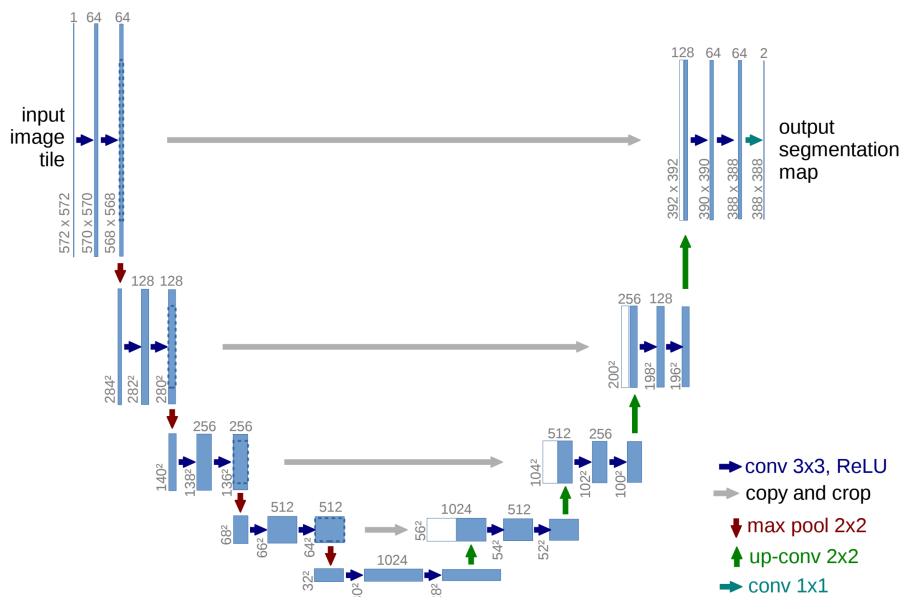


Figure 2.10: Illustration of the UNet architecture. Figure courtesy of [40]

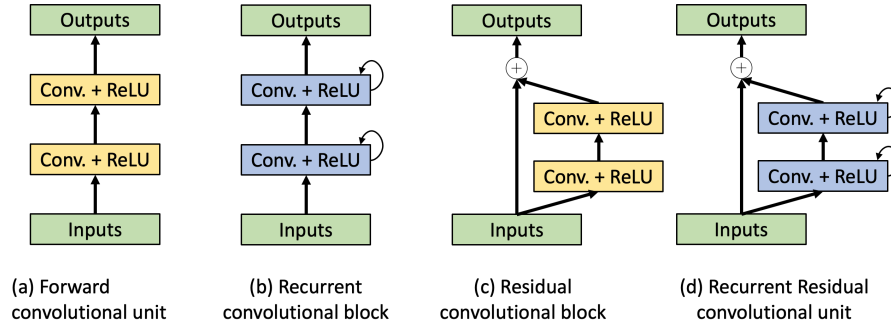


Figure 2.11: Illustration of different variants of the convolutional, residual and recurrent units

difficult to train a deeper architecture due to the problem of vanishing gradients. This problem was solved by the introduction of a deep residual learning framework which employs shortcut connections that perform identity mapping to facilitate training [121]. Zhang *et al.* [122] improved UNet performance by adapting a deep residual UNet architecture (DeepResUNet) that combined the strengths of deep residual learning and UNet architecture wherein the encoder, bridge and decoder were all comprised of residual units. However, such replacement of all feed-forward units with residual blocks may increase the network complexity and tend to overfit the training data [123]. Peretz *et al.* [123] then suggested a hybrid version wherein residual blocks are only used in the encoding path of the UNet. A UNet with residual connections has also been recently proposed for instrument segmentation in endoscopic images [124]. A recurrent CNN wherein both feed-forward and recurrent connections have local connectivity has also been shown to significantly boost the performance of deep learning models [125]. In 2018, Alom *et al.* proposed a Recurrent residual UNet (R2-UNet) [113] that uses recurrent convolutional layers with residual connectivity for stronger feature representation. This network contains the same number of parameters as UNet and was shown to improve the segmentation accuracy of blood vessels, skin cancer and lung segmentation as compared to state-of-the-art UNet.

2.2.1.4.2 Dilated convolutions and ASPP-based networks Dilated convolutions enable enlarging the receptive field of the filters to incorporate a larger context without

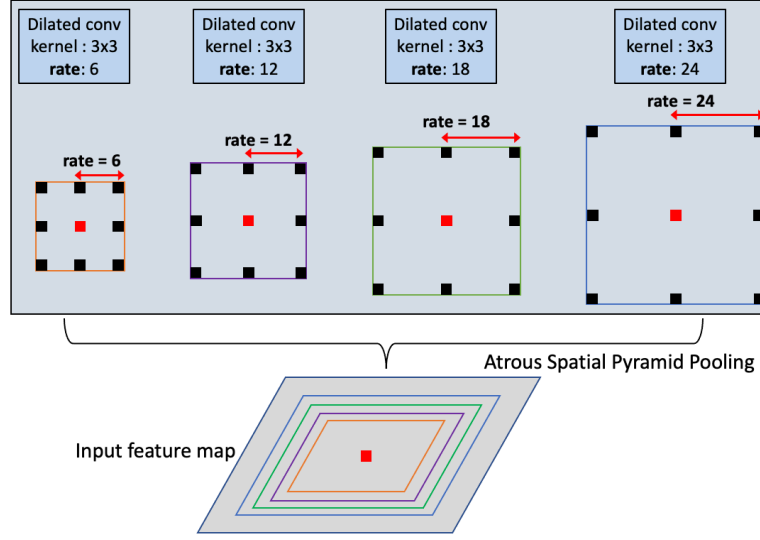


Figure 2.12: Illustration of Atrous spatial pyramid pooling that consists of parallel dilated convolutions with different rates

increasing the computational cost. Several studies have proposed that replacing conventional convolutions in CNN models with dilated convolutions significantly improves the network performance [114, 126, 127]. However, increasing dilation factors tend to increase the sparsity of the kernel and fail to aggregate local features [126]. A novel architecture was then proposed for the segmentation of small object instances in remote satellite imagery by first increasing the dilation factors and then decreasing them [126]. Atrous spatial pyramid pooling (ASPP) that involves parallel atrous convolutions with different dilation rates has also boosted the accuracy of detection and segmentation of objects at different scales [41] (see Figure 2.12).

Recently, a variant of the residual UNet called ResUNet-a [128] was proposed where atrous convolutions and pyramid scene parsing pooling was incorporated in the network to improve the segmentation accuracy. Nyugen *et al.* [129] presented a deep encoder-decoder-based network that consisted of atrous convolutions and depth-wise separable convolution for polyp segmentation in colonoscopy images. Sun *et al.* [130] proposed a UNet with dilated convolutions for the segmentation of colorectal polyps.

2.2.1.4.3 Attention mechanism-based networks In deep learning, attention mechanisms are meant to mimic cognitive attention by diminishing non-relevant information

and only enhancing the important parts of the data. This can help the networks learn better by making them focus on the important parts of input data. Attention mechanisms have proven to be effective in highlighting only the relevant activations during training and are computationally efficient. Attention-based approaches can be classified into two: hard attention and soft attention. Hard attention highlights relevant regions by image cropping or iterative region proposal. However, hard attention is non-differentiable and requires reinforcement learning to train. Soft attention, on the other hand, works by weighting different parts of the image based on their relevance. It is a differentiable approach and can be easily trained using back-propagation.

They have been used in a wide variety of applications: image captioning [131], machine translation [132] and classification tasks [133]. Soft attention can be implemented at the skip connections of UNet to actively suppress activations in irrelevant regions and avoid redundant features to pass through. An attention-gated model (shown in Figure 2.13) was proposed and integrated into the standard UNet for improved pancreas segmentation [115]. As shown in Figure 2.13, attention gate (AG) takes two inputs, $\mathbf{x}$ and $\mathbf{g}$. The gating vector $\mathbf{g}$ contains contextual information to prune the feature maps of lower levels. Input tensors, $\mathbf{x}$ and $\mathbf{g}$ are linearly transformed using channel-wise 1×1 convolutions. The vectors $\mathbf{x}$ and $\mathbf{g}$ are then summed element-wise to increase the weights of the aligned regions. This is followed by ReLU activation, a 1×1 convolution, and a sigmoid layer to scale the output attention coefficients in the range $[0,1]$. The attention coefficients are then upsampled to the original dimensions using trilinear interpolation. Finally, the attention weights are multiplied with the original vector $\mathbf{x}$ to scale the regions based on relevance and passed along in the skip connections of the UNet. AG was also integrated with R2-UNet [113] for improved segmentation [134]. Jha *et al.* [135] proposed ResUNet++ that took advantage of residual units, ASPP, and attention units to provide improved segmentation of colorectal polyps. Methods for detecting bleeding [61, 62], segmentation of various lesions such as ulcer, cancer, etc in the GI tract [66] and polyp segmentation [54] can also be investigated in the context of ureteroscopy.

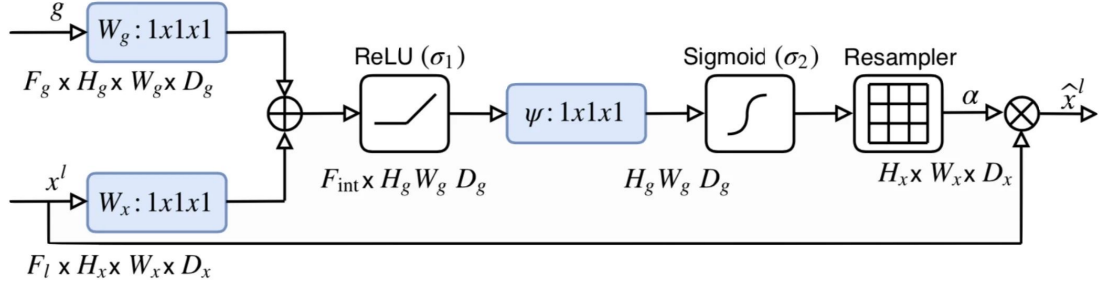


Figure 2.13: Schematic of the attention gate (AG) mechanism. Figure courtesy of [115]

2.2.1.5 Auxiliary tasks to improve segmentation

Image registration and image reconstruction can be used as auxiliary tasks to improve segmentation results. This section presents relevant background on these auxiliary task approaches that have inspired my contribution towards improving deep learning-based segmentation.

2.2.1.5.1 Motion-based image registration Image registration refers to the process of finding an optimal spatial transformation that best aligns the underlying structures in images. It has been widely used for various applications such as image guidance, motion tracking, segmentation, image reconstruction, and so on [136]. Several studies have shown the complementarity of image registration and segmentation tasks, meaning the features learned in image registration can be also used to improve the segmentation result [137, 138]. The registration algorithms can broadly be divided into two types: traditional approaches and learning-based approaches.

Traditional methods. The traditional non-learning approaches rely on the iterative optimisation of transformation based on an energy function. Horn and Schunck laid the foundation of variational methods of optical flow estimation in 1981 [89] and has been an active area of research since then. Illumination constancy and spatial coherence is the most basic assumption used in most optical flow algorithms. But, this assumption does not hold true for real-world applications, especially in cases of medical images due to image artefacts and different imaging protocols. Also, such methods get stuck in local minima and tend to fail against large displacements. This issue was solved by the introduction of Large Displacement Optical Flow that used local descriptor (e.g.

HOG) matching with the variational method [96]. Such matching results are used to initialise the flow which is then optimised using the variational method in a coarse-to-fine pyramid strategy. Optical flow methods based on descriptors such as Weber local descriptor (WLD) [139], complete rank transform (CRT) [140], and Modality independent neighbourhood descriptors (MIND) [100] have been quite successful in handling the illumination variances. CNN-based patch matching that uses convolutions and max pooling has also been proposed in the state-of-the-art DeepFlow method [141]. Rigid transformation models of registration are often too simple to take the tissue deformations into account. To tackle this, deformable image registration (DIR) methods should be used. Deformable image registration can be defined as the non-linear process of estimating the spatial correspondence between source and target images. Some non-learning-based approaches for DIR have gained tremendous popularity such as Demons [142], diffeomorphic demons [143], automatic image registration (AIR) [144], HAMMER [145], FNIRT [146], and so on. Several variants [147, 148] of the efficient Demons algorithm [142] have been proposed over the years depending on how the forces are computed. One of the main limitations of the demons algorithm is that it does not provide diffeomorphic transformations. Diffeomorphisms are powerful as they help preserve the object topology whilst preventing any folding and are considered to be a good working framework when no additional information about spatial transformation is available [143]. Using symmetrical forces helps to achieve faster convergence and makes the result independent of the order of input images. However, standard diffeomorphic registration methods do not formulate the transformation optimisation symmetrically. Avants *et al* [149] proposed a novel symmetric image normalisation method to maximise cross-correlation. Avants *et al* [150] presented a detailed study evaluating the impact of Advanced Neuroimaging Tools (ANTs) similarity metrics such as squared intensity difference, cross-correlation, and voxel-wise mutual information on affine and deformable registration. Traditional methods of image registration are iterative and suffer from low generalisability. These methods often fail when there is a lot of variation in appearance between the source and target images. Deformable multi-modal registration methods have been proposed in the literature to deal with variations in appearance that arise due

to the images being acquired using different image modalities [104]. The limitations of traditional approaches have been overcome to a great extent by the deep learning-based models of image registration [151] that can be categorised into supervised and unsupervised approaches.

Learning-based DIR methods. Several supervised DIR approaches have been developed in the literature. The idea of using CNNs for optical flow estimation in an end-to-end supervised learning manner was proposed with the introduction of FlowNet [152]. This idea was also introduced in several other networks [153, 154]. A CNN-based regression model was developed for brain MR images to directly learn the mapping between source and target images to their corresponding DVFs [155]. Yang *et al.* [156] also proposed a network to predict deformable registration followed by its refinement using a correction network for brain MR images. However, such neural network methods rely on strong supervision for training. The use of supervised methods is majorly limited by the fact that they require ground-truth deformation vector fields (DVFs) for model training which is difficult to obtain, especially in the case of medical datasets. Unsupervised DIR methods have gained tremendous success and popularity over recent years. Shan *et al.* [157] developed an unsupervised end-to-end deep learning framework that can directly predict deformation fields for medical image registration. Zhang proposed an Inverse-Consistent deep Network (ICNet) for unsupervised deformable image registration and evaluated this method on T1-weighted brain MRI scans for tissue segmentation and anatomical landmark detection [151]. In this work, authors have developed an inverse-consistent constraint that encourages a pair of images to be symmetrically deformed toward one another until both warped images are matched [151]. Guha *et al.* proposed an unsupervised CNN model for the brain MRI dataset wherein the deformation fields were up-scaled using linear interpolation [158]. Such models are, however, too simple to properly handle large non-linear deformations. In order to handle large non-linear deformations, De *et al.* [159] used B-spline for transformation and interpolation for the predicted deformation fields and presented results on many more complex datasets. Here, B-spline does not pass through all data points and can often lead to large interpolation errors [160]. Ali *et al.* [160] presented an unsupervised

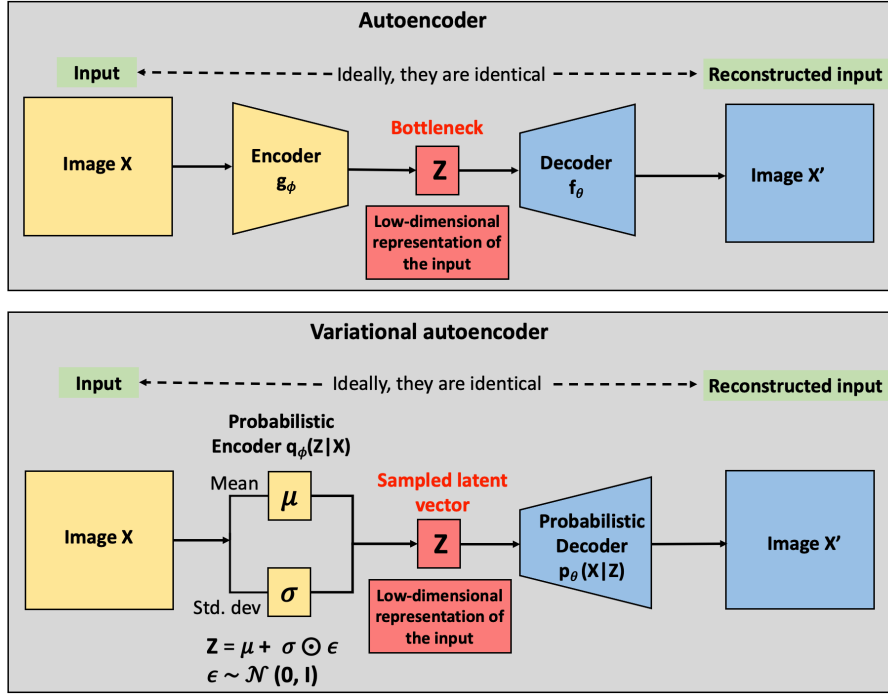


Figure 2.14: Illustration of an autoencoder and variational autoencoder model architecture

end-to-end CNN framework for image registration that used a bi-cubic Catmull-Rom spline re-sampler to reduce the errors in the re-sampling of deformation fields. They also added a series of deformable convolutional filters to better capture complex deformations.

2.2.1.5.2 VAE-based image reconstruction Autoencoder neural network is used to learn an identity function in an unsupervised way to reconstruct the input while compressing the original input into a low-dimensional representation [161]. As shown in Figure 2.14, autoencoder consists of an encoder function $g(\cdot)$ and a decoder function $f(\cdot)$ parameterised by ϕ and θ , respectively. The parameters (ϕ, θ) are learned together to output reconstructed data such that it is ideally similar to the original input data while mapping the input data into a fixed vector, $\mathbf{Z}$. Variational autoencoder [162], on the other hand, involves mapping the original data into a distribution, say, p_θ . The relationship between input data $\mathbf{X}$ and latent vector $\mathbf{Z}$ can be defined by : prior $p_\theta(\mathbf{Z})$, likelihood $p_\theta(\mathbf{X}|\mathbf{Z})$, and posterior $p_\theta(\mathbf{Z}|\mathbf{X})$. As shown in Figure 2.14, $q_\phi(\mathbf{Z}|\mathbf{X})$ and $p_\theta(\mathbf{X}|\mathbf{Z})$ represent the probabilistic encoder and probabilistic decoder, respectively. The encoder maps the original data sample into a compressed low-dimensional latent space distribution. A latent

vector sampled from this distribution is then fed into the decoder that tries to generate data from the conditional distribution $p_{\theta}(\mathbf{X}|\mathbf{Z})$. However, we know that sampling is a stochastic process meaning gradients cannot be backpropagated. To solve this issue, the reparameterization trick is used wherein mean μ and variance σ of latent space distribution is learned while stochasticity comes from the random variable $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The loss function of the Variational autoencoder is a combination of negative log-likelihood and a Kulback-Leibler (KL) divergence-based regulariser, as expressed below:

$$L_{VAE} = -\mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z}|\mathbf{X})} \log p_{\theta}(\mathbf{X}|\mathbf{Z}) + D_{KL}(q_{\phi}(\mathbf{Z}|\mathbf{X}) \parallel p_{\theta}(\mathbf{Z})) \quad (2.4)$$

Equation 2.4 is known as the variational lower bound, or evidence lower bound (ELBO) where we can minimise the loss by maximising the log-likelihood of generating real data (first term in Equation 2.4) and minimising the gap between real and estimated posterior distributions (second term in Equation 2.4).

In the field of deep learning, multiple tasks when learned simultaneously by a shared model can reduce overfitting and aid improved learning by leveraging auxiliary information [163]. It has been shown that by interconnecting the reconstruction and the segmentation networks, one can achieve sharper reconstructions and more accurate segmentation results [164]. Several works that have shown improvement in segmentation are inspired by UNet-based encoder-decoder architecture [40] with some recent works extending network architecture using dual-decoders [165], double encoder-decoder networks [166], residual blocks, and attention mechanisms [167, 168]. Tomar *et al.* [165] proposed a dual-decoders-based approach with one decoder acting as a segmentation network while the other acts as an autoencoder network that helps strengthen feature maps of the encoder. Galdran *et al.* [166] proposed an approach where two pretrained encoder-decoder networks are sequentially stacked, and the output from the first network acts as attention, enabling the second network to provide improved segmentation. Yeung *et al.* [168] proposed a dual-attention gated CNN for polyp segmentation. PraNet [167] used encoder-decoder architecture with added reverse attention mechanism for refining polyp boundaries [167]. Similarly, for the kidney stone segmentation, UNet-based architecture was used together with a CNN-based DVNet that provided dense deformation between

adjacent frames [81]. The authors showed that including such auxiliary tasks in the network improved the segmentation of kidney stones; however, the use of multiple frames to correct the segmentation makes it inefficient. Also, most of these works extended UNet by stacking modules in series [166, 168] which increases network complexity and can adversely affect training on smaller datasets.

Variational auto-encoders (VAE) have been successfully used for representation disentanglement and data generation [169]. Kohl *et al.* proposed a generative segmentation model based on a combination of UNet with a conditional VAE to produce plausible segmentation variants for lung abnormalities segmentation [170]. Similarly, Kong *et al.* proposed a skip connection-based VAE-UNet model to improve 3D brain tumour segmentation [171]. The underlying hypothesis behind this approach is that the VAE regularises the weights of the shared encoder to improve the segmentation performance. These networks attempt to project the embedding to latent space for the entire images rather than a specific region of interest. Additionally, in [170], the ground truth mask is incorporated together with original image samples. However, this will constrain VAE representations towards the samples present in the majority of the training set and can affect the generalisability.

2.2.1.6 Metrics for evaluating image segmentation

Once a model is trained for segmenting images into a set of classes, its performance can be quantitatively evaluated using a range of metrics. Some of the most commonly used metrics are mentioned below. Here, TP (True positive): represents the number of positive pixels classified as positive; TN (True Negative): represents the number of negative pixels classified as negative; FP (False Positive): represents the number of pixels that are actually negative but are incorrectly predicted to be positive; FN (False Negative): represents the number of pixels that are actually positive but are incorrectly predicted to be negative.

- Precision/Positive Predictive value (PPV): defined as the number of true positives divided by the number of all positives

$$Precision = \frac{TP}{TP + FP}$$

- Recall/ Sensitivity: defined as the number of true positive results divided by the number of all samples that should have been identified as positive

$$Recall = \frac{TP}{TP + FN}$$

- Dice Coefficient (DSC) / F1 score: defined as the harmonic mean of precision and recall

$$DSC = \frac{2TP}{TP + FP + FN}$$

- Jaccard Index(JI)/ Intersection over Union (IoU): defined as an area of intersection over the union of the predicted segmentation and the ground truth. IoU penalises the over- and under-segmentation more than DSC

$$IoU = \frac{TP}{TP + FP + FN}$$

- Hausdorff Distance: average Hausdorff distance is a widely used performance measure to calculate the distance between two point sets. The average Hausdorff distance between two finite point sets Y^{gt} and Y where these refer to ground truth and predicted segmentation mask, respectively, can be given by:

$$d(Y^{gt}, Y) = \left(\frac{1}{Y^{gt}} \sum_{y^{gt} \in Y^{gt}} \min_{y \in Y} d(y^{gt}, y) + \frac{1}{Y} \sum_{y \in Y} \min_{y^{gt} \in Y^{gt}} d(y^{gt}, y) \right) / 2$$

2.2.2 Image denoising and quality enhancement

Image denoising and quality enhancement is a classical computer vision problem that researchers have been trying to solve for decades. This section outlines the various traditional and deep learning-based approaches for image enhancement relevant to this thesis.

2.2.2.1 Traditional methods

Traditional denoising methods are based on the modelling of noise distributions, optimisation schemes or priors of natural images such as sparsity [172], total variation [173], and non-local similarity [174]. NLM [174] uses a weighted average of non-local similar patches for image denoising and BM3D [175] employs structural similarity of patches in a

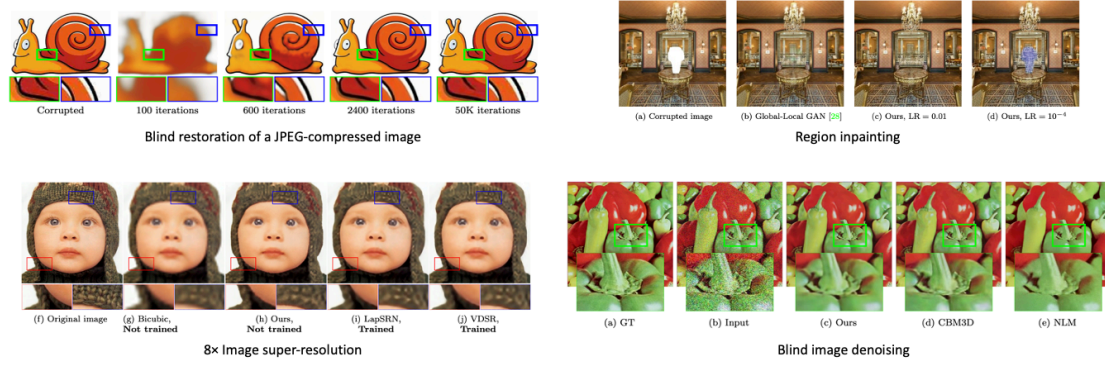


Figure 2.15: Illustration of Deep Image Prior results for diverse image reconstruction problems. Figure courtesy of [187]

transform domain for denoising. Most of the traditional methods for image enhancement work on the basic assumption that an image can be decomposed into several components related to the imaging process and the high-quality image can be obtained by estimating the physical parameters from a single image or by using prior information such as a reference image [176, 177], atmospheric scattering model (ASM) [178, 179], or the Retinex model (RM) [180, 181]. However, these methods often lead to colour deviation, saturation or spurious results as it is very difficult to accurately estimate the physical parameters or find a good reference image [182]. Moreover, most traditional methods have been mostly designed for denoising of natural images only.

2.2.2.2 Learning-based methods

Recently, deep learning-based restoration methods [183, 184] that learn transformation mapping between noisy image and its clean counterpart have outperformed classical denoising methods. LightenNet [185] was proposed to learn a mapping between low-illumination image and the corresponding illumination map for image enhancement. Similarly, DehazeNet [186], an end-to-end model was proposed to recover a haze-free image by estimating the transmission map from an input hazy image. However, such supervised learning-based approaches require high-quality noiseless images during training, thereby rendering them inapplicable for most medical imaging applications.

More recently, there have been several works on unsupervised methods that do not require ground truth and are therefore more useful for medical imaging applications.

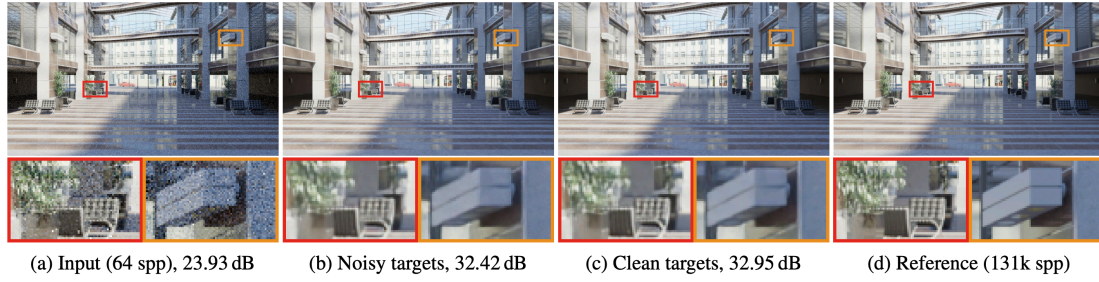


Figure 2.16: Illustration of denoising of a Monte Carlo rendered image using N2N a) input, b) trained using noisy targets, c) trained using clean targets, d) reference image. Figure courtesy of [188]

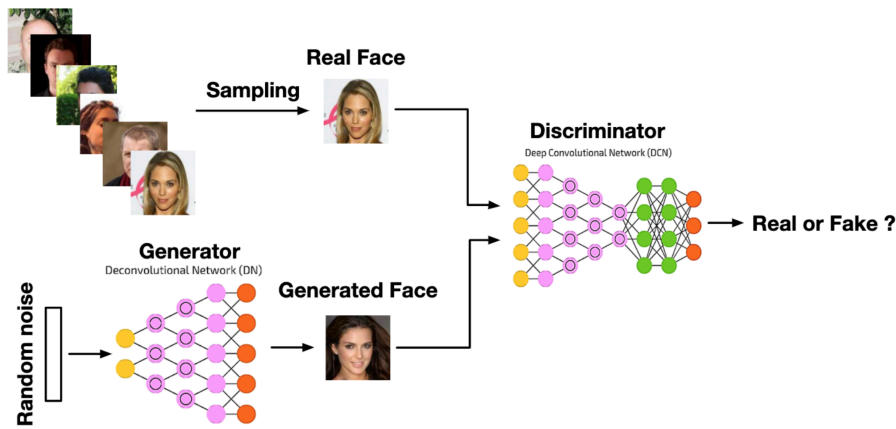


Figure 2.17: Illustration of the GAN architecture consisting of a generator and discriminator that are trained synchronously. Figure courtesy of [189]

One such method is the Deep Image Prior (DIP) [187] that exploits the structure of generator networks and shows that a randomly initialised neural network can be used as a handcrafted prior for restoring image quality. DIP can be employed for various image reconstruction tasks (as shown in Figure 2.15) such as restoration of compressed images, region inpainting, image super-resolution, blind image denoising, and image enhancement. However, DIP only works for less structured artefacts such as additive noise or compression artefacts. Also, DIP needs to be performed separately for each noisy input image, making it computationally expensive and slow. Lehtinen *et al.* [188] proposed a Noise2Noise (N2N) approach to demonstrate that denoising can be obtained by using two independent sets of noise realisations per training sample for various noise models. Although Noise2Noise yields a comparable performance with denoising approaches trained with clean ground truth (see Figure 2.16), it requires two noise realisations per

image which may not be readily available for most medical imaging scenarios. Also, there is no clarification provided on the relationship between noise realisations and the other conditions required for N2N to work well. A self-supervised approach for image denoising, Noise2Void [190] was proposed to exclusively uses only singly noisy images. However, this approach is based on the assumption that noise is pixel-wise independent and that the true intensity of a pixel can be predicted from the local image context. Such assumptions make the applicability of these methods problematic as most medical imaging datasets have complicated noise characteristics [191].

Recently, generative models such as GANs and VAEs have shown impressive image restoration performance for various datasets. GAN architecture consists of two components (See Figure 2.17): a discriminator (D) that learns to distinguish between real and generated (fake) images while the generator (G) learns to create images to fool the discriminator. The two components are trained synchronously during network training. Chen *et al.* [192] proposed a GAN-inspired generative-collaborative framework for smoke removal in laparoscopy images. Similarly, Jiang *et al.* [193] proposed EnlightenGAN that used dual discriminator structure, perceptual loss fusion and attention mechanism for low-light enhancement. CycleGAN [194] inspired several works, such as CycleCBAM [195] that proposed a framework to transform poor quality fundus images to good quality fundus images and CSI-GAN [196] that proposed addition of illumination regularization and structured loss to improve low-quality medical images while preserving structure details. Ali *et al.* [53] also proposed an automatic framework for the detection and segmentation of various artefacts in endoscopy videos followed by a GAN-based model for the restoration of the corrupted frames. However, GAN-based approaches often amplify noise in dark background areas and may produce halo artefacts [196]. Li *et al.* [182] proposed a semi-supervised learning framework that integrated prior physical imaging defect models with the CycleGAN framework for endoscopy image enhancement. However, they only considered limited quality issues and it is often difficult to accurately estimate the parameters of physical imaging models. Although GAN-based generative models produce high-quality output images, they are not suited for representation learning [197, 198]. On the other hand, variational autoencoder

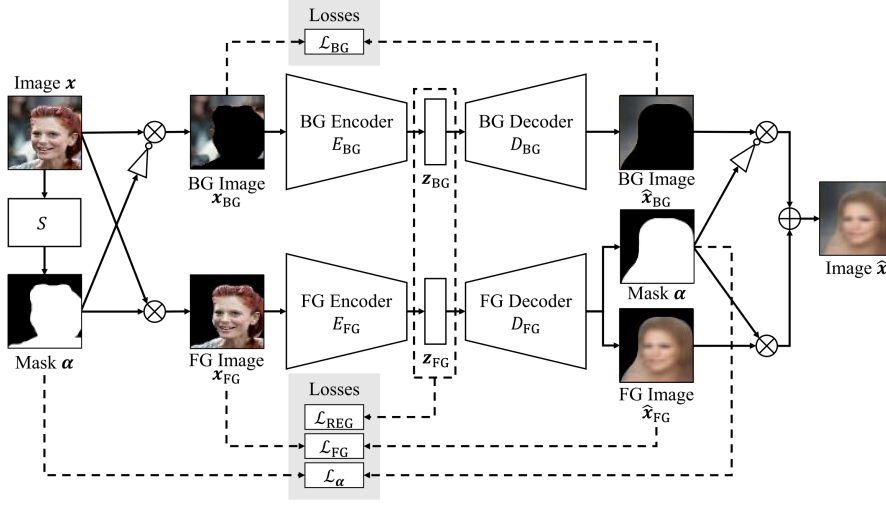


Figure 2.18: VAE-based framework that uses a pretrained segmentation network S to partition the image into two subspaces for object-wise disentanglement. Figure courtesy of [198]

(VAE) [162] has led to a recent wave of generative models as it enables both representation and statistical learning in a single framework [197].

VAE [162] aims at discovering and disentangling factors of variation by forcing independence between various dimensions of latent space while reconstructing the input image. Implicit disentanglement aims to learn disentangled representations by only using the given data and does not use any supervisory information or explicit inductive biases based on the prior knowledge of the target domain [198]. β -VAE [199, 200] is a widely used VAE that encourages disentanglement by increasing the weight of the KL divergence term which thereby increases the independence between various factors of latent space. The loss function for β -VAE can be expressed as:

$$L_{VAE} = -\mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z}|\mathbf{X})} \log p_{\theta}(\mathbf{X}|\mathbf{Z}) + \beta D_{KL}(q_{\phi}(\mathbf{Z}|\mathbf{X}) \parallel p_{\theta}(\mathbf{Z})) \quad (2.5)$$

where β is considered a hyper-parameter such that $\beta > 1$ encourages latent disentanglement by applying a stronger constraint on the latent bottleneck and limiting the representation capacity of latent vector $\mathbf{Z}$. β -TCVAE [201] is an extension of β -VAE with a total correlation loss introduced. Although remarkable results have been obtained by implicit disentanglement methods [199–202], there is no guarantee that the obtained representations are interpretable [203]. Explicit disentanglement involves the explicit separation of latent variables into multiple disjoint groups. Bepler *et al.* [204] proposed a

spatial-VAE framework to explicitly disentangle image rotation and translation from other unstructured latent factors. Skafte *et al.* [205] proposed VITAE framework that separated latent variables into two groups based on appearance and perspective. Parameters of a three-dimensional morphable face model [206] can be assigned to the latent variables for disentanglement of facial images [207]. Similarly, the segmentation of images can be used as an inductive bias to allow groups of latent variables to correspond to various image regions. Instance segmentation of objects can be used as a basis for assigning groups of latent variables [208]. Awiszus *et al.* [209] proposed a framework that used segmentation masks to enforce the disentanglement of various facial attributes into statistically independent sub-spaces. Nakagawa [198] proposed a VAE-based framework that exploited pretrained image segmentation prior and alpha compositing to achieve object-wise disentanglement (see Figure 2.18).

3

Classical computer vision-based image segmentation

Contents

3.1	Introduction	45
3.2	Method	48
3.2.1	Self-similarity descriptors	48
3.2.1.1	Modality independent neighbourhood descriptor	48
3.2.1.2	Normalised correlation transform descriptor	49
3.2.2	Data-term	50
3.2.3	Regularisation	51
3.2.4	Energy minimisation	51
3.2.5	Minimum variance quantisation	53
3.3	Results	54
3.4	Discussion	57
3.5	Conclusion	59

This chapter presents a novel illumination-invariant method for the segmentation of stone fragments in ureteroscopy images. The proposed method computes dense optical flow using the first-order primal-dual approach in a coarse-to-fine strategy enabling the handling of large displacements. To the best of my knowledge, this work is the first optical-flow-based approach towards the segmentation of stone fragments in ureteroscopy datasets.

Contribution/Originality: I independently implemented the proposed framework

in Python with the following novel contributions: (1) used robust correlation transform (CoT) based descriptors as self-similarity metrics for data term, (2) extended the classical two-frame approach to a multi-frame model to deal with small motion and occlusions, (3) normalisation of CoT descriptors to eliminate large deviations and minimise the effect of debris on flow field computations, (4) applied minimum variance quantisation on computed motion field for region aggregation and more pronounced delineation of stone fragments. The clinical dataset used in this study was provided by Dr Ben Turney. I developed the experimental setup for the acquisition of *in vitro* (synthetic data) wherein laser lithotripsy was performed by Dr Louise Goldsmith. I manually annotated the stone fragments in the videos to evaluate the proposed method. This work was selected for oral presentation at SPIE 2020 conference and recognised with an award.

Gupta, Soumya, Sharib Ali, Louise Goldsmith, Ben Turney, and Jens Rittscher. "Motion-induced segmentation of stone fragments in ureteroscopy video." In *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modelling*, vol. 11315, pp. 290-298. SPIE, 2020. **(Award: Runner-up in the Image-Guided Procedures, Robotic Interventions, and Modelling Young Scientist Award at SPIE 2020.)**

3.1 Introduction

Ureteroscopy has evolved into a routine technique for the treatment of kidney stones and diagnosis of the pathology of the upper urinary tract [21]. As mentioned in Chapter 1, automated segmentation in ureteroscopy can help the surgeon in accurate laser targeting that will reduce unnecessary heat buildup from laser lithotripsy and also reduce operative time. In addition, automated segmentation is also the primary step towards the estimation of stone size which is the most important parameter to decide if the fragmentation procedure can be terminated. This will also reduce the chances of any residual fragments and ensure high stone-free rates.

Several methods have been proposed in the literature for automated detection and segmentation of kidney stones in Ultrasound [72–74, 210, 211] and CT [75–79] images. However, unlike radiography, ureteroscopic image quality is deteriorated by the kidney movement due to patient breathing, and the irrigation fluid which is continuously passed

throughout the procedure [18, 212]. A region-growing algorithm for segmentation of renal calculi on ureteroscopic images was proposed [18]. However, such approaches are computationally expensive and require the user to define a set of seed pixels, a similarity criterion, and a stopping criterion. It is difficult to determine this criterion due to the large variability in the types of kidney stones (see Figure 2.1). To the best of my knowledge, not much work has been reported on the segmentation of kidney stones in ureteroscopy images. This is due to the challenges and complexities involved in ureteroscopy data. The automatic segmentation of kidney stones is a complex and challenging task due to significant occlusions from stone debris; dynamic background; high variation in the appearance, size, shape and composition of stone; specular highlights; high motion blur; intra-operative bleeding; and additional image artefacts from turbid fluid inside the target cavity.

Horn and Schunck [89] based dense optical flow approach involves minimising an energy term that consists of a data term and a regularization term. Given, a source image X_s and a target image X_t dense optical flow field $\mathbf{u}$, $\mathbf{u} \in \mathbb{R}^2$ from X_s to X_t is calculated by minimising the energy function $E(\mathbf{u})$ given by:

$$E(\mathbf{u}) = \lambda \psi_{data}(X_s, X_t, \mathbf{u}) + \phi_{reg}(\mathbf{u}) \quad (3.1)$$

where ψ_{data} is the data term that reflects the similarity of the pixels between the two images, ϕ_{reg} is the regularization term that assumes smoothness of the solution $\mathbf{u}$, and λ parameter controls the relative weighting of the data term and the regularization term. The classical data term is based on the brightness constancy assumption (BCA) that does not hold true in real-life scenarios, especially medical imaging. Section 2.2.1.3 presents a comprehensive review of the various data terms and regularisation terms that have been proposed in the literature to improve the original framework proposed by Horn and Schunck [89]. Various data terms based on a combination of different constancy assumptions have been formulated to deal with the illumination variations [91–93]. However, these are not effective in case of complex illumination variations that are present in video sequences [93]. Optical flow methods based on handcrafted image features have been quite successful in handling variances in illumination [83, 84].

Classical coarse-to-fine optical flow formulation helps to deal with large displacements using a multi-scale pyramid strategy [91, 94]. However, the multi-scale strategy may result in the disappearance of small objects or pixels with weakly contrasted textures at coarser levels. To tackle this issue, several large displacement optical flow methods have been proposed [95, 96]. Self-similarity patterns have been very successful in handling large illumination variations, large displacements and motion blur [97, 98]. Data terms based on various descriptors [97, 99, 103, 104, 139, 213] have also been proposed. Normalised correlation transform-based similarity metric has been shown to be more robust to illumination changes for computing deformation fields between histology brain slices and polarised light microscopy images [102]. The spatial regulariser proposed by Horn and Schunck over-smoothes the flow field discontinuities resulting in blurry optical fields. Over the past few years, several optical-flow-driven and image-driven regularisers have been proposed [105–108]. Weighted non-local regularisation methods [97, 109–112] have been proposed to robustly preserve motion discontinuities.

Due to the presence of large illumination variations from view-point changes and non-linear reflections from debris in ureteroscopy datasets, the underlying optical flow computation is largely affected. This work presents a novel multi-frame total-variational optical flow-based approach for segmenting stone fragments in ureteroscopy data. The proposed approach utilises the illumination invariant optical flow framework [213]. To improve the accuracy and robustness of the stone segmentation, the following new terms are introduced: 1) use of robust correlation transform (CoT) based descriptors as self-similarity metric for data-term, 2) extension of the classically used two-frame model to a multi-frame model to deal with very small motion and occlusions, 3) normalisation of CoT descriptors (NCoT) using clipped standard deviation within local neighbourhoods to eliminate large deviations and minimise the effect of debris on flow field computations, and 4) application of minimum variance quantisation on computed motion field for region aggregation to provide more pronounced segmentation of stone fragments. The experiments demonstrate the effectiveness of the proposed method compared to the previous technique [213] on both a two-frame approach and a multi-frame approach.

3.2 Method

The proposed algorithm employs normalised correlation transform (NCoT) based descriptors to formulate the data-term and classical bilateral filtering to enforce regularization of the flow field. The formulated data term ψ_{data} constrained with regularization ϕ_{reg} is then minimised using a first-order primal-dual approach in a coarse-to-fine strategy that allows handling large displacements. This approach is further extended to a multi-frame model that helps to efficiently overcome the effect of debris, and occlusion and also deal with frames having insignificant or no motion at all. The obtained flow field is further processed using minimum variance quantisation to obtain pronounced segmentation boundaries of stone fragments.

3.2.1 Self-similarity descriptors

The concept of self-similarity was first introduced in the domain of image denoising [174]. The approach employs the use of similar image patches across a noisy image to obtain a denoised pixel, which is computed as a weighted average of all other pixels in the image. Self-similarity descriptors are vectors computed for an n -connected neighbourhood window W of radius R and size $(2R+1) \times (2R+1)$ (i.e., with $(2R+1)^2$ pixels and connectivity, $n = ((2R+1)^2 - 1)$, $\forall R \geq 1$) centred at pixel $\mathbf{x}$. Fig. 3.1(a) shows an example where $R=1$, then window W is of size 3×3 and $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_8\}$ are the neighbourhood pixels (represented by black dots) of centre pixel $\mathbf{x}$ (represented as a red dot), where each black dot has an associated patch P_i centred around it. Fig. 3.1(b) represents patch P_3 centred on $\mathbf{x}_3$ with its 8 neighbouring pixels $\mathbf{p}_{\mathbf{x}_3}^i$ (represented by green dots). The descriptors are computed for all pixels in window W using their corresponding patch pixels centred at each pixel. Fig. 3.1(c) illustrates the translation between window W centred on $\mathbf{x}$ in image $X(\mathbf{x})$ and patch P_3 centred on $\mathbf{x}_3$ in $X(\mathbf{x})$.

3.2.1.1 Modality independent neighbourhood descriptor

MIND descriptors [100] can be defined as sum-of-squared differences between pixels in window W centred at $\mathbf{x}$ of an image $X(\mathbf{x})$ and its shifted versions represented by

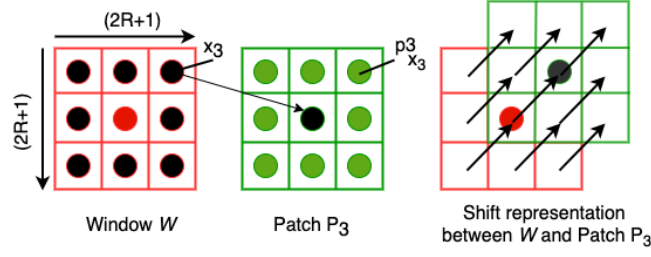


Figure 3.1: a) $\mathbf{x}_i$ represented by black dots ($i \in [1,8]$) are the 8 connected neighbours of the centre pixel $\mathbf{x}$ (represented by red dot) in window W . b) A patch P_i centred at third element $\mathbf{x}_3$ of window W with green dots representing the neighbouring pixels $\mathbf{p}_{\mathbf{x}_3}^i$. c) Illustration of translation between corresponding pixels $\mathbf{x}_i$ and $\mathbf{p}_{\mathbf{x}_3}^i$ ($i \in [1, 8]$) of window W and patch P_3 respectively.

patches P_i centred at each $\mathbf{x}_i$ in image $X(\mathbf{x})$ [213]. Mathematically, similarity measure of window W and patch P_i can then be written as $S_{P_i}(\mathbf{x})$:

$$S_{P_i}(\mathbf{x}) = \int_{j=0}^n (X(\mathbf{x}_j) - X(\mathbf{p}_{\mathbf{x}_i}^j))^2, \quad \text{with } \mathbf{x}_j \in W \text{ and } \mathbf{p}_{\mathbf{x}_i}^j \in P_i \quad (3.2)$$

where $\mathbf{x}_0$ and $\mathbf{p}_{\mathbf{x}_i}^0$ are the centre pixels of a window W in image $X(\mathbf{x})$ and patch P_i of shifted image $X(\mathbf{x}_i)$ respectively. Similarity $S_{P_i}(\mathbf{x})$ can be normalised [174] as:

$$S'_{P_i}(\mathbf{x}) = \exp^{-\frac{S_{P_i}(\mathbf{x})}{\sigma_{\mathbf{x}}^2}}, \quad \text{where } \sigma_{\mathbf{x}}^2 \text{ represents the local variance of image intensities} \quad (3.3)$$

The n -dimensional normalised MIND descriptor vector $\mathbf{ND}(X, \mathbf{x}, n)$ computed for pixel $\mathbf{x}$ in image $X(\mathbf{x})$ can be expressed as :

$$\mathbf{ND}(X, \mathbf{x}, n) = (S'_{P_1}(\mathbf{x}), \dots, S'_{P_i}(\mathbf{x}), \dots, S'_{P_n}(\mathbf{x})) \quad (3.4)$$

3.2.1.2 Normalised correlation transform descriptor

Correlation transforms (CoTs) computed for each pixel in window W using local neighbourhood information can be used to obtain self-similarity neighbourhood descriptors. Computation of CoTs is based on patch mean μ_P and patch standard deviation σ_P of image intensities within a local image region and are therefore invariant to local intensity variations. This standard deviation can be clipped to avoid division by small values. In the present case, this clipping technique also helps to deal with the issue of debris and dynamic background that can significantly affect the quality of segmentation. Post

clipping correlation transform $NCOT_{P_i}(\mathbf{x})$ can be computed for each pixel in window W from a local patch surrounding that pixel and can be expressed as :

$$NCOT_{P_i}(\mathbf{x}) = |X(\mathbf{x}_j) - \mu_{P_i}| / \text{clip}(\sigma_{P_i}, [m_1, m_2]), \quad \text{with } \mathbf{x}_j \in W \quad (3.5)$$

Here, W represents the neighbourhood and $P_i(\mathbf{x})$ denotes shifted version of image patches centred at $\mathbf{x}_i$ in image $X(\mathbf{x})$. μ_{P_i} and σ_{P_i} represent the mean and standard deviation in the image patch P_i . m_1 and m_2 are the minimum and maximum values for which σ_{P_i} is clipped. For this work, I have empirically chosen $m_1 = 0.01$ and $m_2 = 0.5$.

The n-dimensional normalised CoT descriptor vector $\mathbf{ND}(X, \mathbf{x}, n)$ computed for pixel $\mathbf{x}$ in image $X(\mathbf{x})$ can be expressed as :

$$\mathbf{ND}(X, \mathbf{x}, n) = (NCOT_{P_1}(\mathbf{x}), \dots, NCOT_{P_i}(\mathbf{x}), \dots, NCOT_{P_n}(\mathbf{x})) \quad (3.6)$$

3.2.2 Data-term

Neighbourhood descriptors are separately computed for the source X_s and the target X_t images and then used to formulate the data-term for optimisation of flow field $\mathbf{u}$. The coordinates of superimposed pixels are $\mathbf{x}$ in source image X_s and $\mathbf{x} + \mathbf{u}$ in target image X_t , where $\mathbf{u}$ is the displacement vector at $\mathbf{x} \in \mathbb{R}^2, \mathbf{x} = (x, y)$. Data-term is calculated by the sum of absolute differences between corresponding descriptors $\mathbf{ND}(X_s, \mathbf{x}, i)$ and $\mathbf{ND}(X_t, \mathbf{x} + \mathbf{u}_\mathbf{x}, i)$, which is expressed as:

$$\psi'_{data} = \int_{\mathbf{x} \in \Omega} \left\{ \frac{1}{n} \sum_{i=1}^n |\mathbf{ND}(X_s, \mathbf{x}, i) - \mathbf{ND}(X_t, \mathbf{x} + \mathbf{u}_\mathbf{x}, i)| \right\} \quad (3.7)$$

The data term is then linearised using a first-order Taylor series expansion to obtain a convex term and can be expressed as :

$$\psi'_{data} = \int_{\mathbf{x} \in \Omega} \left(\frac{1}{n} \sum_{i=1}^n |\mathbf{ND}(X_s, \mathbf{x}, i) - \mathbf{ND}(X_t, \mathbf{x} + \mathbf{u}_\mathbf{x}^0, i) + \nabla \mathbf{ND}(X_t, \mathbf{x} + \mathbf{u}_\mathbf{x}^0, i)(\mathbf{u}_\mathbf{x} - \mathbf{u}_\mathbf{x}^0)| \right) \quad (3.8)$$

where, $\mathbf{u}_\mathbf{x}^0$ is an approximation of $\mathbf{u}_\mathbf{x}$ and $\nabla = [\frac{\partial}{\partial x}, \frac{\partial}{\partial y}]^T$.

To extend the classical two-frame model to a three-frame approach, two separate data terms are calculated - ψ_{data_1} (corresponding to Frame-1 and Frame-3) and ψ_{data_2}

(corresponding to Frame-2 and Frame-3). The two data-terms are then averaged together to obtain the final data-term ψ_{data} :

$$\psi_{data} = \frac{1}{2} \left[\int_{\mathbf{x} \in \Omega} \left\{ \frac{1}{n} \sum_{i=1}^n |\mathbf{ND}(X_s^{frame-1}, \mathbf{x}, i) - \mathbf{ND}(X_t^{frame-3}, \mathbf{x} + \mathbf{u}_{\mathbf{x}}, i)| \right\} + \int_{\mathbf{x} \in \Omega} \left\{ \frac{1}{n} \sum_{i=1}^n |\mathbf{ND}(X_s^{frame-2}, \mathbf{x}, i) - \mathbf{ND}(X_t^{frame-3}, \mathbf{x} + \mathbf{u}_{\mathbf{x}}, i)| \right\} \right] \quad (3.9)$$

3.2.3 Regularisation

The role of the regulariser in the flow field computation is to enforce smooth optical flow while preserving the flow discontinuities at the edges. Bilateral filtering formalises the similarity between two locations and has been used in various optical flow and image registration methods [97, 110–112, 213]. It helps to reduce the over-smoothing of boundaries by weighting the flow field of pixels based on their colour and spatial distance. A non-local regularization ($\phi_{reg}(\mathbf{u})$) that uses a L1 penalty to allow sharp solutions and preserve motion discontinuities [97, 213] is given by :

$$\phi_{reg}(\mathbf{u}) = \int_{\mathbf{x} \in \Omega} \int_{\forall \mathbf{x}' \neq \mathbf{x} \in W_x} w_{\mathbf{x}}^{\mathbf{x}'} |\mathbf{u}_{\mathbf{x}} - \mathbf{u}_{\mathbf{x}'}|, \quad \text{with } w_{\mathbf{x}}^{\mathbf{x}'} = e^{-\left(\frac{|\mathbf{x} - \mathbf{x}'|^2}{2\gamma_1^2} + \frac{|X(\mathbf{x}) - X(\mathbf{x}')|^2}{2\gamma_2^2}\right)}, \quad (3.10)$$

where (γ_1, γ_2) are normalisation factors and $w_{\mathbf{x}}^{\mathbf{x}'}$ represents the weight assigned to each pixel $\mathbf{x}'$ in the neighbourhood window W centred around pixel $\mathbf{x}$. This weight term is a measure of how likely the pixels $\mathbf{x}$ and $\mathbf{x}'$ belong to the same object. It depends on the spatial distance and the colour distance between the centre pixel $\mathbf{x}$ and the neighbour pixels $\mathbf{x}'$. Value of the weight term ($w_{\mathbf{x}}^{\mathbf{x}'}$) will be small for pixels that are spatially distant ($|\mathbf{x} - \mathbf{x}'|$ large) from each other or those that belong to different regions (large colour difference $|X(\mathbf{x}) - X(\mathbf{x}')|$), thus preventing smoothing of the edge pixels [213].

3.2.4 Energy minimisation

The flow field $\mathbf{u}$ between the source and the target images requires minimisation of the energy term $E(\mathbf{u})$ (see Eq. (3.1)) using primal-dual optimisation approach [214]. Key steps of the primal-dual convex optimisation for the aforementioned energy formulation in Eq. (3.1) are discussed below.

The smoothing error represented by $\phi_{reg}(\mathbf{u})$ in Eq. 3.10 is a convex function in the flow field $\mathbf{u}$. The flow field at each pixel $\mathbf{u}_x$ is evaluated for all other pixels in the neighbourhood W . The spanning of smoothing error to neighbouring pixels is represented by a linear operator K . In order to compute a flow field for all pixels in a given neighbourhood, a linear operator $K: \mathbb{R}^{2 \cdot |\Omega|} \rightarrow \mathbb{R}^{2 \cdot |\Omega| \cdot |W_x|}$ is introduced [213]:

$$K\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 - \mathbf{u}_{1,1} & \cdots & \cdots & \mathbf{u}_{|\Omega|} - \mathbf{u}_{1,1} \\ \vdots & \ddots & & \vdots \\ \vdots & \ddots & & \vdots \\ \mathbf{u}_1 - \mathbf{u}_{m,n} & \cdots & \cdots & \mathbf{u}_{|\Omega|} - \mathbf{u}_{m,n} \end{pmatrix}, \quad (3.11)$$

where $\{\mathbf{u}_{1,1}, \dots, \mathbf{u}_{m,n}\} \in W_x$ represents the flow vectors of neighbourhood pixels surrounding the flow field $\mathbf{u}$ and $\{\mathbf{u}_1, \dots, \mathbf{u}_{|\Omega|}\}$ represents the pixel indexes of the estimated flow field $\mathbf{u}$ itself. The smoothness error after spanning it to neighbouring pixels can be re-written as $E_s(K\mathbf{u})$, expressed as :

$$E_s(K\mathbf{u}) = \sum_{\mathbf{x} \in \Omega} \sum_{\mathbf{x}' \in W_x} w_{\mathbf{x}}^{\mathbf{x}'} |\mathbf{u}_{\mathbf{x}} - \mathbf{u}_{\mathbf{x}'}| \quad (3.12)$$

The optical flow energy term of Eq. 3.1 can now be re-defined as:

$$\min_{\mathbf{u}} \{\lambda \psi_{data}(\mathbf{u}) + \phi_s(K\mathbf{u})\} \quad (3.13)$$

The presence of a linear operator introduces complications and the flow optimisation problem is therefore formulated into a saddle point min-max problem, expressed as:

$$\min_{\mathbf{u}} \max_{\mathbf{w}} \{\lambda \psi_{data}(\mathbf{u}) + \langle K\mathbf{u}, \mathbf{w} \rangle - \phi_{reg}^*(K\mathbf{w})\}, \quad (3.14)$$

where ϕ_{reg}^* is the complex conjugate of ϕ_{reg} , $\mathbf{w} \in \mathbb{R}^{2 \cdot |\Omega| \cdot |W_x|}$ is the dual variable in the closed convex set $\mathbf{S}_w$ which represents the non-negative entries of the bilateral filter weights (mentioned in Eq. (3.10)) and $\langle \cdot \rangle$ represents the dot product. Eq. (3.14) is convex in $\mathbf{u}$ and concave in $\mathbf{w}$. Proximal Point algorithm [215] is used to solve the saddle point problem mentioned in Eq. (3.14). Let τ and α be positive scalars such that $\tau\alpha = \|K^T K\|$ and $(\hat{u}, \hat{w})$ be the initial estimates (usually set to zero). The primal-variable solution is then given by:

$$\tilde{\mathbf{u}}^k = \arg \min_{\mathbf{u} \in \mathbf{S}_u} \left\{ \lambda \psi_{data}(\mathbf{u}) + \frac{\tau}{2} \left\| \mathbf{u} - \left[\mathbf{u}_k - \frac{1}{\tau} K^T \mathbf{w}^k \right] \right\|_2^2 \right\}, \quad (3.15)$$

where S_u is the convex set of $\mathbf{u}$, $\tilde{\mathbf{u}}^k$ is the current estimate, $\mathbf{u}_k = \hat{\mathbf{u}}$ is the initial estimate, $\mathbf{u}$ represents the update values and K^T is the adjoint linear operator. The dual-variable estimate $\tilde{\mathbf{w}}^k$ is given by:

$$\tilde{\mathbf{w}}^k = \arg \min_{\mathbf{w} \in S_w} \left\{ \lambda \phi_{reg}^*(\mathbf{w}) + \frac{\alpha}{2} \left\| \mathbf{w} - \left[\mathbf{w}_k - \frac{1}{\alpha} K(2\tilde{\mathbf{u}}^k - \mathbf{u}_k) \right] \right\|_2^2 \right\}, \quad (3.16)$$

Since Eq: (3.15) and (3.16) are both differentiable, their gradients are zero ($\frac{\partial \tilde{\mathbf{u}}^k}{\partial \mathbf{u}}=0$, $\frac{\partial \tilde{\mathbf{w}}^k}{\partial \mathbf{w}}=0$) at desired points. This gives a set of equations which are then solved to obtain the final solution.

The presented model is implemented in a coarse-to-fine approach to handle large displacements. The minimisation process starts with setting the dual variable equal to zero at the coarsest level. Bi-linear interpolation has been used to build the pyramids with no anti-aliasing in the fine levels and bi-cubic interpolation for upsampling the flow field. The flow field at finer levels of the pyramid is an up-scaled version of the coarser levels. The flow field obtained at each level is used to warp the source image X_s . Eq. (3.8) is used to calculate the data-term at each level of the pyramid and Eq. (3.10) is used to calculate the weights required to perform non-local regularization of the computed flow field in the primal-dual minimisation scheme. Each up-scaling step involves the application of a 3×3 median filter to make the flow field smooth and minimise the interpolation error. The list of algorithm parameters and their corresponding empirically established values are mentioned in Table 3.1. The optimal value of some parameters was different for different images and so has been indicated as an interval in Table 3.1.

3.2.5 Minimum variance quantisation

The flow field obtained by this algorithm is further processed using minimum variance quantisation to reduce the number of colours in the segmentation result and obtain pronounced boundaries of stone fragments. To perform minimum variance quantisation, the number of maximum colours has been specified to be 3 to group the colours of the optical flow image into stone fragments, moving debris and stationary background. Minimum variance quantisation works by dividing the RGB colour cube into smaller boxes of different sizes wherein the size of the boxes depends on the distribution of

Parameters	Values
Weight of data term	[0.85-0.95]
Size of data window	3
Size of non-local propagation window	5
Pyramid factor	[0.6-0.75]
Anti-aliasing start level	[3-7]
Number of warps per level	[2-4]
Number of equation iterations per warp	[8-20]
Down-sampling method	Bi-cubic
Up-sampling method	Bi-cubic
Warping method	Bi-cubic
Normalisation factors (γ_1, γ_2)	5,3

Table 3.1: A list of algorithm parameters and their corresponding empirically established values

Method	Frames	Sim. Metrics	Datasets					Mean
			# 1	# 2	# 3	# 4	# 5	
MIND	2	Dice C.	0.748	0.683	0.652	0.163	0.943	0.638
		Jaccard	0.597	0.518	0.484	0.088	0.892	0.516
		Corr. Coef	0.718	0.568	0.668	0.084	0.913	0.590
	3	Dice C.	0.912	0.804	0.725	0.806	0.923	0.834
		Jaccard	0.838	0.672	0.569	0.675	0.857	0.722
		Corr. Coef	0.893	0.735	0.710	0.788	0.884	0.802
NCoT	2	Dice	0.946	0.819	0.567	0.680	0.907	0.784
		Jaccard	0.898	0.693	0.396	0.515	0.831	0.666
		Corr. Coef	0.935	0.754	0.542	0.675	0.861	0.753
	3	Dice	0.889	0.806	0.902	0.852	0.933	0.876
		Jaccard	0.800	0.675	0.821	0.742	0.875	0.783
		Corr. Coef	0.866	0.760	0.897	0.838	0.905	0.853

Table 3.2: Quantitative evaluation of the motion induced segmentation after minimum variance quantisation

colours in the given image. To improve the accuracy, it allocates more colour-map entries to frequent colours and fewer entries to infrequent colours.

3.3 Results

The proposed algorithm has been evaluated on 4 synthetic and 1 real clinical ureteroscopy dataset. The synthetic dataset was obtained by performing in-vitro lithotripsy of a human kidney stone with a laser and imaged using a clinical ureteroscope. 5 sets of three images

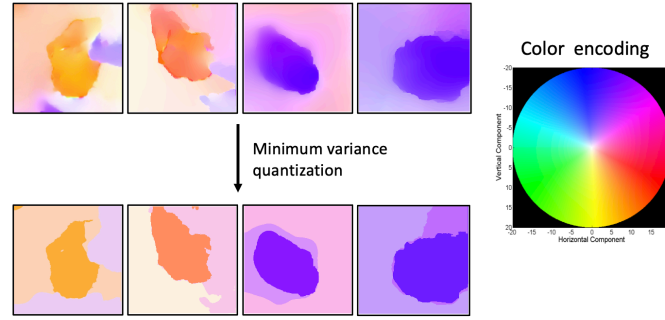


Figure 3.2: Optical flow images and the corresponding results after application of minimum variance quantisation. The saturation and hue in the optical flow images represent the magnitude of displacement and orientation of the fragment respectively. The colour encoding is shown on the right with the horizontal and vertical components of displacement represented along the x-axis and y-axis respectively.

each, were randomly picked from a collection of the mentioned datasets. MIND and CoT descriptors were then evaluated for each test set and analysed quantitatively for 2 frames and 3 frames respectively.

To establish a quantitative evaluation of results, minimum variance quantisation was applied to the optical flow results to obtain pronounced boundaries of the segmented stone fragments as shown in Fig: 3.2. In Table 3.2 it can be observed that the correlation transform-based descriptors give consistent improvements over almost all metric values. It is also evident that taking three frames into account significantly improves the segmentation accuracy. In the case of three frames, the mean Dice and Jaccard are 0.876 and 0.783, respectively for CoT and 0.834 and 0.722, respectively for MIND descriptors. There is also an improvement of over 5% with CoT for the correlation coefficient-based evaluation metric.

For qualitative analysis, optical flow outputs for random samples from one clinical dataset and three synthetic datasets are shown in Fig. 3.3. The saturation and hue in the optical flow images represent the magnitude of displacement and orientation of the fragment respectively. It can be observed that this multi-frame optical flow estimation produces more accurate segmentation of large stone fragments as compared to the two-frame model. The three-frame approach involves the computation of two data terms, thereby making the algorithm more robust to debris and other imaging artefacts like motion blur. The multi-frame approach also aids in improving the segmentation for cases

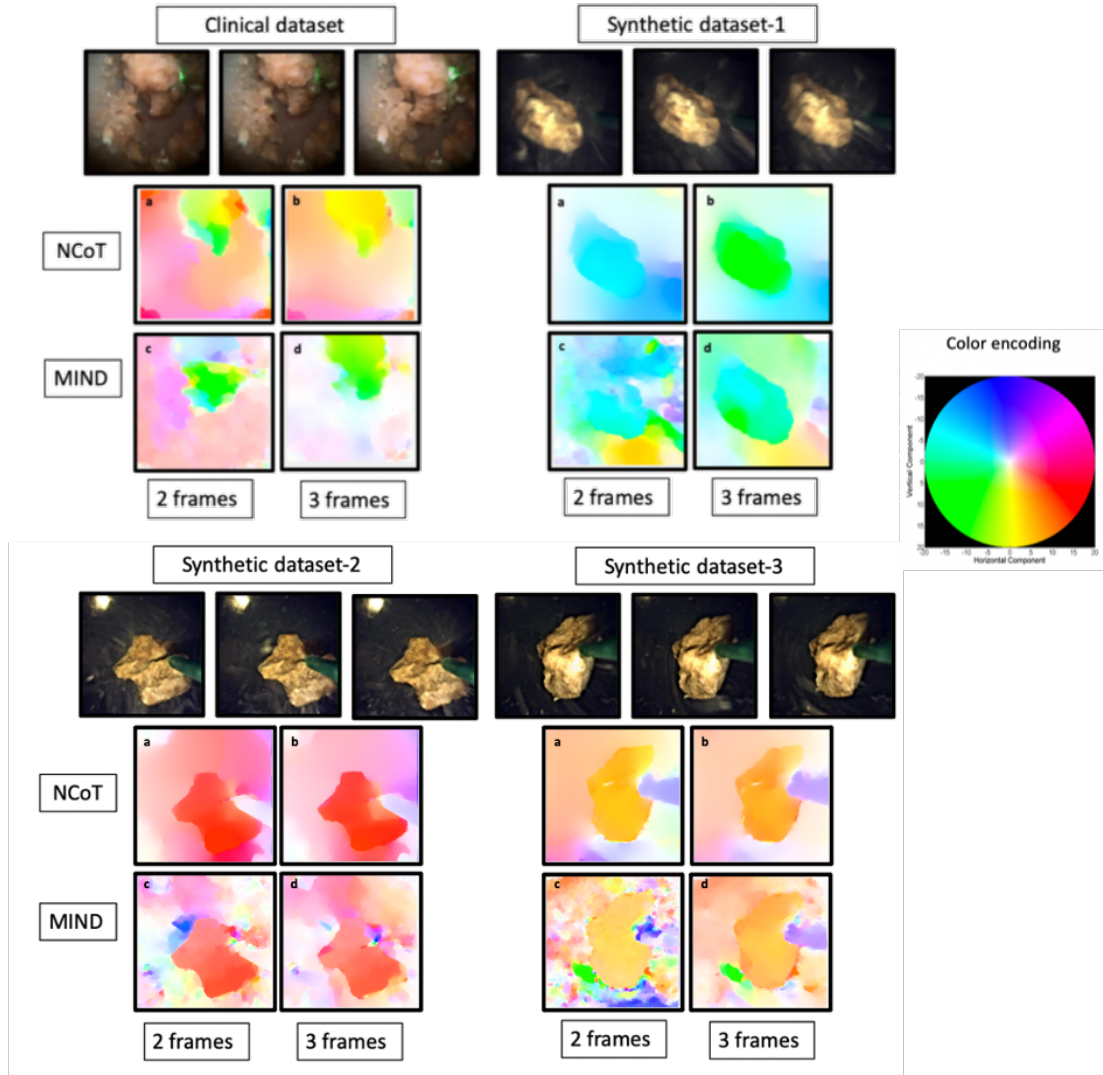


Figure 3.3: Qualitative evaluation of optical flow estimation (before minimum variance quantisation) with NCoT and MIND descriptors using both two-frame and three-frame strategy. The saturation and hue in the optical flow images represent the magnitude of displacement and orientation of the fragment respectively. The colour encoding is shown on the right with the horizontal and vertical components of displacement represented along the x-axis and y-axis respectively.

where there is no significant motion between two consecutive frames. In addition, it can be seen that the use of CoT descriptors enables us to obtain pronounced flow field discontinuities w.r.t. MIND descriptors for the *target* stone. Overall, it is evident that the CoT descriptor delivers consistent performance in overcoming the effect of debris to provide a more accurate segmentation of the stone fragments as compared to the MIND descriptors wherein stone debris is also getting segmented along with the stone.

3.4 Discussion

Segmentation of kidney stones has been well explored for radiology datasets as opposed to ureteroscopy. Here, I have presented a novel illumination-invariant method for the segmentation of stone fragments in ureteroscopy images. As evident from Fig 3.3 and Table 3.2, the proposed algorithm can effectively overcome the challenges of the ureteroscopy dataset and provide an accurate segmentation of stone fragments. However, there are some limitations to this traditional approach for stone segmentation in ureteroscopy.

Fig 3.4 illustrates the optical flow results (after application of minimum variance quantisation) across different parameter values for two samples from the synthetic dataset. It can be seen from Fig 3.4 that the optical flow results change significantly when the values of different parameters are varied. Also, the changes in different parameter values affect the two samples differently. This inconsistency can be mainly attributed to the difference in illumination, motion, stone composition, and amount of debris present in the two samples. It is important to observe here that the results are more sensitive to different parameter values in the case of 'sample 1' as opposed to 'sample 2'. The ureteroscopy dataset has a lot of variability in terms of varying illumination conditions, and stone heterogeneity in terms of shape, texture, colour and position. Therefore, a major limitation of this approach is that the parameter settings have low generalisability and may need to be uniquely established for different kinds of ureteroscopy samples. This indicates that the space for hyper-parameter selection is unique for different kinds of samples and therefore one may often obtain sub-optimal results. Another major limitation of this approach is that the run-time of the proposed algorithm is > 25 secs per image

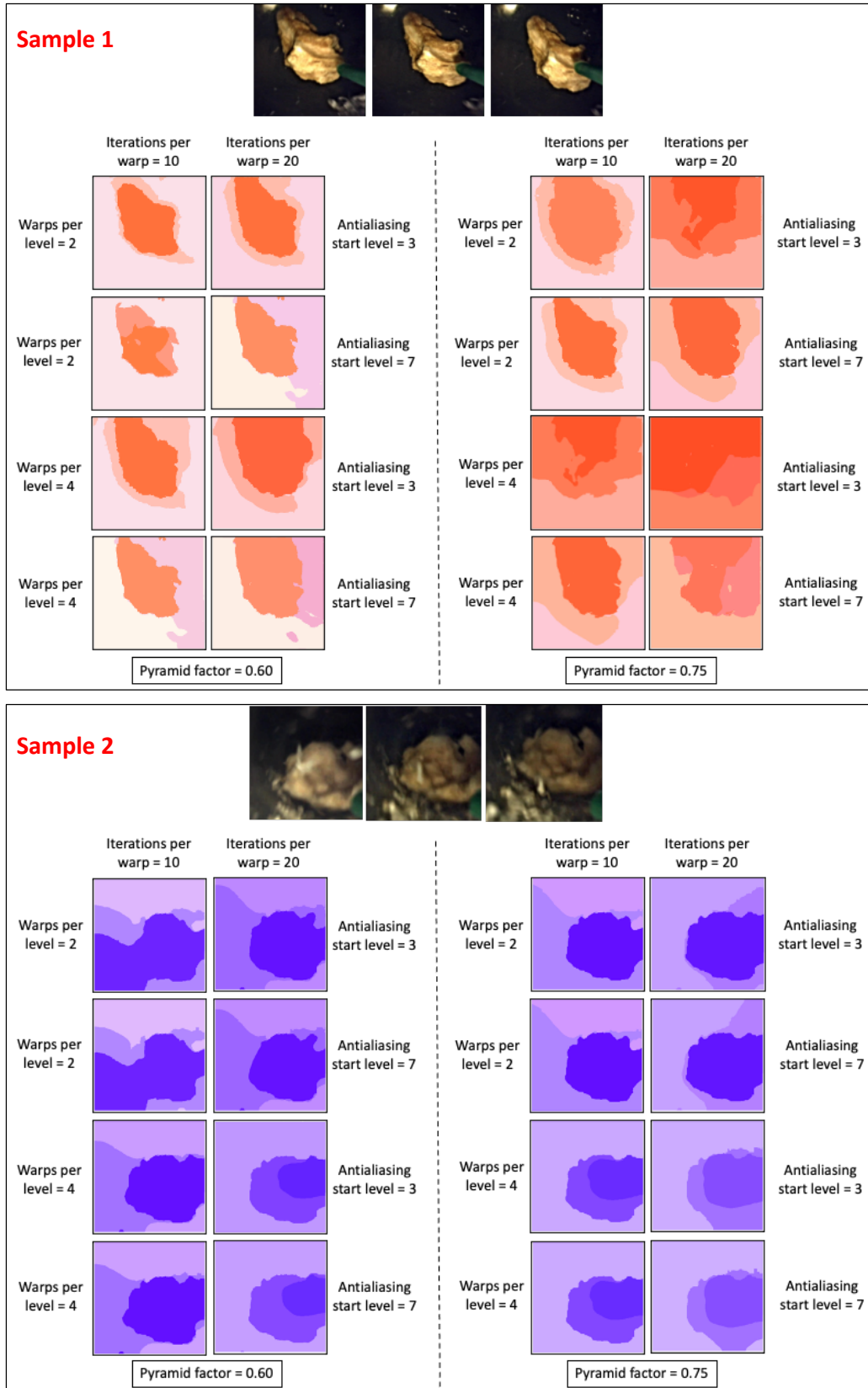


Figure 3.4: Optical flow results (after minimum variance quantisation) obtained using different combinations of values for parameters such as warps per level, iterations per warp, antialiasing start level, and pyramid factor

(depending on the number of warps, iterations and pyramid levels) due to the iterative coarse-to-fine framework. This makes it unsuitable for ureteroscopy applications where segmentation is required in real-time to provide assistance to the surgical team.

3.5 Conclusion

In this work, I have presented a novel illumination-invariant method for the segmentation of stone fragments in ureteroscopy images. To the best of my knowledge, a multi-frame dense optical flow algorithm employing a normalised correlation transform-based data-term in a primal-dual optimisation framework has not been proposed before. The proposed algorithm leverages the motion fields between multiple frames to tackle problems related to illumination variability, blur, occlusions and deformations. Qualitative and quantitative results from my experiments demonstrate that this algorithm can overcome the effect of stone debris and provide accurate segmentation of stone fragments.

The traditional optical flow-based approach presented in this chapter laid the foundation for the segmentation of stone fragments. However, deep learning approaches have been proven very efficient for optical flow estimation as they can overcome the limitations of traditional approaches. The multi-layer architecture of deep learning networks allows them to better extract the intricate multi-scale features and model non-linearities in the images. Deep learning has made it possible to avoid the complex optimisations of the traditional approaches and can run in real-time, thereby making them suitable for medical applications like ureteroscopy and laser lithotripsy. Future work focuses on combining the motion information with deep-learning-based instance segmentation to further improve the segmentation quality and processing speed.

4

Motion-based semantic segmentation using deep learning

Contents

4.1	Introduction	62
4.2	Materials and Method	64
4.2.1	Materials	64
4.2.2	Method	67
4.2.2.1	HybResUNet	68
4.2.2.2	DVFNet	69
4.2.2.3	Loss function	69
4.2.2.4	Proposed framework	72
4.3	Results	72
4.3.1	Experimental setup	74
4.3.1.1	Network implementation	74
4.3.1.2	Independent runs for <i>in vitro</i> and <i>in vivo</i>	74
4.3.1.3	Single vs sequence samples in the test set	75
4.3.2	Data augmentation	76
4.3.3	Loss function	77
4.3.4	Quantitative results	78
4.3.5	Qualitative results	80
4.3.6	Extended out-of-sample test	81
4.3.7	Inference time analysis	83
4.3.8	Ablation study of network design	86
4.4	Discussion	86
4.5	Conclusion	92

Deep learning has made it possible to avoid the complex optimisations of traditional

approaches and can run in real-time, thereby making them suitable for medical applications such as ureteroscopy and laser lithotripsy. This chapter presents the first-ever approach towards multi-class segmentation of ureteroscopy and laser lithotripsy datasets. Such automated segmentation can help the surgical team in accurate laser targeting, thereby reducing the operative time and excess heat generation from laser lithotripsy. In addition, automated segmentation of stone fragments and laser fibre also forms the primary step towards automated quantitative assessments such as the estimation of stone size. The proposed approach effectively makes use of residual connections and motion information between adjacent frames to produce robust and reliable segmentation of stone fragments and laser fibre in real-time.

Contribution/Originality: I independently designed and implemented the proposed framework in Pytorch with the following novel contributions: (1) proposed a novel end-to-end CNN-based learning framework with residual connections that leverages motion between image pairs to overcome challenges such as motion blur, stone debris, and other artefacts, to provide real-time multi-class segmentation of both stone and laser fibre. (2) proposed a compound loss function that outperforms traditional loss functions on ureteroscopy and laser lithotripsy datasets. (3) performed experimental validations on diverse and challenging *in vitro* and *in vivo* ureteroscopy datasets to demonstrate the effectiveness of the proposed multi-class segmentation approach. I recorded the clinical procedures performed by Dr Ben Turney using a frame grabber to formulate the clinical datasets used in this study. I collected the animal datasets from Boston Scientific. I developed the experimental setup for the acquisition of *in vitro* (synthetic data) wherein laser lithotripsy was performed by Dr Louise Goldsmith. I independently annotated the entire dataset used in this study. This work resulted in an oral presentation at IEEE ISBI 2020 and one journal paper.

- **Gupta, Soumya**, Sharib Ali, Louise Goldsmith, Ben Turney, and Jens Rittscher. "MI-UNet: Improved segmentation in ureteroscopy." In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 212-216. IEEE, 2020.

- **Gupta, Soumya, Sharib Ali, Louise Goldsmith, Ben Turney, and Jens Rittscher.**
"Multi-class motion-based semantic segmentation for ureteroscopy and laser lithotripsy."
Computerized Medical Imaging and Graphics(2022): 102112.

4.1 Introduction

Automated segmentation of stone fragments in ureteroscopy and laser lithotripsy can help surgeons improve laser targeting, thereby reducing the extra heat buildup from unnecessary lasering and reducing overall operative time. Additionally, automated segmentation is required to perform quantitative analysis of stone fragments such as stone size which is the most important factor to decide if further fragmentation is required or if the procedure can be terminated. Chapter 3 presented a traditional approach for the segmentation of stone fragments. However, this requires complex optimisation of parameter values and cannot run in real-time. Deep learning-based automated segmentation can overcome the limitations of traditional approaches, thereby making it more suitable for medical applications like ureteroscopy. In addition to the segmentation of stone fragments, segmentation of laser fibre is also important considering that inaccurate laser targeting can result in undesirable excess heat, cause inter-operative bleeding and lead to extended operative times. It can be difficult to visualise the laser fibre in certain frames either due to its transparent tip or additional occlusions from stone debris and turbid fluid. The type of dataset used in this study has not been well-explored in literature and is significantly different and challenging as compared to the other endoscopy datasets. The ureteroscopy video has a small field of view, occlusions from stone debris, intra-operative bleeding, motion blur from unavoidable kidney movements and camera motion, specular highlights, dynamic background, varying illumination conditions, stone heterogeneity, and artefacts from turbid fluid [12, 18]. Sample images from the *in vitro* and *in vivo* datasets are shown in Figure 4.1 wherein the stone debris and blood are highlighted in yellow and red rectangles, respectively.

Automated detection and segmentation of kidney stones have been well explored for Ultrasound (US) [72–74, 210, 211] and Computed tomography (CT) images [75–79]. Rosa *et al.* [18] proposed a region-growing algorithm for the segmentation of renal calculi

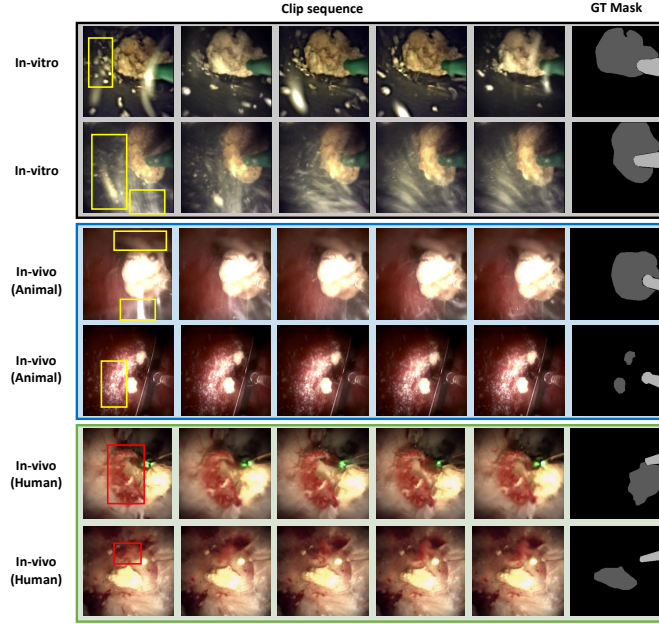


Figure 4.1: Example images for *in vitro*, *in vivo* (animal) and *in vivo* (human) clip sequences with corresponding ground truth masks showing stone and laser fibre. Stone debris and blood are highlighted in yellow and red rectangles (on the left), respectively.

on ureteroscopic images but such approaches are computationally expensive and require the user to define a set of seed pixels, a similarity criterion, and a stopping criterion that is typically difficult to determine due to high variability of kidney stones. To the best of my knowledge, multi-class segmentation that delineates both stones and the laser fibre in ureteroscopy videos has not been addressed before. This chapter proposes a novel end-to-end CNN-based learning framework that uses residual information and motion information for robust segmentation of both stones and laser fibre in both *in vitro* and *in vivo* datasets.

Section 2.2.1.4 presents a comprehensive review of various advancements in encoder-decoder networks such as residual and recurrent connections [113, 123, 128], dilated convolutions [126] and atrous spatial pyramid pooling [41], and attention mechanisms [115]. In this work, I have experimented with various combinations of residual and recurrent connections, ASPP, dilated convolutions and attention gate model to identify the semantic segmentation network with competitive performance. Several studies have shown that the features learned in image registration can be also used to improve the segmentation result [137, 138]. Section 2.2.1.5.1 presents a review of motion-based image registration

approaches. An unsupervised CNN framework that used a bi-cubic Catmull-Rom spline re-sampler to reduce the errors in the re-sampling of deformation fields and a series of deformable convolutional layers to capture complex deformations was proposed for image registration in brain and lung datasets [160]. Inspired by this work, I further improved the segmentation performance by using a sub-network called DVFNet that leverages motion between the adjacent frames to compute end-to-end deformation vector field (DVF) predictions. This motion information is then used to prune the segmentation map obtained from the semantic segmentation network, resulting in a context-aware edge-enhanced multi-class segmentation. In this work, the proposed method has been compared with different state-of-the-art (SOTA) methods to exhibit its competitiveness on both *in vitro* and *in vivo* datasets. An extended out-of-sample study has also been presented on unseen animal data and patient data to measure the robustness compared to the SOTA methods. Finally, an extensive ablation study is provided that validates the network choices and data augmentation strategies.

4.2 Materials and Method

This section presents a description of the dataset which is used in this study and details the proposed framework for multi-class segmentation in ureteroscopy and laser lithotripsy.

4.2.1 Materials

Sub-sequences containing intense lasering and stone movement were extracted and clip sequences containing 5 adjacent frames of these sub-sequences were used. One frame from each clip sequence (1 out of 5) was manually labelled using the following three class labels: stone fragments, laser fibre, and background. In Figure 4.1, *in vitro*, *in vivo* (animal), and *in vivo* (human) samples have been highlighted in gray, blue, and red boxes, respectively. The VGG Image Annotator (VIA) tool [216] was used to obtain a ground truth mask for each clip sequence as shown in Figure 4.1. The annotations used in this study were performed by me and independently verified by two experts, a senior research associate with over 5 years of experience in endoscopy and a senior

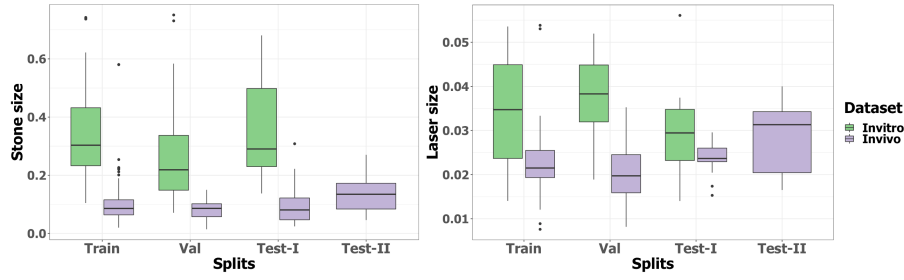


Figure 4.2: Box-plots showing the size distribution of stone and laser class across training, validation, and test sets in *in vitro* and *in vivo* datasets. It can be seen that the size distribution of kidney stones is different across train, validation, and test sets, indicating that they come from different videos. The large variability between *in vitro* and *in vivo* datasets is also clearly evident, explaining why experiments were run independently for the *in vitro* and *in vivo* datasets.

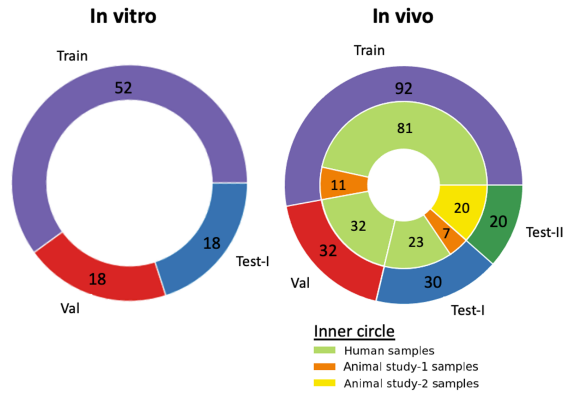


Figure 4.3: Pie charts showing the division of *in vitro* and *in vivo* datasets. The proportion of both human and animal samples has also been illustrated for the *in vivo* dataset

urologist. A few samples did have disagreements but the senior urologist's decision was considered to be final in case of any disagreement.

The division of both *in vitro* and *in vivo* datasets into training, validation and test sets has been illustrated in Figure 4.3. The *in vitro* dataset consists of 52, 18 and 18 clip sequences in train, validation and test splits, respectively while the *in vivo* consists of data from 4 subjects with 92, 32, and 30 clip sequences in the train, validation, and Test-I splits, respectively. As shown in Figure 4.3, 13 % of the combined train, validation, and test-I set consist of ureteroscopy images obtained from animal studies performed at Boston Scientific, and 87% consists of clinical ureteroscopy images collected at the Oxford University Hospitals. In the case of *in vivo*, an extra dataset (Test-II) that consists of 20 unseen new samples is also included and is used in the final part of this study to validate the proposed model as compared to the existing SOTA approaches. Test-II consists of

unseen ureteroscopy images obtained from another set of animal studies performed at Boston Scientific. As can be seen from Figure 4.3, the models were trained on mainly human ureteroscopy samples while the Test-II set consists of images of ureteroscopy performed on animal subjects. In addition, hardware settings such as LED illumination and software configuration settings used at the time of surgery and image acquisition are different for the animal studies in Test-I and Test-II datasets. This indicates that the trained models do not cover all the possible variance scenarios arising from different image acquisition settings, resulting in a performance decrease in the Test-II dataset.

An overview of the dataset configuration has also been shown in Table 4.1. For a better understanding of the datasets, box plots showing the relative size distribution of stone and laser class across training, validation, and test sets in the *in vitro* and *in vivo* datasets, respectively, have been shown in Figure 4.2. Here, the size distribution refers to the ratio of the object size in pixels to the size of the image (256×256). As evident from Table 4.1 and Figure 4.2, the high standard deviation of the stone indicates high variability of kidney stones in each dataset. It can also be seen that the mean of stone is different for train, validation, and test sets, indicating that they come from different videos.

Patient-specific features can be very important and unique in certain image modalities. However, in the case of ureteroscopy, images from the same patient can have a lot of variability in terms of tissue appearance, illumination conditions, stone variability, and different viewpoints with respect to the camera. These factors together create a dynamic scene in every frame. Figure 4.8 illustrates random image frames from three patients with observable Intra and inter-patient variability across data. Therefore, the dataset was randomly split on sample level into the train (60%), validation (20%), and test (20%) for both *in vivo* and *in vitro* data. To further justify the experiments and demonstrate that no data leak has occurred in the test samples, we: i) evaluated model performance on 18 samples obtained from a separate unseen test patient data (see patient 3 in Figure 4.8) that was not included in the training and held-out test dataset, and ii) performed patient wise 4-fold cross-validation on the *in vivo* data (see Table 4.12).

The *in vitro* dataset supports preliminary investigation using different stone shapes, sizes, compositions, and debris levels. This allowed us to understand the performance

Dataset	Split	Number of samples		Distribution ($\mu \pm \sigma$)	
		Single	Sequence ($\times 5$)	Stone	Laser
<i>In vitro</i>	Train	52	260	0.3570 ± 0.1702	0.0345 ± 0.0107
	Validation	18	90	0.3517 ± 0.1635	0.0295 ± 0.0099
	Test-I	18	–	0.2858 ± 0.1975	0.0369 ± 0.0101
<i>In vivo</i>	Train	92	460	0.1084 ± 0.0865	0.0222 ± 0.0067
	Validation	32	160	0.0836 ± 0.0337	0.0205 ± 0.0069
	Test-I	30	–	0.0949 ± 0.0628	0.0240 ± 0.0033
	Test-II	20	–	0.1384 ± 0.0566	0.0277 ± 0.0076

Table 4.1: Overview of the train, validation, and test splits for the *in vitro* and *in vivo* experiments. Mean (μ) and standard deviation (σ) for each stone and laser annotations in each split are also provided. Here, Test-I refers to the split samples from the curated dataset as 20% split and Test-II refers to the unseen animal samples.

of segmentation methods and their ability to differentiate between background, stone, and instrument under controlled settings. Clinical human kidney stones were used for *in vitro* experiments to mimic more realistic imaging conditions. As can be seen from Figure 4.2 and t-SNE plot in Figure 4.6, there is a large variability between the *in vitro* and *in vivo* datasets. Therefore, independent experiments for *in vitro* and *in vivo* datasets were conducted in this work. To further justify the approach and understand the role of *in vitro data*, I trained the HybResUNet model with the *in vitro* and *in vivo* datasets: trained separately and trained together (see Table 4.3 and Figure 4.5).

4.2.2 Method

The proposed framework utilises two sub-networks: HybResUNet, which is a hybrid version of a residual UNet that uses residual connections in the encoder path only; and a DVFNet that leverages motion between adjacent frames to compute end-to-end deformation vector fields. This motion information obtained from a DVFNet is used to prune the segmentation mask obtained from the semantic segmentation network resulting in an improved multi-class segmentation. The entire framework is designed as an end-to-end CNN model that optimises the proposed compound loss function. In this section, the semantic segmentation network module, DVFNet, and the compound loss function are presented.

4.2.2.1 HybResUNet

Base network. The first sub-network of the segmentation framework is an encoder-decoder-based network called HybResUNet [123]. It is basically a 9-level deep UNet architecture wherein residual blocks are used instead of traditional feed-forward units in the contracting (encoder) path as shown in Figure 4.4. This is because the replacement of all feed-forward units in a conventional encoder-decoder network with residual blocks increases the network complexity and tends to overfit the training data [123]. Each of the four residual units in the encoder consists of repeated application of two 3×3 convolutions, followed by a Batch Normalisation (BN) and a Rectified Linear Unit (ReLU). These convolutions are followed by the addition of output to its initial input as residual units and a 2×2 max pooling operation with a stride of 2 for downsampling. The decoder in the HybResUNet uses transposed and regular convolutions to gradually increase the image size while reducing the number of features. The network also consists of skip connections that circumvent the information loss during downsampling by concatenating the output obtained from each residual block with the output of transposed convolution from the up-scaled features at the decoder layers. Each of these concatenations is further followed by the sequential application of two regular convolutions. For a fair comparison, other networks namely UNet [40], DeepResUNet [122], and R2-UNet [113] are all implemented as 9 levels deep architecture to obtain the best performing base network for the dataset.

Dilations, ASPP, and Attention gate. To further improve the base network, I tried incorporating dilated convolutions, Atrous Spatial Pyramid Pooling (ASPP), and attention gate mechanisms [115]. Replacing conventional convolutions in CNN models with dilated convolutions is known to improve the aggregation of multi-scale contextual information without losing resolution [114, 126, 127]. Hamaguchi *et al.* [126] introduced a novel architecture for the segmentation of small object instances in remote satellite imagery by first increasing the dilation factors and then decreasing them. Inspired by this idea, a series of dilation rates that worked best for the data was empirically obtained: [1, 2, 3, 4, 3, 2, 1, 2, 1]. These networks also consist of an Atrous Spatial Pyramid Pooling (ASPP) module at the end of the encoding path with an output stride (ratio of the input image size to the output feature map size) of 16 and six parallel convolutions with dilation

rates [1, 2, 4, 8, 16, 32]. To sum up, the experiments labelled with ASPP in Table 4.8, Table 4.9 and Table 4.14 comprise a series of dilated convolutions and an ASPP module at the end of the encoding path. To leverage the attention gate mechanism [115], attention gates were added to the skip connections just before the concatenation operation (refer to Att in Table 4.8, Table 4.9 and Table 4.14). Such a mechanism suppresses the propagation of irrelevant and noisy responses in the network. I further tried to incorporate ASPP, series of dilated convolutions, and attention gate all together to observe if this improves the segmentation accuracy (refer to Att-ASPP in Table 4.14).

4.2.2.2 DVNet

Building on the work of [160], the proposed DVNet is also based on an encoder-decoder architecture where the parameters of the spline re-sampler are learned from training data. It consists of a total of 12 layers, which include three linear convolutional layers, two average pooling layers, and two deformable convolutional layers in the encoder. Each of these convolutional layers is combined with Batch Normalisation (BN) and exponential linear unit (ELU) as shown in Figure 4.4. Although ELU is slower to compute as compared to ReLU, its training time is compensated by its faster convergence as it pushes the mean of the activation function closer to zero. The decoder layer consists of a Catmull-Rom spline resampler to rescale the obtained DVF from the encoder, which is then further resampled with two additional layers that consist of a convolution layer, an ELU activation, and the spline resampler. The final DVF is then applied to the original image to obtain the corresponding warped image X_{warp} .

4.2.2.3 Loss function

Although commonly used, the cross-entropy loss does not differentiate between correctly classified and misclassified samples causing the easily classified negatives to compromise the majority of the loss and dominate the gradient [217]. Focal loss (FL), which is an improved version of CE loss was introduced by [217]. The focal Loss is defined as:

$$L_{FL} = -(1 - p_y)^\gamma \log(p_y), \quad (4.1)$$

where $\gamma \geq 0$ is the tunable focusing parameter that smoothly adjusts the rate at which easy examples are down-weighted, p is the probability. FL is equivalent to CE when $\gamma=0$. Bokhovkin *et al.* [218] proposed a novel loss function that is essentially a differentiable surrogate of a metric accounting accuracy of boundary detection. Let's say $\mathbf{y}$ and $\mathbf{y}^{gt}$ represent the binary map predicted by a neural network and ground truth map, respectively for arbitrary class c for an image. The boundaries $\mathbf{y}^{gtb}$ and $\mathbf{y}^b$ can then be defined as:

$$\begin{aligned}\mathbf{y}^{gtb} &= pool(1 - \mathbf{y}^{gt}, \theta_o) - (1 - \mathbf{y}^{gt}) \quad \text{and} \\ \mathbf{y}^b &= pool(1 - \mathbf{y}, \theta_o) - (1 - \mathbf{y}),\end{aligned}\tag{4.2}$$

where $(1-\mathbf{y}_{gt,pl})$ refers to the inversion of any pixel of the map and $pool(\cdot, \cdot)$ denotes a pixel-wise max-pooling operation to the inverted binary map with a sliding window of size, θ_o set to 3. The Euclidean distances between pixels to boundaries require computation of a supporting map which is the map of extended boundary given by $\mathbf{y}^{gtb,ext} = pool(\mathbf{y}^{gtb}, \theta_\Delta)$ and $\mathbf{y}^{b,ext} = pool(\mathbf{y}^b, \theta_\Delta)$, where θ_Δ set to 5.

The precision P^c and recall R^c can then be defined as:

$$P^c = \frac{sum(\mathbf{y}^b \circ \mathbf{y}^{gtb,ext})}{sum(\mathbf{y}^b)}\tag{4.3}$$

$$R^c = \frac{sum(\mathbf{y}^{gtb} \circ \mathbf{y}^{b,ext})}{sum(\mathbf{y}^{gtb})},\tag{4.4}$$

where ‘ $\circ$ ’ denotes the pixel-wise multiplication of two binary maps and $sum(\cdot)$ refers to the pixel-wise summation of a binary map. The reconstructed boundary metric, B^c is averaged over all classes and is then used to formulate the loss function, $L_{boundary}$ that can be defined as:

$$L_{boundary} = 1 - B^c \text{ with boundary metric } B^c = \frac{2P^c R^c}{P^c + R^c}\tag{4.5}$$

where P^c and R^c refer to the precision and recall. [218] performed a comparative analysis of their proposed boundary loss with various loss functions such as IoU loss and Dice loss. In the first part of this study where non-sequence data is used, I aim to find the best-performing baseline network by using a compound loss function that combines this

boundary loss with SOTA Focal loss. For the second part of this study where I have integrated DVFNet for motion estimation, I have proposed to use an extended novel compound loss function which is a combination of Focal loss, Boundary loss, Cross-correlation loss between warped and target image, and smoothness constraint [151] on the predicted deformation fields. The cross-correlation loss is computed between the warped images ($X_{\text{warp}_{i \leftarrow (i+2)}}$ and $X_{\text{warp}_{(i+2) \leftarrow (i+4)}}$) with their corresponding source images (X_i and $X_{(i+2)}$), respectively and is given by:

$$L_{sim} = L_{sim1}(X_i, X_{\text{warp}_{i \leftarrow (i+2)}}) + L_{sim2}(X_{(i+2)}, X_{\text{warp}_{(i+2) \leftarrow (i+4)}}) \quad (4.6)$$

i.e.,

$$L_{sim} = \frac{1}{2N} \sum \left(\frac{X_i(x) - \mu_1}{\sqrt{\sigma_1^2 + \epsilon^2}} - \frac{X_{\text{warp}_{i \leftarrow (i+2)}}(x) - \mu_{\text{warp}_{i \leftarrow (i+2)}}}{\sqrt{\sigma_{\text{warp}_{i \leftarrow (i+2)}}^2 + \epsilon^2}} \right)^2 + \frac{1}{2N} \sum \left(\frac{X_{(i+2)}(x) - \mu_2}{\sqrt{\sigma_2^2 + \epsilon^2}} - \frac{X_{\text{warp}_{(i+2) \leftarrow (i+4)}}(x) - \mu_{\text{warp}_{(i+2) \leftarrow (i+4)}}}{\sqrt{\sigma_{\text{warp}_{(i+2) \leftarrow (i+4)}}^2 + \epsilon^2}} \right)^2, \quad (4.7)$$

where μ and σ are the mean and standard deviation, N is the total number of pixels and $\epsilon = 10^{-3}$ to avoid division by zero. The estimated deformation vector fields (DVF) can be locally smoothed using a smoothness constraint on its spatial gradients. Further, the smoothness constraint on the estimated deformation vector fields ($DVF_{i \leftarrow (i+2)}$ and $DVF_{(i+2) \leftarrow (i+4)}$) can be expressed as:

$$L_{smo} = \sum (\|\nabla DVF_{i \leftarrow (i+2)}\|_2^2 + \|\nabla DVF_{(i+2) \leftarrow (i+4)}\|_2^2), \quad (4.8)$$

where ∇ indicates the gradient of flow fields and $\|\cdot\|$ indicates its L_2 norm.

Finally, the proposed compound loss function can be formulated as:

$$L = L_{FL} + \kappa L_{boundary} + \xi L_{sim} + \zeta L_{smo}, \quad (4.9)$$

where κ , ξ , and ζ are the hyper-parameters used to balance the contribution from the boundary loss, similarity loss and smoothness loss, respectively with initial values set to 1, 0.5 and 1, respectively.

4.2.2.4 Proposed framework

The proposed framework is shown in Figure 4.4. Each clip sequence comprises five images, X_i to $X_{(i+4)}$ wherein image pairs $(X_i, X_{(i+2)})$ and $(X_{(i+2)}, X_{(i+4)})$, respectively, are provided as an input in grayscale format to different DVFNet networks during training. As shown in Figure 4.4, $X_{(i+1)}$ and $X_{(i+3)}$ are skipped over in order to have significant stone and debris movement across images fed into DVFNet. Each DVFNet computes a Deformation Vector Field (DVF) map ($DVF_{i \leftarrow (i+2)}$ and $DVF_{(i+2) \leftarrow (i+4)}$) and their corresponding warped image ($X_{warp_{i \leftarrow (i+2)}}$ and $X_{warp_{(i+2) \leftarrow (i+4)}}$). The obtained DVFs are locally smoothed a smoothness constraint L_{smo} on its spatial gradients is introduced. Further, Normalised Cross Correlation (NCC) is used as a similarity metric to minimise shape differences between the obtained warped images ($X_{warp_{i \leftarrow (i+2)}}$ and $X_{warp_{(i+2) \leftarrow (i+4)}}$) and their corresponding source images X_i and $X_{(i+2)}$, respectively, as shown in the Figure 4.4 ($Lsim^1$ and $Lsim^2$ correspond to these losses).

As shown in Figure 4.4, mean of the warped images ($X_{warp_{i \leftarrow (i+2)}}$ and $X_{warp_{(i+2) \leftarrow (i+4)}}$) is fed to the HybResUNet network to obtain the first semantic map, $\mathbf{y}_i^1$. A second semantic map, $\mathbf{y}_i^2$ is obtained by using the fifth input image $X_{(i+4)}$ in the RGB format to another HybResUNet. Finally, the two semantic maps are averaged to obtain a final map, $\mathbf{y}_i$. The network then optimises the final output semantic map by minimising a combined loss function represented in Eq. (4.9). The DVFNet part of the framework is only used during network training (indicated by solid path in Figure 4.4) while the learned weights of the HybResUNet are used during frame-wise inference (indicated by dotted path in Figure 4.4).

Table 4.2 presents the number of trainable parameters for different networks discussed in this paper.

4.3 Results

This section first outlines the experimental set-up, a description of the data augmentation strategy and an objective comparison of the proposed compound loss function against the SOTA focal loss [217]. Subsequently, quantitative and qualitative results of the proposed

Network	# Trainable parameters
UNet [40]	31,390,851
HybResUNet [123]	31,564,995
DeepResUNet [122]	32,613,193
R2-UNet [113]	39,091,523
HybResUNet +DVFNet [†]	31,916,421
ASPP-HybResUNet+DVFNet [†]	59,194,757
Att-HybResUNet+DVFNet [†]	32,442,033
Att-ASPP-HybResUNet+DVFNet [†]	59,720,369

Table 4.2: Number of trainable parameters for different CNN networks explored in this work

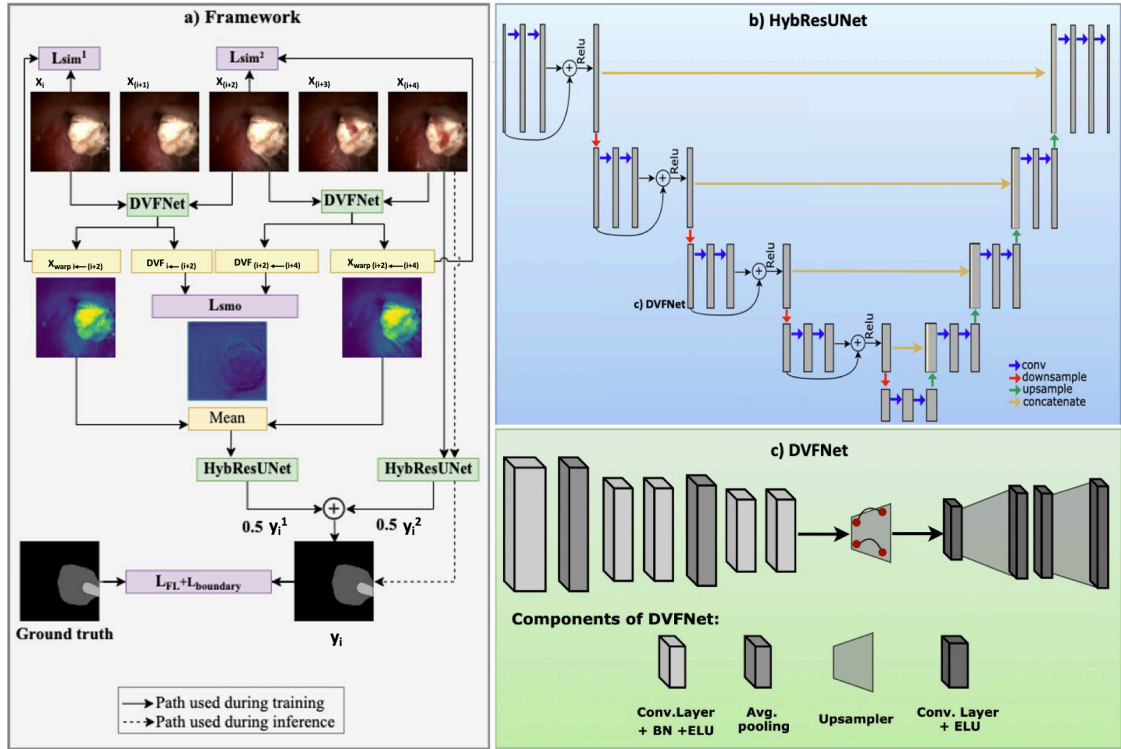


Figure 4.4: Proposed framework: a) Overall proposed semantic segmentation framework illustrating both training and inference paths presented in this study. Here, DVFNet utilising image pairs is only used for training shown by solid path while the learned weights of HybResUNet are used during frame-wise inference indicated by the dotted line. b) Network representation of the HybResUNet used in this work. It is a 9-level deep UNet architecture with residual connections in the encoder path and c) an encoder-decoder DVFNet architecture used to compute deformation vector fields (DVF) between image pairs that are used to prune the segmentation results obtained from the HybResUNet.

networks against existing SOTA networks on the laser lithotripsy and ureteroscopy datasets (both *in vitro* and *in vivo*) are presented. Then, an extended out-of-sample test of the proposed framework is provided that also includes cross-validation results of models trained on the patient-wise split in *in vivo* data. This is followed by a quantitative comparison of the inference time of the networks and finally, ablation studies for network design are presented.

4.3.1 Experimental setup

This section presents network implementation details, the rationale behind not merging *in vitro* and *in vivo* datasets and the justification behind not using sequences during network inference.

4.3.1.1 Network implementation

All image samples were each resized to 256×256 pixels in a 3-channel RGB format to train the deep learning models (original image size depending on the acquisition settings of the source ranging from 394×392 to 1080×1080). Networks were trained with a batch size of 2 on NVIDIA Quadro RTX 6000 for 100 epochs using Adam optimiser with a learning rate of $1e^{-3}$, initial decay rates were set to default 0.9 and 0.999 for estimation of the first and second moments of gradient respectively, with validation performed after every epoch.

4.3.1.2 Independent runs for *in vitro* and *in vivo*

Initially, the model was trained on a combination of both *in vitro* and *in vivo* datasets. Table 4.3 and Figure 4.5 present a quantitative and qualitative comparison between two scenarios where *in vitro* and *in vivo*: i) were trained separately and ii) were trained together with HybResUNet. The performance decrease in the case of joint training can be attributed to the significant variability between the *in vitro* and *in vivo* datasets. Figure 4.6 shows a t-SNE plot for better visualisation of the variability between *in vitro* and *in vivo* datasets. The study then proceeded with independent runs for each of them.

Training strategy	DSC			
	<i>In vitro</i>		<i>In vivo</i>	
	Stone	Laser	Stone	Laser
Independent	0.8838	0.8724	0.8339	0.8241
Joint	0.8192	0.9222	0.8027	0.8095

Table 4.3: Quantitative comparison of independent runs against joint training of *in vitro* and *in vivo* datasets

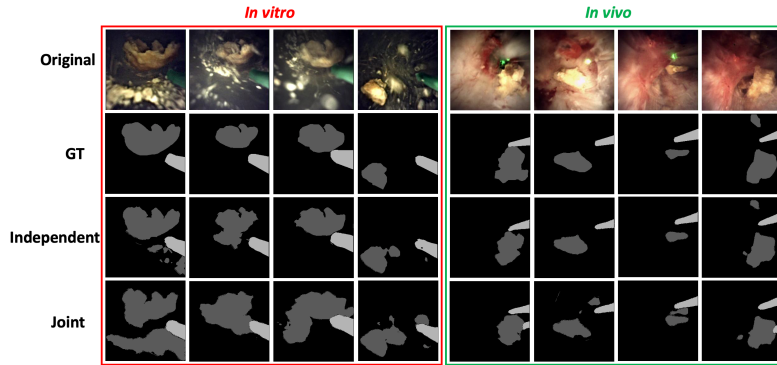


Figure 4.5: Qualitative comparison of independent runs against joint training of *in vitro* and *in vivo* datasets

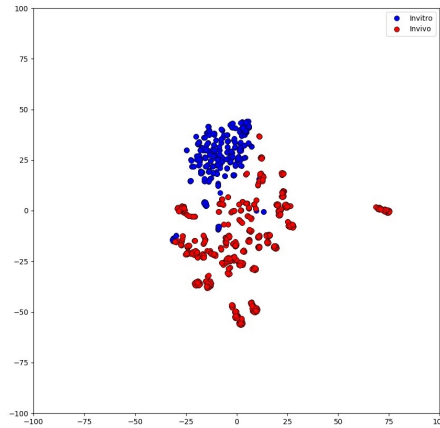


Figure 4.6: t-SNE plot of features extracted from *in vitro* (shown in blue) and *in vivo* (shown in red) samples. The clear separation in their distribution indicates the variability between the *in vitro* and *in vivo* datasets

4.3.1.3 Single vs sequence samples in the test set

Table 4.4 below presents a Dice score-based comparison with their corresponding computation time. Here, one can observe that using sequence samples does not show significant improvement on Dice, however, inference time exceeds 6 times that when

computed with a single frame. This indicates that the weights of HybResUNet updated during training using sequence samples are sufficient to provide good performance in the test set. I have therefore not used sequence samples in the test dataset.

Sample	Network (during inference)	DSC				Compute time(s)
		<i>In vitro</i>		<i>In vivo</i>		
		Stone	Laser	Stone	Laser	
Single	HybResUNet	0.9068 ±0.09	0.8702±0.14	0.8203±0.14	0.8568±0.21	0.86
Sequence	HybResUNet+DVFNet	0.9067±0.10	0.8731 ±0.16	0.8221 ±0.14	0.8603 ±0.21	5.23

Table 4.4: Comparison of dice score and computation time (for 10 samples on NVIDIA Quadro RTX 6000) for HybResUNet+DVFNet applied on sequence data vs single sample data in the test dataset

4.3.2 Data augmentation

Data augmentation techniques such as flips, random crops, rotate and color jittering has been commonly used for deep learning in medical imaging. In this study, I first intend to determine the optimal augmentation choices by studying the effect of different strategies on segmentation accuracy. Initially, I performed 8 training experiments for both *in vitro* and *in vivo* datasets, where each experiment involved training HybResUNet on various augmentation types. The list of augmentation techniques with their corresponding settings used in this study is mentioned below in Table 4.5. Table 4.6 below presents the ablation study associated with determining the optimal data augmentation strategy for our dataset. It can be seen that the *Random Brightness Contrast(RBC)* and *Equalize* transformation improve the segmentation accuracy in the case of *in vitro* dataset as compared to no augmentation scenario. On the other hand, *RBC* and *CLAHE* both provide a higher DSC compared to no augmentation in the case of the *in vivo* dataset. The dice similarity coefficient (DSC) was recorded for each experiment for the stone and laser class (refer to Table 4.6). It can be seen that the *Random Brightness Contrast(RBC)* and *Equalize* transformation improve the segmentation accuracy in the *in vitro* datasets as compared to the no augmentation scenario. On the other hand, *RBC* and *Contrast Limited Adaptive Histogram Equalization (CLAHE)* both provide a higher DSC compared to no augmentation in the *in vivo* datasets. The difference in the results of the *in vitro* and

in vivo can be majorly attributed to the difference in background appearance between the two datasets (refer to Figure 4.1). Further experiments in this study have therefore used *RBC+Equalize* and *RBC+CLAHE* for augmentation of the *in vitro* and *in vivo* datasets, respectively.

Augmentation type	Settings
Horizontal Flip	Probability of application (p) = 0.5
Vertical Flip	Probability of application (p) = 0.5
Shift scale rotate (SSR)	Shift limit=0.0625, Scale limit=0.1, Rotate limit=45, p=0.5
Sharpen	Range [0.2,0.5], p=0.5
Gaussian Blur (GB)	Gaussian filter with the size of the kernel in the range [3,7]
Random Brightness Contrast (RBC)	Range [-0.2, 0.2], p=0.5
Equalize	Probability of application (p) = 0.5
Contrast Limited Adaptive Histogram Equalization (CLAHE)	Contrast clip limit = 4.0, p=0.5

Table 4.5: Augmentation settings used in this study for both *in vitro* and *in vivo* datasets

Augmentation type	DSC					
	<i>In vitro</i>			<i>In vivo</i>		
	Stone	Laser	Average	Stone	Laser	Average
None	0.8732	0.8438	0.8585	0.8319	0.7936	0.81275
Horizontal Flip	0.8241	0.8309	0.8275	0.8203	0.7947	0.8075
Vertical Flip	0.8179	0.8054	0.81165	0.8102	0.7194	0.7648
Shift scale rotate (SSR)	0.8505	0.8217	0.8361	0.8261	0.7875	0.8068
Sharpen	0.849	0.8177	0.83335	0.8065	0.8001	0.8033
Gaussian Blur (GB)	0.8629	0.8467	0.8548	0.8126	0.8072	0.8099
Random Brightness Contrast (RBC)	0.8743	0.8774	0.87585	0.8265	0.8168	0.8217
Equalize	0.8895	0.8257	0.8576	0.775	0.8217	0.7983
CLAHE	0.8571	0.8251	0.8411	0.839	0.832	0.8355

Table 4.6: Comparison of augmentation techniques (applied on HybResUNet) for both *in vitro* and *in vivo* datasets

4.3.3 Loss function

A novel compound loss function is introduced that is a combination of focal loss, boundary loss, cross-correlation loss, and smoothness loss for improved segmentation performance as described in Section 3.2.3. Table 4.7 presents a quantitative comparison of the segmentation results obtained when a UNet and HybResUNet was trained with the focal loss vs when trained with a combination of other losses in the case of both *in vitro* and *in vivo* test datasets.

Network	Loss function	DSC					
		<i>In vitro</i>			<i>In vivo</i>		
		Stone	Laser	Mean	Stone	Laser	Mean
UNet	L_{FL}	0.8544	0.7643	0.8094	0.7257	0.6991	0.7124
	$L_{FL}+L_{boundary}$	0.8631	0.8401	0.8516	0.7948	0.7657	0.7803
	$L_{FL}+L_{boundary}+L_{sim}$	0.8832	0.8413	0.8622	0.7977	0.7874	0.7925
	$L_{FL}+L_{boundary}+L_{sim}+L_{smo}$	0.8892	0.8582	0.8737	0.7825	0.8144	0.7985
HybResUNet	L_{FL}	0.8698	0.8026	0.8362	0.7712	0.7110	0.7411
	$L_{FL}+L_{boundary}$	0.9037	0.8854	0.8945	0.8011	0.8100	0.8055
	$L_{FL}+L_{boundary}+L_{sim}$	0.8997	0.8898	0.8947	0.8150	0.8243	0.8197
	$L_{FL}+L_{boundary}+L_{sim}+L_{smo}$	0.9055	0.8842	0.8948	0.8251	0.8413	0.8332

Table 4.7: Comparison of loss functions showing accuracy improvement with compound loss as opposed to state-of-the-art focal loss

It can be observed for UNet that the combination of focal loss and boundary loss improved the mean DSC by nearly 5.2 % and 9.5 % in the case of *in vitro* and *in vivo* datasets, respectively. While for the HybResUNet, this improvement was recorded to be 6.5 % and 7.9 % for *in vitro* and *in vivo* datasets, respectively. In the case of UNet, it can be observed that incorporating similarity loss to the combination of focal and boundary loss boosted the DSC for both *in vitro* and *in vivo* datasets by nearly 1.25 % and 1.56%. While in the case of HybResUNet, this improvement can be seen as a boost of 1.76% in the case of *in vivo*. It can also be seen that incorporating both smoothness and similarity loss to the combination of focal loss and boundary loss further increases the mean DSC by 2.6 % and 2.3 % for the UNet and in the case of HybResUNet, it provided a DSC boost of nearly 0.03 % and 3.3 % in case of *in vitro* and *in vivo* datasets, respectively. When combined losses are used instead of focal loss, the DSC for the laser class is boosted by a higher margin as compared to the boost in the stone class.

4.3.4 Quantitative results

I have evaluated the proposed method and compared it with other existing SOTA approaches using standard computer vision metrics such as Dice similarity coefficient (DSC), Jaccard index (JI), Hausdorff distance (HD), positive predictive value (PPV) and sensitivity (defined per pixel) in Table 4.8 and Table 4.9. The details on the test datasets are provided in Table 4.1. In Table 4.8 and Table 4.9 below, I have established a quantitative comparison of the proposed framework against SOTA network architectures.

Table 4.14 presents an ablation study for the integration of dilations, ASPP, and attention gate in the network.

- (I) **Base network:** SOTA methods are compared as the base network of choice for the proposed assembled network. For this case, it can be observed that the HybResUNet provided a significantly higher DSC with a value of 0.8781 for *in vitro* and 0.829 for the *in vivo*. A higher JI, lower HD, higher overall PPV, and sensitivity were also seen as compared to other baseline networks included in the experiment (UNet, DeepResUNet, R2UNet, DeepLabv3+, and Joint model).
- (II) **Base network with DVFNet (with DVF):** In this set of experiments, I propose to incorporate DVFNet together with the best base network in (I), i.e., HybResUNet. For this network, the mean of $DVF_{i \leftarrow (i+2)}$ and $DVF_{(i+2) \leftarrow (i+4)}$ is fed as input to the first HybResUNet as shown in Figure 4.4. In the case of *in vitro*, DVFNet (with DVF) significantly improved the DSC, JI, and HD of HybResUNet by nearly 1.2%, 1%, and 4%, respectively. It can also be seen that the DVFNet further improved the overall PPV and sensitivity of most networks in this set. Although DVFNet (with DVF) can be seen to improve the stone segmentation of HybResUNet in *the in vivo*, it showed no overall improvement in the segmentation results.
- (III) **Base network with DVFNet (with warped images):** This set of experiments involved the incorporation of DVFNet (with warped image) together with the best base network HybResUNet and its derivatives. Here, warped image corresponds to the case when the mean of $X_{warp_{i \leftarrow (i+2)}}$ and $X_{warp_{(i+2) \leftarrow (i+4)}}$ is fed to the input of the first HybResUNet as shown in Figure 4.4. In the case of the *in vitro*, DVFNet (with warped image) can be seen to improve the mean of DSC and JI of Att-HybResUNet(refer to Table 4.14) and Att-ASPP-HybResUNet by 1% and 1.63%, respectively. However, the overall performance is still lower than that obtained from the use of DVFNet (with DVF). For the *in vivo* data, it can be observed that DVFNet (with warped image) improved the performance of all networks: HybResUNet, ASPP-HybResUNet, Att-HybResUNet, Att-ASPP-HybResUNet, particularly for HybResUNet where DSC, JI, and HD were improved by nearly 1.15%, 2.2%, and

<i>In vitro (Test-I)</i>							
Class	Method	DSC	<i>p</i> -values	JI	HD	PPV	Sensitivity
Stone	I UNet [40]	0.8521±0.19	0.0790	0.7736±0.19	6.3096±1.83	0.9155±0.11	0.8479 ±0.20
	I HybResUNet [123]	0.8838±0.14	0.1063	0.8143±0.18	6.0125±2.37	0.9123±0.10	0.8914±0.17
	I DeepResUNet [122]	0.8828±0.12	0.1781	0.8054±0.15	6.3286±1.55	0.9239±0.10	0.8732±0.15
	I R2-UNet [113]	0.8215±0.20	0.0258*	0.7344±0.22	6.7252±2.08	0.8855±0.13	0.8280±0.22
	I DeepLabv3+(ResNet-50) [41]	0.8004±0.18	0.0153*	0.6982±0.21	7.2936±1.96	0.8411±0.21	0.8110±0.20
	I Joint model [137]	0.7484±0.17	0.0042*	0.6268±0.21	7.5695±1.77	0.7470±0.23	0.8057±0.15
	I ML-UNet [81]	0.8391±0.15	0.0139*	0.7485±0.20	6.8246±2.66	0.8383±0.18	0.8985±0.17
	II HybResUNet+DVFNet (with DVF) [†]	0.9068 ±0.09	—	0.8413 ±0.14	5.7454 ±2.11	0.9299±0.10	0.9040±0.13
	II ASPP-HybResUNet+DVFNet (with DVF) [†]	0.8253±0.14	0.0201*	0.7250±0.19	7.0101±2.17	0.8244±0.20	0.8847±0.15
	II Att-HybResUNet+DVFNet (with DVF) [†]	0.8872±0.14	0.1821	0.8180±0.17	6.2642±2.32	0.9136±0.09	0.8919±0.17
	II Att-ASPP-HybResUNet+DVFNet (with DVF) [†]	0.8238±0.19	0.0417*	0.7380±0.23	6.3834±2.15	0.8620±0.19	0.8585±0.20
	III HybResUNet+DVFNet (with warped image) [†]	0.8801±0.14	0.2451	0.7985±0.18	6.2620±1.93	0.8899±0.19	0.8931±0.15
	III ASPP-HybResUNet+DVFNet (with warped image) [†]	0.8295±0.19	0.1426	0.7447±0.22	6.5654±1.92	0.8165±0.23	0.8768±0.17
	III Att-HybResUNet+DVFNet (with warped image) [†]	0.8899±0.13	0.3757	0.8211±0.16	5.8026±1.97	0.9534 ±0.05	0.8626±0.18
	III Att-ASPP-HybResUNet+DVFNet (with warped image) [†]	0.8860±0.10	0.4739	0.8089±0.15	6.2607±2.10	0.8898±0.14	0.9113 ±0.12
Laser	I UNet [40]	0.8388±0.25	0.2771	0.7787±0.27	3.3187 ±1.05	0.8393±0.26	0.8408±0.24
	I HybResUNet [123]	0.8724±0.17	0.8292	0.8031±0.20	3.6028±1.30	0.8792±0.20	0.8780±0.15
	I DeepResUNet [122]	0.8276±0.23	0.1048	0.7541±0.25	3.6314±1.15	0.8609±0.24	0.8065±0.21
	I R2-UNet [113]	0.8061±0.27	0.1080	0.7368±0.28	4.1269 ±1.52	0.8380±0.22	0.8013±0.28
	I DeepLabv3+(ResNet-50) [41]	0.8446±0.20	0.2520	0.7717±0.23	3.8366±1.43	0.8626±0.18	0.8452±0.22
	I Joint model [137]	0.7122±0.23	0.0001*	0.5943±0.23	4.6308±1.11	0.7399±0.25	0.6991±0.24
	I ML-UNet [81]	0.8094±0.23	0.0914	0.7307±0.26	4.0275±1.68	0.7821±0.23	0.8561±0.24
	II HybResUNet+DVFNet (with DVF) [†]	0.8702±0.14	—	0.7923±0.17	3.4938±1.16	0.8510±0.18	0.9069±0.10
	II ASPP-HybResUNet+DVFNet (with DVF) [†]	0.8278±0.24	0.1349	0.7572±0.25	3.8447±1.27	0.8320±0.26	0.8320 ±0.22
	II Att-HybResUNet+DVFNet (with DVF) [†]	0.9004 ±0.08	0.1069	0.8263 ±0.11	3.2977 ±0.95	0.8891±0.12	0.9227±0.04
	II Att-ASPP-HybResUNet+DVFNet (with DVF) [†]	0.8475±0.22	0.3302	0.7809±0.24	3.5967±1.08	0.8326±0.24	0.8740±0.19
	III HybResUNet+DVFNet (with warped image) [†]	0.8708±0.19	0.4718	0.7898±0.17	3.6420±1.42	0.8955 ±0.15	0.8697±0.13
	III ASPP-HybResUNet+DVFNet (with warped image) [†]	0.8758±0.16	0.6223	0.8066±0.19	3.5898±1.27	0.8556±0.20	0.9118±0.11
	III Att-HybResUNet+DVFNet (with warped image) [†]	0.8393±0.23	0.2714	0.7711±0.24	3.7712±1.73	0.8312±0.24	0.8536±0.22
	III Att-ASPP-HybResUNet+DVFNet (with warped image) [†]	0.8769±0.12	0.5193	0.7984±0.16	3.6418±0.98	0.8548±0.17	0.9233 ±0.05

Table 4.8: Quantitative comparison of proposed network architectures against existing approaches on the ureteroscopy and laser lithotripsy *in vitro* test set (Test-I) where listed values represent average performances. Here, ‘I’ in the table represents the performance of various baseline models, ‘II’ represents the combination of the best-performing baseline model (HybResUNet) and DVFNet (with DVF) under different configurations, and ‘III’ represents the combination of the best-performing baseline model (HybResUNet) and DVFNet(with warped image) under different configurations. The SOTA methods have been accordingly referenced and the experimental methods have been indicated by a [†]superscript. *p*-values that represent statistical significance between the proposed method and other implementations with *p*-value < 0.05 are computed using paired t-test (highlighted by a * superscript)

2.83%, respectively. In addition to this, DVFNet (with warped image) can be seen to increase the sensitivity of all networks in this set.

4.3.5 Qualitative results

In this section, I have presented the qualitative results of segmentation obtained from the proposed framework as opposed to the ground truth and other SOTA approaches in both *in vitro* and *in vivo* cases. Figure 4.7 demonstrates that the proposed model outperforms the existing approaches by overcoming the challenges and providing a more accurate delineation of stone and laser fibre in the case of both *in vitro* and *in vivo* datasets. As evident in Figure 4.7, it can be clearly seen that laser fibre is nearly segmented well by all models except for some difficult frames like the first image in the *in vivo* dataset wherein it

In vivo (Test-I)									
Class	Method	DSC	p-values	JI	HD	PPV	Sensitivity		
Stone	I	UNet [40]	0.8129±0.13	0.5556	0.7025±0.16	5.4626±1.40	0.8312±0.19	0.8374±0.12	
		HybResUNet [123]	0.8339±0.15	0.2862	0.7381 ±0.17	5.3870±1.75	0.8462 ±0.21	0.8525±0.09	
		DeepResUNet [122]	0.8214±0.13	0.8819	0.7140±0.15	5.4958±1.33	0.8167±0.18	0.8543±0.11	
		R2-UNet [113]	0.7734±0.16	0.0162*	0.6536±0.18	6.0680±1.35	0.7580±0.21	0.8328±0.11	
		DeepLabv3+(ResNet-50) [41]	0.7653±0.16	0.0008*	0.6438±0.19	5.7150±1.32	0.7573±0.24	0.8375±0.13	
		Joint model [137]	0.6900±0.22	0.0007*	0.5660±0.23	5.8173±1.44	0.7218±0.26	0.7003±0.23	
	II	MI-UNet [81]	0.7126±0.20	0.0017*	0.5852±0.21	6.0840±1.43	0.6875±0.24	0.7938±0.19	
		HybResUNet+DVFNet (with DVF) [†]	0.8347 ±0.14	0.3036	0.7357±0.17	5.3560 ±1.47	0.8429±0.19	0.8473±0.09	
		ASPP-HybResUNet+DVFNet (with DVF) [†]	0.8115±0.18	0.4706	0.7127±0.20	5.3123 ±1.73	0.8148±0.23	0.8430±0.15	
		Att-HybResNet+DVFNet (with DVF) [†]	0.7996±0.16	0.2450	0.6914±0.19	5.6316±1.47	0.7506±0.22	0.9026 ±0.09	
		Att-ASPP-HybResUNet+DVFNet (with DVF) [†]	0.8072±0.17	0.3592	0.7036±0.19	5.4765±1.60	0.7881±0.22	0.8668±0.13	
		III	HybResUNet+DVFNet (with warped image)[†]	0.8203±0.14	—	0.7158±0.17	5.4264±1.43	0.8226±0.20	0.8562±0.12
			ASPP-HybResUNet+DVFNet (with warped image) [†]	0.7992±0.16	0.1227	0.6911±0.19	5.5953±1.55	0.7817±0.23	0.8682±0.11
			Att-HybResUNet+DVFNet (with warped image) [†]	0.8183±0.16	0.8483	0.7156±0.18	5.5431±1.65	0.8032±0.21	0.8713±0.09
			Att-ASPP-HybResUNet+DVFNet (with warped image) [†]	0.8016±0.15	0.0975	0.6918±0.18	5.3543±1.59	0.8062±0.21	0.8382±0.12
Laser	I	UNet [40]	0.7974±0.21	0.0548*	0.7043±0.24	4.1557±1.48	0.8050±0.24	0.8137±0.20	
		HybResUNet [123]	0.8241±0.18	0.0331*	0.7328±0.21	3.9540±1.35	0.8640±0.17	0.8214±0.20	
		DeepResUNet [122]	0.7851±0.27	0.0277*	0.7023±0.26	4.2869±1.40	0.8084±0.27	0.7783±0.27	
		R2-UNet [113]	0.7678±0.22	0.0315*	0.6636±0.23	4.4229±1.14	0.7875±0.22	0.7979±0.23	
		DeepLabv3+(ResNet-50) [41]	0.7144±0.29	0.0148*	0.6200±0.29	4.4581±1.06	0.7287±0.29	0.7111±0.29	
		Joint model [137]	0.6348±0.30	0.0001*	0.5271±0.29	4.7468±0.98	0.6991±0.28	0.6111±0.31	
		MI-UNet [81]	0.7249±0.22	0.0009*	0.6082±0.24	4.8526±1.31	0.7445±0.23	0.7635±0.19	
	II	HybResUNet+DVFNet (with DVF) [†]	0.7834±0.26	0.0245*	0.6965±0.25	4.3907±1.41	0.7748±0.23	0.8409 ±0.27	
		ASPP-HybResUNet+DVFNet (with DVF) [†]	0.8048±0.18	0.0999	0.7034±0.21	4.4334±1.04	0.7826±0.17	0.8385±0.20	
		Att-HybResUNet+DVFNet (with DVF) [†]	0.8017±0.24	0.0234*	0.7144±0.23	4.3347±1.20	0.8352±0.19	0.8300±0.26	
		Att-ASPP-HybResUNet+DVFNet (with DVF) [†]	0.8374±0.13	0.5780	0.7397±0.17	4.2352±0.81	0.7979±0.17	0.8975 ±0.12	
	III	HybResUNet+DVFNet (with warped image) [†]	0.8568±0.21	—	0.7878 ±0.21	3.6501±0.99	0.8894 ±0.18	0.8487±0.22	
		ASPP-HybResUNet+DVFNet (with warped image) [†]	0.8658 ±0.15	0.7439	0.7852±0.17	3.6361 ±0.64	0.8634±0.16	0.8770±0.15	
		Att-HybResUNet+DVFNet (with warped image) [†]	0.8389±0.19	0.0701	0.7558±0.20	3.9573±0.95	0.8798±0.17	0.8370±0.21	
		Att-ASPP-HybResUNet+DVFNet (with warped image) [†]	0.8478±0.18	0.7321	0.7662±0.20	3.9358±0.90	0.8276±0.20	0.8784±0.18	

Table 4.9: Quantitative comparison of proposed network architectures against existing approaches on the ureteroscopy and laser lithotripsy *in vivo* test set (Test-I) where listed values represent average performances. Here, 'I' in the table represents the performance of various baseline models, 'II' represents the combination of the best performing baseline model (HybResUNet) and DVFNet (with DVF) under different configurations, and 'III' represents the combination of the best performing baseline model (HybResUNet) and DVFNet(with warped image) under different configurations. The SOTA methods have been accordingly referenced and the experimental methods have been indicated by a [†]superscript. *p*-values that represent statistical significance between the proposed method and other implementations with *p*-value < 0.05 are computed using paired t-test (highlighted by a * superscript)

is only the proposed model that is able to clearly segment the laser. It can also be observed from Figure 4.7 that the existing approaches are not able to provide a clear segmentation of stone in most *in vitro* and *in vivo* frames and hence some debris is segmented as part of the stone, resulting in either underestimation or overestimation of the stone size.

4.3.6 Extended out-of-sample test

I have evaluated the proposed *in vivo* framework on a separate out-of-sample dataset (Test-II) which is acquired from a second site at Boston Scientific. As shown in Table 4.10, the proposed model HybResUNet+DVFNet (with warped image) provided a significant DSC improvement of 9.61%, 11%, 8.31%, 9.86% and 5.24% over UNet, HybResUNet, DeepResUNet, R2-UNet and DeepLabv3+, respectively in case of the stone class. While

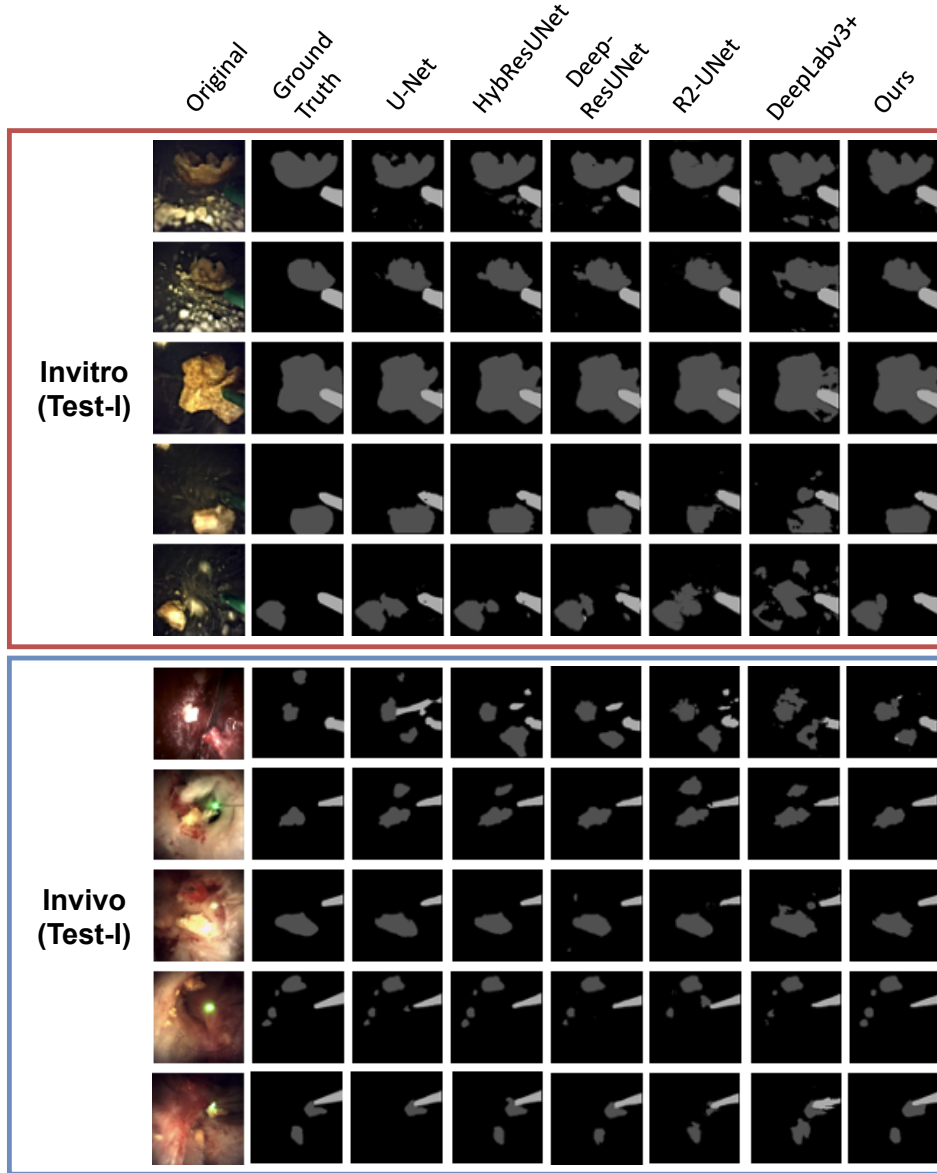


Figure 4.7: Qualitative analysis of the proposed method (Att-HybResUNet+DVFNet (with DVF) for the *in vitro* and HybResUNet+DVFNet (with warped image) for the *in vivo*) against existing SOTA methods on the ureteroscopy and laser lithotripsy test sets (Test-I). Each row shows a test image, followed by its ground truth segmentation mask (showing laser fibre and stone), followed by SOTA approaches: UNet, HybResUNet, Deep-ResUNet, R2-UNet, and Deeplabv3+, and finally the proposed model which is Att-HybResUNet+DVFNet (with DVF) for the *in vitro* and HybResUNet+DVFNet (with warped image) for the *in vivo*.

for the laser class, DSC saw an improvement of 31.79%, 22.15%, 30.13%, 14.52% and 10.42% over UNet, HybResUNet, DeepResUNet, R2-UNet and Deeplabv3+, respectively. Further, one can also see from Table 4.10 that the ASPP module further improved the performance of HybResUNet+DVFNet by 1.62% in dice score and 1.74% in Jaccard

<i>In vivo (Test-II)</i>							
Class	Method	DSC	<i>p</i> -values	JI	HD	PPV	Sensitivity
Stone	UNet [40]	0.3516±0.22	0.3185	0.2382±0.19	8.8137±1.69	0.4192±0.30	0.3426±0.22
	HybResUNet [123]	0.3471±0.25	0.0400*	0.2413±0.21	9.0627±1.95	0.3977±0.31	0.3451±0.24
	DeepResUNet [122]	0.3558±0.24	0.1138	0.2471±0.21	8.5908±1.63	0.4071±0.31	0.3518±0.23
	R2-UNet [113]	0.3508±0.23	0.0954	0.2406±0.20	8.9529±1.73	0.4221±0.30	0.3513±0.26
	DeepLabv3+(ResNet-50) [41]	0.3662±0.24	0.0062*	0.2520±0.20	8.6843±1.73	0.4288±0.30	0.3480±0.21
	HybResUNet+DVFNet (with warped image) [†]	0.3854 ±0.25	0.1227	0.2728±0.22	8.6852±1.90	0.4262±0.29	0.3873 ±0.25
	ASPP-HybResUNet+DVFNet (with warped image) [†]	0.3852±0.25	—	0.2733 ±0.22	8.5904 ±1.80	0.4383 ±0.31	0.3775±0.24
	Att-HybResUNet+DVFNet (with warped image) [†]	0.3757±0.25	0.1599	0.2656±0.22	8.9544±2.03	0.4228±0.30	0.3794 ±0.24
Laser	Att-ASPP-HybResUNet+DVFNet (with warped image) [†]	0.3607±0.25	0.8447	0.2538±0.22	8.7278±1.96	0.4157±0.32	0.3432±0.23
	UNet [40]	0.5564±0.29	0.0284*	0.4385±0.27	5.7087±0.90	0.5824 ±0.31	0.5746±0.31
	HybResUNet [123]	0.6003±0.24	0.0988	0.4673±0.22	6.0062±1.11	0.6443±0.29	0.5969±0.25
	DeepResUNet [122]	0.5635±0.30	0.0459*	0.4465±0.27	5.9400±0.85	0.6255±0.31	0.5452±0.31
	R2-UNet [113]	0.6403±0.21	0.0086*	0.5050±0.22	5.4860±1.10	0.6860±0.25	0.6417±0.23
	DeepLabv3+(ResNet-50) [41]	0.6641±0.11	0.0020*	0.5071±0.12	5.4833±0.91	0.8078±0.14	0.5854±0.14
	HybResUNet+DVFNet (with warped image) [†]	0.7333±0.13	0.7438	0.5956±0.16	5.2917 ±1.09	0.8131±0.17	0.7057 ±0.18
	ASPP-HybResUNet+DVFNet (with warped image) [†]	0.7452 ±0.11	—	0.6060 ±0.14	5.4093±1.14	0.8417 ±0.15	0.7016±0.15
	Att-HybResUNet+DVFNet (with warped image) [†]	0.6742±0.17	0.2658	0.5332±0.19	5.8803±1.01	0.7300±0.22	0.6732±0.20
	Att-ASPP-HybResUNet+DVFNet (with warped image) [†]	0.7164±0.14	0.1468	0.5765±0.16	5.4492±0.95	0.7712±0.17	0.7005±0.17

Table 4.10: Quantitative comparison of proposed network architecture for *in vivo* data against existing approaches on an unseen *in vivo* test dataset (Test-II) where listed values represent average performances. *p*-values represent statistical significance between proposed method and other implementations with *p*-value < 0.05 (highlighted by a * superscript)

index for the laser class. Qualitative results in Figure 4.10 on this dataset *also* demonstrate that the proposed model for the *in vivo* is able to provide the most accurate segmentation of all frames as compared to the other existing approaches.

To justify the experiments, I have also evaluated the trained models on 18 samples obtained from a separate unseen test patient data (example samples of patient 3 are shown in 4.8). It can be seen in Table 4.11 and Figure 4.9 that the proposed approach outperforms SOTA approaches on out-of-sample patient data. Additionally, in order to demonstrate that no data leak has occurred in the test samples, patient-wise 4-fold cross-validation was also performed on the *in vivo* data and the proposed approach provided an average (across folds) DSC improvement of 9.98% and 11.5% over SOTA UNet for the stone and laser classes, respectively (see Table 4.12).

4.3.7 Inference time analysis

In this section, computation time analysis of different networks involved in this study has been shown with the results presented in Table 4.13. The inference time was calculated by running 10 test images on NVIDIA Quadro RTX 6000. As mentioned in Section 4.2.2.4, sequence samples and DVFNet are only employed during training and not used during test time. In Table 4.4, I have provided a comparison of DSC and computation time

<i>In vivo (Test on unseen patient data)</i>					
Class	Method	DSC	J1	PPV	Sensitivity
Stone	UNet	0.8883 $\pm$ 0.10	0.8121 $\pm$ 0.15	0.8820 $\pm$ 0.14	0.9103 $\pm$ 0.09
	HybResUNet	0.9179 $\pm$ 0.03	0.8500 $\pm$ 0.06	0.9328 $\pm$ 0.06	0.9091 $\pm$ 0.06
	HybResUNet+DVFNNet(with warped image)	0.9205 $\pm$ 0.03	0.8545 $\pm$ 0.06	0.9267 $\pm$ 0.04	0.9180 $\pm$ 0.06
Laser	UNet	0.8697 $\pm$ 0.08	0.7775 $\pm$ 0.11	0.8600 $\pm$ 0.09	0.8850 $\pm$ 0.10
	HybResUNet	0.8581 $\pm$ 0.04	0.7539 $\pm$ 0.06	0.8446 $\pm$ 0.08	0.8800 $\pm$ 0.07
	HybResUNet+DVFNNet(with warped image)	0.8702 $\pm$ 0.07	0.7769 $\pm$ 0.11	0.8671 $\pm$ 0.11	0.8843 $\pm$ 0.07

Table 4.11: Quantitative comparison of proposed network architecture for *in vivo* data against baseline methods on 18 samples from unseen patient data where listed values represent average performances.

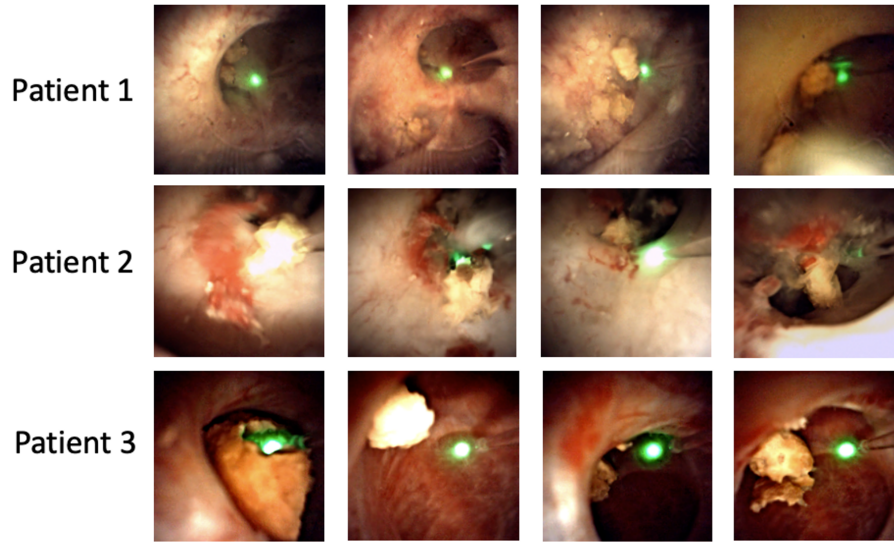


Figure 4.8: Random frames from different patients illustrating the intra and inter-patient variability across data. Here, patient 3 represents the out-of-sample test dataset used for generalisability assessment.



Figure 4.9: Qualitative comparison of proposed network architecture for *in vivo* data against baseline methods on 18 samples from an unseen patient data (example images shown in row 3 of Figure 4.8)

	Network	Fold1		Fold2		Fold3		Fold4		Average	
		DSC	JI	DSC	JI	DSC	JI	DSC	JI	DSC	JI
Stone	UNet	0.668	0.531	0.626	0.483	0.579	0.445	0.663	0.532	0.634±0.04	0.498±0.04
	HybResUNet	0.697	0.557	0.615	0.470	0.649	0.512	0.657	0.521	0.655±0.03	0.515±0.04
	HybResUNet +DVFNet (warped img)	0.678	0.538	0.622	0.484	0.751	0.633	0.738	0.603	0.697±0.06	0.565±0.07
Laser	UNet	0.685	0.565	0.544	0.427	0.058	0.036	0.837	0.724	0.531±0.34	0.438±0.29
	HybResUNet	0.692	0.568	0.590	0.473	0.122	0.073	0.860	0.756	0.566±0.32	0.467±0.29
	HybResUNet +DVFNet (warped img)	0.668	0.543	0.672	0.552	0.241	0.150	0.787	0.653	0.592±0.24	0.475±0.22

Table 4.12: Performance comparison of 4-fold patient-wise cross-fold validation performed on *in vivo* data where each fold represents validation results on different subjects (Fold1, Fold2 and Fold4 are human patient data while Fold3 represents animal data).

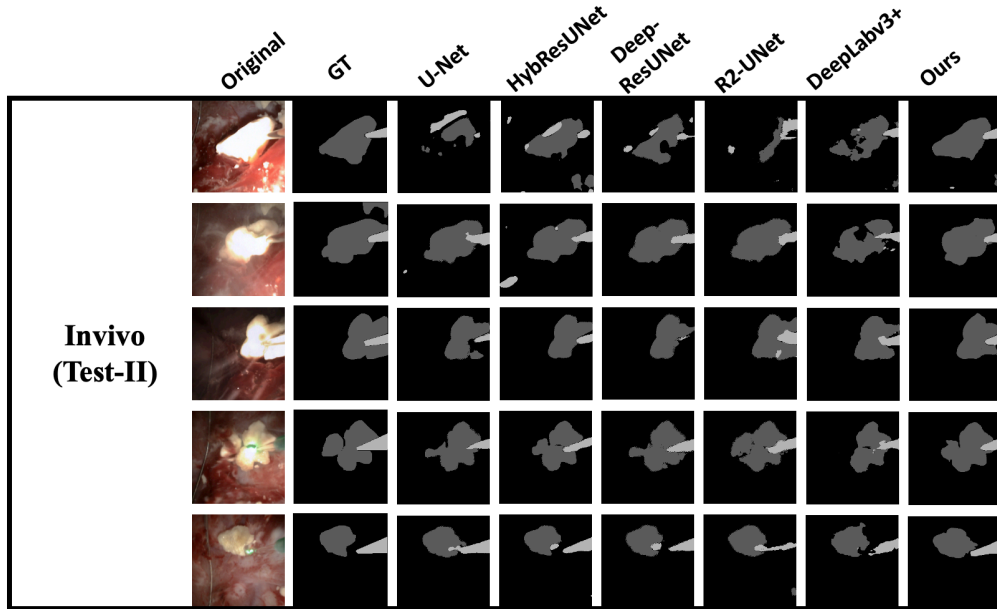


Figure 4.10: Qualitative analysis of the proposed method (HybResUNet+DVFNet (with warped image)) for *in vivo* against existing SOTA methods on the unseen *in vivo* test dataset (Test-II).

(for 10 samples on NVIDIA Quadro RTX 6000) for HybResUNet+DVFNet applied on sequence data vs single sample data in the test dataset. Here, one can observe that using sequence samples does not show significant improvement with respect to Dice. However, the inference time for the sequence sample exceeds 6 times as compared to that when computed with a single frame. Therefore, the weights of the HybResUNet learned using sequence samples during training are used during frame-wise inference. This allows the network to perform inference in real time.

Network	Computation time (s)
UNet [40]	0.8427
HybResUNet [123]	0.8623
DeepResUNet [122]	0.8736
R2-UNet [113]	1.2164
DeepLabv3+(ResNet50) [41]	0.7440
HybResUNet +DVFNet	5.2272
ASPP-HybResUNet+DVFNet	5.1638
Att-HybResUNet+DVFNet	5.2366
Att-ASPP-HybResUNet+DVFNet	5.0530

Table 4.13: Inference time of base network for 10 images of the test set on NVIDIA Quadro RTX 6000. DVFNet is only used during training and not in test inference

4.3.8 Ablation study of network design

Table 4.14 presents an ablation study for the integration of various combinations of dilations, ASPP and attention gate in different networks.

- (I) **Base network with ASPP:** Here, we compare the effect of dilated convolutions and ASPP module connected to each base network presented in (I) in Table 4.14. It can be seen that in the case of ASPP-HybResUNet, ASPP does not provide any overall improvement as compared to only HybResUNet. It can also be seen that ASPP increases the sensitivity for both stone and laser classes but decreases the PPV of the stone class in most networks for the *in vitro* case.
- (II) **Base network with Attention(Att):** In the second set of experiments in Table 4.14, one can see that using attention increases the overall sensitivity of most networks.
- (III) **Base network with Att-ASPP:** In the third set of experiments, one can see that the combination of both Attention and ASPP does not lead to any improvement in the segmentation of stone or laser in both *in vitro* and *in vivo* cases.

4.4 Discussion

This study is the first attempt to develop a multi-class segmentation approach for ureteroscopy and laser lithotripsy. In this work, I conducted experiments for *in vitro* and *in vivo* datasets independently. This is due to the large variability that exists between the

Dataset	Method	DSC		JI		HD		PPV		Sensitivity		
		Stone	Laser	Stone	Laser	Stone	Laser	Stone	Laser	Stone	Laser	
In vitro (Test-I)	I	ASPP-UNet [†]	0.8504	0.8656	0.7580	0.7958	6.563	3.5086	0.8845	0.8612	0.8553	0.8780
		ASPP-HybResUNet [†]	0.8777	0.8794	0.8015	0.8075	6.0554	3.6527	0.8917	0.8735	0.8997	0.9141
		ASPP-DeepResUNet [†]	0.7934	0.8647	0.6825	0.7891	7.5642	3.7578	0.7611	0.8537	0.9094	0.9006
		ASPP-R2-UNet [†]	0.8506	0.8832	0.7597	0.8106	6.7954	3.6081	0.8785	0.8717	0.8654	0.9192
	II	Att-UNet [115]	0.8730	0.8488	0.7925	0.7777	6.3598	3.49	0.9021	0.8426	0.8796	0.8657
		Att-HybResUNet [†]	0.8844	0.8264	0.8192	0.7592	5.6916	4.0531	0.9402	0.8402	0.8730	0.8168
		Att-DeepResUNet [†]	0.8960	0.8491	0.8244	0.7729	5.8547	3.5731	0.9392	0.8404	0.8776	0.8746
		Att-R2-UNet [134]	0.8788	0.8481	0.8007	0.7731	5.8442	3.6752	0.9118	0.8291	0.8779	0.8814
	III	Att-ASPP-UNet [†]	0.8301	0.8643	0.7357	0.7852	6.8324	3.8761	0.8340	0.8615	0.8733	0.8807
		Att-ASPP-HybResUNet [†]	0.8641	0.8693	0.7790	0.8038	6.4245	3.4374	0.8339	0.8794	0.9350	0.8788
		Att-ASPP-DeepResUNet [†]	0.8571	0.8578	0.7648	0.7775	6.457	3.7789	0.8788	0.8430	0.8635	0.9025
		Att-ASPP-R2-UNet [†]	0.8338	0.8782	0.7370	0.8056	6.6628	3.6545	0.8292	0.8577	0.8832	0.9340
In vivo (Test-I)	I	ASPP-UNet [†]	0.7963	0.7865	0.6854	0.6928	5.5427	3.9079	0.7991	0.7834	0.8416	0.801
		ASPP-HybResUNet [†]	0.8035	0.8334	0.6939	0.7453	5.7695	3.7815	0.814	0.8249	0.833	0.8558
		ASPP-DeepResUNet [†]	0.8234	0.7968	0.7152	0.699	5.2131	3.5706	0.8282	0.816	0.861	0.7855
		ASPP-R2-UNet [†]	0.7939	0.7409	0.6805	0.6453	5.5599	4.0746	0.7986	0.7261	0.8334	0.7638
	II	Att-UNet [115]	0.8255	0.8094	0.7157	0.7292	5.3276	4.2002	0.8526	0.8325	0.8358	0.8316
		Att-HybResUNet [†]	0.819	0.8288	0.7388	0.7761	5.3323	4.096	0.8033	0.8763	0.8725	0.8231
		Att-DeepResUNet [†]	0.8119	0.7726	0.6994	0.6832	5.3048	4.5833	0.8334	0.7736	0.8141	0.7825
		Att-R2-UNet [134]	0.7985	0.6966	0.6794	0.5858	5.745	4.5984	0.7902	0.7343	0.8293	0.6893
	III	Att-ASPP-UNet [†]	0.8079	0.8364	0.7019	0.7478	5.4895	3.6628	0.7923	0.8222	0.8712	0.8606
		Att-ASPP-HybResUNet [†]	0.803	0.8229	0.6842	0.7282	5.5082	3.9781	0.8391	0.8119	0.7982	0.8495
		Att-ASPP-DeepResUNet [†]	0.8185	0.8087	0.712	0.7071	5.3754	4.1094	0.8025	0.7999	0.87	0.8332
		Att-ASPP-R2-UNet [†]	0.7814	0.7913	0.6617	0.6794	5.5792	4.1598	0.7837	0.7943	0.8278	0.8032

Table 4.14: Quantitative comparison of proposed network architectures against existing approaches on the ureteroscopy and laser lithotripsy test sets (Test-I). The published networks have been accordingly cited and the experiment networks have been indicated by a [†]superscript.

two datasets, as evident from Figure 4.2. Table 4.3 and Figure 4.5 present quantitative and qualitative comparisons, respectively, for both *in vitro* and *in vivo*, both trained independently and jointly. The performance decrease in the case of the joint training can be attributed to the significant variability between the *in vitro* and *in vivo* datasets as illustrated by a t-SNE plot in Figure 4.6 showing a distinct disjoint between *in vitro* and *in vivo* dataset. The variability between the *in vitro* and *in vivo* datasets can also be seen from Figure 4.1, which shows that the *in vivo* background has more textural information, higher heterogeneity, debris from tissue and blood, and other image artefacts that are almost absent in the *in vitro* dataset.

Firstly, a HybResUNet [123] network was trained on the *in vitro* and *in vivo* datasets by using different augmentation techniques and DSC for each class was recorded to determine the optimised augmentation strategy that best captures the target data (refer to the Table 4.6). It can be observed from Figure 4.1 that the laser fibre is always present in a certain orientation and in the right part of the image. This explains why networks trained with spatial-level transforms such as flips, shift and rotate are not able to perform

well on the test dataset. It can also be seen from Table 4.6 that the *random brightness contrast (RBC)* and the histogram equalisation in case of *in vitro*, and *RBC* and *CLAHE* in the case of *in vivo* dataset seem to either improve the segmentation accuracy or provide a competitive performance as compared to the case with no augmentations. Therefore, *RBC+Equalize* and *RBC+CLAHE* were used for data augmentation in all experiments for *in vitro* and *in vivo* datasets, respectively.

A compound loss function has been used to enable better differentiation between easy or hard examples and improve the segmentation performance. In Table 4.7, one can see that the boundary loss when combined with focal loss leads to an improvement in the case of *in vitro* and *in vivo* datasets, respectively. The *in vitro* dataset has less scope for improvement as it has no textural information in the background and relatively more pronounced boundaries for stone and laser class as compared to the *in vivo* dataset. This explains the relatively higher improvement for the *in vivo* as compared to the *in vitro* dataset. The incorporation of similarity loss and smoothness constraint further boosted the DSC for both *in vitro* and *in vivo* datasets. It can also be noted that the accuracy of the laser class was boosted by an overall higher margin as compared to the stone class. I hypothesise that such discrepancy in the improvement percentage between the stone class and the laser class is possibly due to the stone class having more variability in terms of shape and texture as compared to the laser class.

HybResUNet has a relatively simpler model with residual blocks in the encoder path only as opposed to the DeepResUNet [122] and R2-UNet [113]. It can also be seen from Table 4.2 that the number of trainable parameters is significantly less for HybResUNet as compared to DeepResUNet and R2-UNet. I hypothesize that DeepResUNet and R2-UNet make the UNet network more complicated and tend to overfit the training data, thereby leading to poor performance (refer to Table 4.8 and Table 4.9) as compared to HybResUNet. Atrous Spatial Pyramid Pooling (ASPP) technique uses a series of different dilation rates in parallel to capture multi-scale contextual information. The use of dilated convolutions in an increasing order followed by decreasing order helps better aggregation of local features and improves detection of small objects [126]. This explains the improvement in the Dice similarity coefficient and Jaccard index of the laser class

for ASPP-HybResUNet in both *in vitro* and *in vivo* datasets (refer to the Table 4.14). It can also be observed that when the attention gate (AG) model is integrated into the encoder-decoder HybResUNet network, it improved the segmentation of the stone class and also improved the sensitivity by suppressing irrelevant responses in the network. However, when I tried integrating the attention gate mechanism, dilated convolutions, and ASPP, the networks tend to become overly complicated and hence there was no improvement observed in the segmentation performance (refer to the Table 4.14).

To further improve the segmentation performance of the proposed network, I integrated DVNet into the framework. As can be observed in Table 4.8 and Table 4.9 respectively, DVNet (with DVF) and DVNet (with warped image) when integrated with HybResUNet improved the performance in the *in vitro* and *in vivo* datasets, respectively. In the case of *in vitro*, HybResUNet+DVNet (with DVF) provided an improvement of 1.18%, 1%, and 3.9 % in DSC, JI, and HD, respectively. This network when combined with the attention gate model, Att-HybResUnet+DVNet (with DVF), slightly improved the segmentation performance, thereby providing the best results on the *in vitro* dataset. On the other hand, HybResUNet+DVNet (with warped image) in case of the *in vivo* performed the best and can be seen to provide an improvement of 1.15%, 2.22%, and 2.83 % in DSC, JI, and HD, respectively.

It can also be seen in Table 4.8 that for the *in vitro* data, there is not much difference in the results for DVNet (with DVF) and DVNet (with warped image). On the other hand, it can be seen in Table 4.9 that for the *in vivo* samples, DVNet (with warped image) provided significantly better results as compared to DVNet (with DVF). Therefore, I hypothesise that as the *in vitro* lacks any background texture, it does not make a big difference as to which DVNet approach should be adopted. But, for the *in vivo* that has the texture-rich background, DVNet (warped image) performs better as it retains the high-level background information.

It can also be seen in Table 4.8 that in the case of the stone class, the proposed method “HybResUNet+DVNet (with DVF)” has the least standard deviation (± 0.09 in DSC) and p-value < 0.05 for most SOTA baseline methods. Similarly, for the laser

class, our method achieved a standard deviation of only ± 0.14 in DSC. Its variant (Att-HybResUNet+DVNet (with DVF)) improved the performance further and showed the least standard deviation (± 0.08 in DSC) as compared to the SOTA baseline methods. The *in vitro* dataset has no textural information in the background, making it relatively easier for the networks to learn the segmentation of stone and laser class. This leaves less scope for improvement, especially for the laser class which has lesser variability as opposed to the stone class. We, therefore, do not observe any statistically significant improvement ($p\text{-values} > 0.05$) in the segmentation of laser fibre. In Table 4.9 for the clinical data, one can observe that the proposed method provided a significant improvement ($p\text{-value} < 0.05$) over all SOTA baseline methods for the laser class and over most SOTA methods for the stone class.

In Figure 4.7, the proposed framework is able to outperform other existing networks by overcoming the effect of debris and other artefacts in the case of both *in vitro* and *in vivo* datasets. The first image in the *in vivo* dataset in Figure 4.7 gets one of the worst-performing results across all networks. It can be observed that the small stone on the top part of the image is less visible due to relatively less light falling on it and is therefore not getting captured by any models. It can also be observed that the bright-red tissue protrusion in this image is being misclassified as a stone by all the networks. This shows that the networks are sensitive to any tissue protrusions in the images. In addition to this, it should also be observed in this image that the proposed model is able to provide the most accurate segmentation of the laser fibre as opposed to the other networks. I hypothesise that the other models fail to segment the laser accurately due to the fact that it is transparent in appearance and the laser light is not activated making it difficult for the network to spot it. Further, the segmentation prediction on the second and third *in vivo* images shows that the proposed model is not only able to overcome the effect of stone debris but also blood and white tissue debris. Based on the results of the fourth *in vivo* image, I hypothesise that the proposed model is able to successfully pick up secondary small stones and perform well in dynamic illumination situations.

In addition, to further justify that the sample-level split presented in the paper did not cause a data leak in the test results, I evaluated the trained models on unseen patient

data (Table 4.11 and Figure 4.9). It is evident that the proposed model outperforms the SOTA baseline models. Similarly, on a 4-fold patient-wise cross-validation using the *in vivo* data, the proposed approach showed consistent performance outperforming the SOTA methods (Table 4.12). I also performed an extended out-of-sample test study on only animal data referred to as “Test-II” of the proposed framework. These data were acquired from a new site (Boston Scientific) and the quantitative results and qualitative results are presented in Table 4.10 and Figure 4.10, respectively. As shown in Table 4.10, the proposed framework outperformed the existing SOTA approaches on Test-II samples as well, thereby demonstrating that the proposed model not only outperforms other SOTA networks but also is robust to new ureteroscopy and lithotripsy datasets. Even though Table 4.10 shows a decrease in metric values on the Test-II as compared to the Test-I dataset, it is important to note that the models were trained on mainly human ureteroscopy samples while Test-II consists of images of ureteroscopy performed on animal subjects only (Figure 4.3). Additionally, the hardware settings such as LED illumination and software configuration settings used at the time of surgery and image acquisition are different for some animal studies present in Test-I compared to the Test-II dataset. This indicates that the trained models do not cover all the possible variance scenarios arising from different image acquisition settings, resulting in a performance decrease in the Test-II dataset. Increasing the number of animal samples in the training set and adding variability in training data, for example, in terms of kidney stone heterogeneity and background scene, can improve model robustness. Additionally, using learning-based data augmentation techniques such as generative adversarial networks can be further investigated to enhance model accuracy and robustness. Although I have tried to mimic real scenarios in the *in vitro* dataset by using real kidney stones for fragmentation, it has not been possible to capture every aspect of the clinical settings such as tissue background, proper flow of irrigation fluid, movement constraints on stone fragments, blood debris and other image artefacts. This possibly explains why the results and proposed framework is different for both *in vitro* and *in vivo* datasets.

Stone and laser localisation and segmentation can assist clinical stone fragmenting procedures. While the video frame rate of most ureteroscopy is 30 FPS, the proposed

framework achieves 12 FPS. As can be seen from Table 4.13, the DVFNet part of the framework requires over $6\times$ more computation time than that for a single frame. Also, Table 4.4 shows DVFNet during test inference does not significantly improve the results but adversely affects the inference time. The DVFNet part of the framework, when used in conjunction with the HybResUNet during training, helps to prune the segmentation mask by adjusting/refining the weights of the network layers of HybResUNet. Therefore, even though a combination of HybResUNet and DVFNet is required for training, only HybResUNet is used during test inference minimising the required time for the segmentation of laser and stone.

4.5 Conclusion

This study presents the first-ever attempt towards a multi-class segmentation method for ureteroscopy and laser lithotripsy imaging that is supported by a comprehensive experimental analysis. The proposed framework effectively makes use of residual connections and motion information between adjacent frames to produce robust and reliable segmentation of renal stones and laser fibre in real-time. The qualitative and quantitative results demonstrate that the presented algorithm can efficiently tackle the challenging vision quality within the kidney, resulting in improved segmentation performance as compared to the existing methods. Future research direction includes using different forms of recurrent neural networks in order to improve the temporal information and use it to further improve the segmentation results. In order to perform accurate size estimation, the present study serves as a preliminary work that needs to be extended to 3-D. The segmentation method can also be used in conjunction with depth estimation techniques in order to estimate parameters like the stone size. Future work also includes a quantitative assessment of the segmented stone fragments and laser fibre in order to help the clinician gain a better understanding of the target and improve patient outcomes. Through this study, I would also highlight that the clinical workload of endoscopists can be tackled by the development of medical image analysis tools that can shorten procedure time whilst improving diagnosis and therapy.

5

Use of VAEs beyond simple image generation

Contents

5.1	Introduction	95
5.2	Improved generalisation using VAE-embodied learning (UNet-eVAE)	97
5.2.1	Method	97
5.2.1.1	UNet module with VAE	97
5.2.1.2	Joint loss minimisation	99
5.2.2	Experiments and results	100
5.2.2.1	Implementation Details and setup.	100
5.2.2.2	Datasets.	100
5.2.2.3	Quantitative results	101
5.2.2.4	Qualitative results	103
5.2.3	Discussion	104
5.3	Unsupervised domain adaptation in endoscopy (EndoUDA)	106
5.3.1	Method	108
5.3.1.1	VAE training on source domain	108
5.3.1.2	Segmentation module	109
5.3.1.3	VAE latent search optimisation via new joint Loss	109
5.3.2	Experiments and results	110
5.3.2.1	Implementation Details	110
5.3.2.2	Datasets	111
5.3.2.3	Benchmark comparison and ablation study	111
5.3.2.4	Comparison with SOTA method	112
5.3.2.5	Effect of out-of-distribution data in supervised training	113
5.4	Conclusion	113

Machine learning models perform well when processing samples from the same distribution of training data but often suffer from the so-called domain shift when they encounter samples from different distributions, resulting in a deteriorated performance. VAEs are powerful generative models that have shown an incredible ability to generate highly realistic pieces of new data. This chapter presents the use of VAE-based generative models for unsupervised domain adaptation and generalisation. Unsupervised domain adaptation refers to the task of leveraging a fully annotated source dataset and an unlabelled target dataset with the goal of performing well on the latter. Generalisation refers to the task of learning a visual representation using the source domain only with the goal of performing well on the unseen domain. The chapter begins with the presentation of a novel VAE-embodied learning framework (UNet-eVAE) for improved generalisation of segmentation models. In this proposed approach, VAE is used to strengthen the feature representations to amplify features in the regions that contribute to the segmentation map generation from the UNet, thereby enhancing the segmentation performance on various ureteroscopy and colonoscopy datasets. This is followed by a subsection that presents a novel VAE-based target-independent unsupervised domain adaptation framework (EndoUDA) for improved segmentation on unseen Narrow-Band Imaging (NBI) modality images when trained with White Light Imaging (WLI) modality images for both Barrett’s oesophagus (BE) and polyp datasets.

Contribution/Originality: For the UNet-eVAE approach, I independently designed and implemented the proposed framework in Pytorch with the following original contributions: (1) proposed a novel end-to-end CNN framework wherein feedback from VAE enhances the ROI-specific feature representations of the segmentation task to provide improved segmentation performance and generalisation, (2) performed experimental validation on diverse publicly available endoscopy datasets, *in vitro* and *in vivo* ureteroscopy datasets to demonstrate the effectiveness of the proposed framework. I recorded the clinical procedures performed by Dr Ben Turney using a frame grabber to formulate the clinical datasets used in this study. I collected the animal datasets from Boston Scientific. I developed the experimental setup for acquiring *in vitro* (synthetic data) wherein Dr

Louise Goldsmith performed laser lithotripsy. I independently annotated the ureteroscopy datasets used in this study. In the EndoUDA approach, I contributed to the design of the proposed framework and loss function.

The UNet-eVAE work resulted in an oral presentation at the Machine Learning in Medical Imaging workshop held in conjunction with MICCAI 2022. The EndoUDA work resulted in a publication at MICCAI 2021.

Chapter 5- Gupta, Soumya, Sharib Ali, Ziang Xu, Binod Bhattarai, Ben Turney, and Jens Rittscher, "UNet-eVAE: Iterative refinement using VAE embodied learning for endoscopic image segmentation"- In *Machine Learning in Medical Imaging*, pp. 161–170. Springer, Cham, 2022.

Chapter 5- Celik, Numan, Sharib Ali, Soumya Gupta, Barbara Braden, and Jens Rittscher. "EndoUDA: a modality independent segmentation approach for endoscopy imaging." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 303-312. Springer, Cham, 2021.

5.1 Introduction

The lack of generalisable models is a significant concern in medical datasets such as ureteroscopy and colonoscopy. This is particularly due to the large differences in the type of endoscopes and instruments used, various clinical settings, differences in the anatomy of patients, and image artefacts. In addition, the quality of the data is primarily affected by unavoidable tissue movements, illumination changes, occlusions from bodily fluids, and significant variations in organ shape and appearance [53, 219]. These artefacts influence the visual interpretation of endoscopy data and may result in misdiagnosis or ineffective treatment. Artificial intelligence-based systems can play a significant role in improving the quality of diagnosis via automated segmentation of abnormalities. However, the lack of generalisation and domain adaptation capability of AI algorithms is a significant concern in medical datasets that can have significant variations in data. This chapter presents use cases of Variational autoencoder (VAE)-based generative models for improved generalisation and unsupervised domain adaptation of segmentation models.

The first framework, UNet-eVAE aims to improve the generalisation of segmentation models in ureteroscopy. In addition to ureteroscopy, this work also focuses on improving the generalisation of polyp segmentation in colonoscopy which is an essential diagnostic modality for colorectal cancer (CRC) screening with significantly high miss-rates for polyp detection [220].

It has been shown that by interconnecting the reconstruction and the segmentation networks, one can achieve sharper reconstructions and more accurate segmentation results [164]. Several works have extended the UNet-based encoder-decoder architecture [40] using dual-decoders [165], double encoder-decoder networks [166], residual blocks, and attention mechanisms [167, 168]. Most of these works extended UNet by stacking modules in series [166, 168] which increases network complexity and can adversely affect training on smaller datasets. VAEs have been successfully used for representation disentanglement and data generation [169]. Several works [170, 171] have proposed a shared encoder-based combination of UNet and VAE. However, these networks attempt to project the embedding to latent space for the entire images rather than a specific region of interest. Additionally, in [170], the ground truth mask is incorporated together with original image samples. However, this will constrain VAE representations towards the samples present in the majority of the training sets and can affect the generalisability.

To tackle such a problem and increase the bandwidth of the representations that the VAE can learn to boost segmentation, two major steps are taken: 1) reconstruct the region-of-interest after applying the predicted mask from the previous epoch and aim to minimise loss with the ground truth region-of-interest (ROI) area. 2) enrich the features of the UNet decoder layers by concatenating the features from various decoder layers of the VAE. Here, the former strategy helps to learn unseen representations during input and generalise to hard samples iteratively. At the same time, the latter strengthens the feature representations to amplify features in the regions that contribute to the segmentation map generation. Thus, this work introduces a VAE-embodied learning framework “UNet-eVAE” that combines VAE and UNet and proposes an iterative mechanism that efficiently enhances the segmentation performance. VAE is used for performing auxiliary reconstruction tasks for learning ROI-specific representations and provides a feedback

mechanism from the main segmentation task. Extensive experiments have been provided to demonstrate the performance improvement on test datasets of the same distribution and generalisability on out-of-sample datasets for colonoscopy and ureteroscopy datasets.

After demonstrating the success of VAEs in improving the generalisation of segmentation models, this chapter presents a subsection on another VAE-based framework for target-independent unsupervised domain adaptation in endoscopy. In addition to the conventional white light imaging (WLI) modality used in endoscopy, other techniques such as narrow-band imaging (NBI) have been developed to improve the early detection of cancer and its precursors [221]. Therefore, it is important to develop modality-agnostic approaches that can be directly integrated into the existing clinical workflow and generalise well on unseen data.

5.2 Improved generalisation using VAE-embodied learning (UNet-eVAE)

This section presents a detailed description of the proposed VAE-embodied learning framework (UNet-eVAE) for improved generalisation of segmentation in various endoscopy datasets. This is followed by the presentation and discussion of results obtained using the proposed method and state-of-the-art approaches.

5.2.1 Method

The proposed framework, as shown in Figure 4.4 is a novel integration of two sub-networks: UNet and VAE (“UNet-eVAE”), such that the VAE enhances ROI-specific feature representations as an auxiliary task, and the fine-grained spatial information from the VAE is fused with the UNet decoder. This technique enriches the feature representations and improves segmentation performance.

5.2.1.1 UNet module with VAE

The UNet is an encoder-decoder architecture. The input image $\mathbf{X} \in \mathbb{R}^{3 \times h \times w}$ is downsampled at consecutive encoder layers while learning feature representations through a series of convolution, batch normalisation and activation function. The decoder is then used

to upsample these features and predict segmentation map $\mathbf{y}_j \in \mathbb{R}^{h \times w}$ that corresponds to each pixel classification in $\mathbf{X}$ where j refers to j^{th} epoch. The skip connections (represented by dotted lines in Figure 4.4) help to recover spatial information lost during downsampling. The proposed approach follows a joint-optimisation strategy where image $\mathbf{X}$ is fed into the UNet model to obtain a segmentation prediction mask $\mathbf{y}_j$ for j^{th} epoch.

A variational autoencoder (VAE) has been used to perform reconstruction of the region-of-interest area obtained from UNet architecture. For this, the segmentation mask obtained from the previous epoch $\mathbf{y}_{j-1}^{th}$ is multiplied by the original image $\mathbf{X}$ to obtain $\mathbf{X}_{FG}$ ($= \mathbf{X} \odot \mathbf{y}_{j-1}$) as shown in Figure 4.4. Thus, a masked input $\mathbf{X}_{FG}$ is compared with the ground truth masked input $\mathbf{X}_{FG}^{gt}$ ($= \mathbf{X} \odot \mathbf{y}^{gt}$) to constrain VAE reconstruction. The presented auxiliary task enriches feature representations used to obtain segmentation masks by further concatenating them with the UNet decoder layers. As illustrated in Figure 4.4, four connections from the VAE-decoder are fused with the segmentation-decoder. Since the initial decoder layers have small spatial details ($1^2, 4^2, 8^2$), the latter layers are chosen ($16^2, 32^2, 64^2, 128^2$) for concatenation to enrich the fine-grained feature representation of the region-of-interest. Let $q_\phi(\mathbf{Z}|\mathbf{X}_{FG})$ be the latent space distribution of the sample $\mathbf{X}_{FG}$. $\mathbf{Z}$ (with parameters $\mu_{\mathbf{Z}}$ and $\sigma_{\mathbf{Z}}$) sampled from this distribution is fed

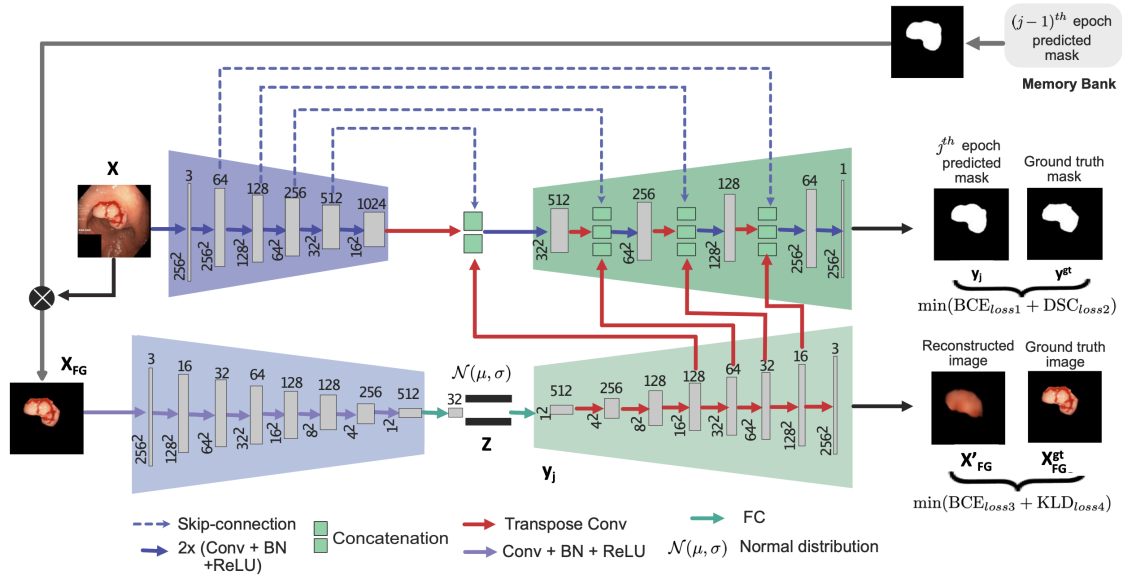


Figure 5.1: Proposed UNet2VAE framework: Here, four connections from the VAE-decoder are fused with the segmentation-decoder. Since the initial decoder layers have small spatial details ($1^2, 4^2, 8^2$), latter layers are chosen ($16^2, 32^2, 64^2, 128^2$).

β -value	DSC	mIOU	Precision	Recall
1	0.686 ± 0.254	0.574 ± 0.275	0.636 ± 0.309	0.876 ± 0.185
5	0.724 ± 0.231	0.612 ± 0.252	0.691 ± 0.280	0.867 ± 0.192
10	0.786 ± 0.225	0.693 ± 0.250	0.822 ± 0.247	0.824 ± 0.224
15	0.717 ± 0.260	0.614 ± 0.275	0.688 ± 0.303	0.853 ± 0.227

Table 5.1: Ablation study for hyper-parameter β

into the decoder, $p_\theta(\mathbf{X}_{FG}^{gt}|\mathbf{Z})$. The loss function of the VAE is given by:

$$\mathcal{L}(\theta, \phi) = -\mathbb{E}_{q_\phi}[\ln(p_\theta(\mathbf{X}_{FG}^{gt}|\mathbf{Z}))] + \mathbb{KL}(q_\phi(\mathbf{Z}|\mathbf{X}_{FG})||p(\mathbf{Z})) \quad (5.1)$$

5.2.1.2 Joint loss minimisation

The proposed model aims to jointly optimise loss functions responsible for the main segmentation task and an auxiliary reconstruction task. For the segmentation task, I have used a combination of binary cross-entropy (BCE) BCE_{loss1} and Dice similarity coefficient DSC_{loss2} [222] losses that are calculated between the predicted mask $\mathbf{y}_j$ and its ground truth mask $\mathbf{y}^{gt}$. The proposed loss for segmentation $\mathcal{L}_{seg}$ is given by:

$$\mathcal{L}_{seg} = \underbrace{\frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i^{gt}(\log(\mathbf{y}_{ji})) + (1 - \mathbf{y}_i^{gt})(\log(1 - \mathbf{y}_{ji}))) + 1}_{BCE_{loss1}} - \underbrace{\frac{2 * \sum_{i=1}^N (\mathbf{y}_i^{gt} * \mathbf{y}_{ji})}{\sum_{i=1}^N \mathbf{y}_i^{gt} + \sum_{i=1}^N \mathbf{y}_{ji}}}_{DSC_{loss2}}, \quad (5.2)$$

where N is the total number of pixels, $\mathbf{y}^{gt}$ is the label (1 or 0) and $\mathbf{y}^j$ is the predicted probability. For the auxiliary reconstruction task, a combination of BCE reconstruction loss BCE_{loss3} and KL-divergence KLD_{loss4} is used and denoted by $\mathcal{L}_r$. Here, BCE_{loss3} is calculated between the reconstructed image from VAE network $\mathbf{X}'_{FG}$ and the ground truth image $\mathbf{X}_{FG}^{gt}$. $\mathcal{L}_r$ is given by:

$$\mathcal{L}_r = \underbrace{\frac{1}{N} \sum_{i=1}^N (\mathbf{X}_{FGi}^{gt}(\log(\mathbf{X}'_{FGi})) + (1 - \mathbf{X}_{FGi}^{gt})(\log(1 - \mathbf{X}'_{FGi})))}_{BCE_{loss3}} - \underbrace{0.5\beta \sum_{k=1}^Z (1 + \log \sigma_k^2 - \sigma_k^2 - \mu_k^2)}_{KLD_{loss4}}, \quad (5.3)$$

where $\beta = 10$ (see ablation study in Table 5.1) is used to weight KLD_{loss4} . The final joint loss $0.5 \times (\mathcal{L}_{seg} + \mathcal{L}_r)$ is optimised using Adam optimiser with learning rate of $1e^{-4}$.

5.2.2 Experiments and results

This section presents a description of the datasets used and network implementation details of this study. This is followed by a comparison of the quantitative and qualitative results of the proposed network against state-of-the-art approaches.

5.2.2.1 Implementation Details and setup.

The experiments in this work were implemented using the PyTorch framework and trained on NVIDIA Quadro RTX 6000. All image samples were resized to $256 \times 256 \times 3$ pixels and trained using Adam optimiser with a learning rate of $1e^{-4}$ and batch sizes of 4 and 32 were used, respectively, for the Ureteroscopy dataset and Kvasir-SEG dataset. In the Adam optimiser, initial decay rates were set to default 0.9 and 0.999 for estimation of the first and second moments of gradient computations, respectively. ReduceLROnPlateau callback was used to monitor the learning rate and tune it to obtain optimal performance. Validation was performed after each epoch. Early Stopping was also used to monitor validation loss and stop the training when no improvement is observed.

For UNet-eVAE, the previous epoch segmentation masks $\mathbf{y}_{j-1}$ were initialised using Otsu thresholding [30] of original image $\mathbf{X}$ during the first training epoch, and the same was done for the model inference. Standard medical image segmentation metrics are used, such as mean intersection over union (mIoU), mean Dice coefficient (DSC), precision, and recall.

5.2.2.2 Datasets.

Both publicly available datasets and in-house datasets have been used in this study. For the polyp segmentation task in colonoscopy, publicly available Kvasir-SEG [223] is used for training, validation, and testing (same distribution). The Kvasir-SEG dataset consists of 1000 polyp images that were randomly split into a training set (80%), validation set (10%), and testing set (10%) with a random seed at 42. To assess the generalisability of the proposed method, two other publicly available datasets that include CVC-ClinicDB [224] and PolypGen-C6 [225] for out-of-sample (unseen distribution) tests have been used. CVC-ClinicDB consists of 66 samples, and PolypGen-C6 consists of 88 samples.

For the ureteroscopic laser lithotripsy, *in vitro* and *in vivo* datasets collected in-house are used. The ureteroscopic laser lithotripsy *in vitro* dataset was acquired under controlled settings. The *in vitro* dataset consists of 70 training samples, nine validation samples, and nine test samples. The *in vivo* dataset consists of 136 clinically acquired images from procedures on humans. This dataset was split into a training set (80%), validation set (10%), and testing set (10%) with a random seed at 42. To assess generalisability, 18 samples have been used from an animal study (performed at Boston Scientific sites) as an out-of-sample dataset to test.

5.2.2.3 Quantitative results

Table 5.2 presents quantitative results on various test sets wherein dataset 1 represents Kvasir-SEG, while dataset 2 and dataset 3 represent ureteroscopy test samples with inference from checkpoints that were trained on *in-vitro* and *in-vivo*, respectively. For dataset 1, results of training on 50% samples and all 900 samples are presented and compared with the proposed model and SOTA approaches such as Probabilistic UNet [170], DDANet [165] and UNet [40]. It can be observed that when trained with 450 samples only, our method provided a DSC improvement of about 24.5%, 10.4%, and 7.26%, on the Kvasir-SEG test set(100 samples) as opposed to Probabilistic UNet, DDANet and UNet, respectively. It can also be seen that the proposed approach showed improved generalisability with a 23%, and 8.13% DSC improvement on out-of-sample CVC-ClinicDB (66 samples) and PolypGen-C6 (88 samples), respectively as opposed to UNet. Similarly, in the case of 900 samples, the proposed approach provided a DSC improvement of 33.8%, 2.9%, and 1.7% over Probabilistic UNet, DDANet, and UNet respectively on the Kvasir-SEG test set. It can also be observed that the proposed method showed improved generalisability with 3.47% DSC improvement on the out-of-sample CVC-ClinicDB dataset. It can also be observed that the performance of existing SOTA approaches significantly lowers when the training set is halved while the proposed method still performs comparatively well. It can also be seen in this section that the segmentation results of the proposed method have a lower standard deviation across all datasets for

Network	Class	Train set	Test set	DSC	mIOU	Precision	Recall
Dataset 1 – Colonoscopy for polyp segmentation: Trained on Kvasir-SEG							
Probabilistic UNet [170]	Polyp	Kvasir-SEG (450)	Kvasir-test	0.5833 ± 0.2838	0.457 ± 0.2633	0.774 ± 0.2675	0.5782 ± 0.3151
DDANet [165]	Polyp	Kvasir-SEG (450)	Kvasir-test	0.6577 ± 0.2966	0.5552 ± 0.3002	0.7647 ± 0.3055	0.6493 ± 0.3221
UNet [40]	Polyp	Kvasir-SEG (450)	Kvasir-test	0.6772 ± 0.2873	0.5720 ± 0.2813	0.7711 ± 0.2580	0.7235 ± 0.3107
			CVC-ClinicDB	0.3686 ± 0.3314	0.2828 ± 0.2826	0.5572 ± 0.3824	0.5430 ± 0.4356
			PolypGen-C6	0.3689 ± 0.3428	0.2874 ± 0.2952	0.4902 ± 0.3941	0.5376 ± 0.4237
Ours	Polyp	Kvasir-SEG (450)	Kvasir-test	0.7264 ± 0.2699	0.6278 ± 0.2757	0.7543 ± 0.2853	0.7916 ± 0.2683
			CVC-ClinicDB	0.4534 ± 0.3148	0.3503 ± 0.2840	0.4255 ± 0.3370	0.6919 ± 0.3858
			PolypGen-C6	0.3989 ± 0.3385	0.3131 ± 0.3069	0.4046 ± 0.3709	0.6660 ± 0.3945
Probabilistic UNet [170]	Polyp	Kvasir-SEG (900)	Kvasir-test	0.5877 ± 0.287	0.4669 ± 0.2781	0.7916 ± 0.2612	0.6323 ± 0.3327
DDANet [165]	Polyp	Kvasir-SEG (900)	Kvasir-test	0.7641 ± 0.2524	0.6711 ± 0.2641	0.7701 ± 0.2655	0.8302 ± 0.2553
UNet [40]	Polyp	Kvasir-SEG (900)	Kvasir-test	0.7731 ± 0.2770	0.6933 ± 0.2846	0.8490 ± 0.2581	0.7882 ± 0.2770
			CVC-ClinicDB	0.4716 ± 0.3731	0.3902 ± 0.3383	0.7487 ± 0.3316	0.5249 ± 0.4190
			PolypGen-C6	0.4775 ± 0.3967	0.4101 ± 0.3747	0.7504 ± 0.3484	0.5283 ± 0.4206
Ours	Polyp	Kvasir-SEG (900)	Kvasir-test	0.7863 ± 0.2248	0.6934 ± 0.2501	0.8220 ± 0.2474	0.8241 ± 0.2242
			CVC-ClinicDB	0.4880 ± 0.3052	0.3803 ± 0.2896	0.5014 ± 0.3565	0.6806 ± 0.3660
			PolypGen-C6	0.4582 ± 0.3603	0.3750 ± 0.3380	0.4741 ± 0.3850	0.6591 ± 0.4005
Dataset 2 – Ureteroscopic laser lithotripsy (In-vitro)							
UNet [40]	Stone	<i>In vitro</i>	<i>In vitro</i>	0.9281 ± 0.0276	0.8670 ± 0.0476	0.9199 ± 0.0675	0.9414 ± 0.0339
Ours	Stone	<i>In vitro</i>	<i>In vitro</i>	0.9444 ± 0.0307	0.8962 ± 0.0537	0.9468 ± 0.0529	0.9441 ± 0.0294
UNet [40]	Laser	<i>In vitro</i>	<i>In vitro</i>	0.7307 ± 0.3045	0.6474 ± 0.3035	0.7704 ± 0.3318	0.8253 ± 0.1726
Ours	Laser	<i>In vitro</i>	<i>In vitro</i>	0.7671 ± 0.3120	0.6998 ± 0.3095	0.7638 ± 0.3056	0.8844 ± 0.1758
Dataset 3 – Ureteroscopic laser lithotripsy (In-vivo) : Trained on clinical							
UNet [40]	Stone	<i>In vivo</i> (clinical)	<i>In vivo</i> (clinical)	0.8864 ± 0.0748	0.8033 ± 0.1096	0.9146 ± 0.0567	0.8757 ± 0.1312
			<i>In vivo</i> (animal)	0.4224 ± 0.3272	0.3236 ± 0.2733	0.9281 ± 0.1054	0.3543 ± 0.3025
Ours	Stone	<i>In vivo</i> (clinical)	<i>In vivo</i> (clinical)	0.9004 ± 0.0607	0.8238 ± 0.0914	0.9294 ± 0.0372	0.8832 ± 0.1113
			<i>In vivo</i> (animal)	0.4458 ± 0.3041	0.3354 ± 0.2515	0.8385 ± 0.2442	0.3681 ± 0.2797
UNet [40]	Laser	<i>In vivo</i> (clinical)	<i>In vivo</i> (clinical)	0.9358 ± 0.0321	0.8811 ± 0.0550	0.9466 ± 0.0431	0.9304 ± 0.0663
Ours	Laser	<i>In vivo</i> (clinical)	<i>In vivo</i> (clinical)	0.9313 ± 0.0359	0.8734 ± 0.0617	0.9127 ± 0.0599	0.9539 ± 0.0403

Table 5.2: Quantitative evaluation of segmentation results obtained from UNet and proposed UNet-eVAE (ours) trained on Kvasir-SEG, *in vitro* and *in vivo* ureteroscopic laser lithotripsy datasets

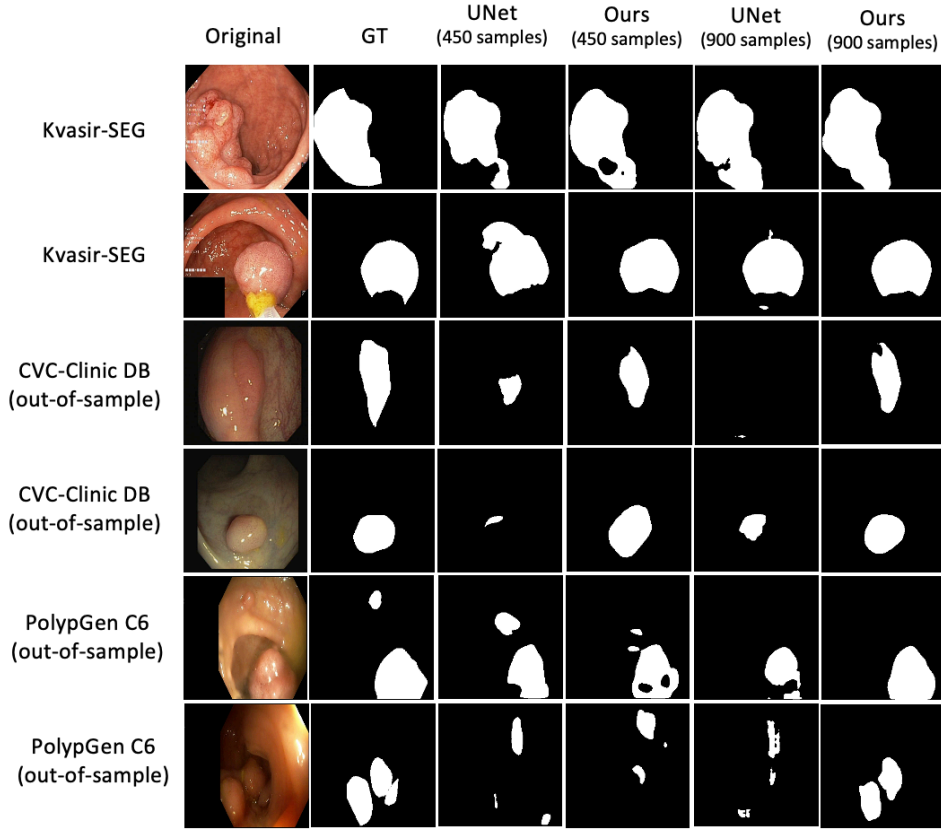


Figure 5.2: Qualitative evaluation of segmentation results obtained from UNet and UNet2VAE trained on Kvasir-SEG dataset

polyp segmentation as compared to the results obtained from SOTA. Similarly, for Dataset-2 in Table 5.2, it can be observed that the proposed method boosted the DSC of both stone and laser classes by 1.76% and 4.98%, respectively. Finally, in the third section of Table 5.2 (refer to Dataset 3) for which results of models trained using 136 *in vivo* (clinical) samples, it can be seen that for the stone class, the proposed method provided a DSC improvement of 1.6% on the clinical test set (14 samples) and also improved the generalisation on out-of-sample animal data (18 samples) by 5.53% as opposed to SOTA UNet. For the laser class, our model provided competitive results to UNet.

5.2.2.4 Qualitative results

Figure 5.2 illustrates the segmentation results of our model trained on the Kvasir-SEG dataset as opposed to the SOTA baseline UNet model. The first two rows in the figure show the network inference results on Kvasir-SEG, while the next four rows illustrate the

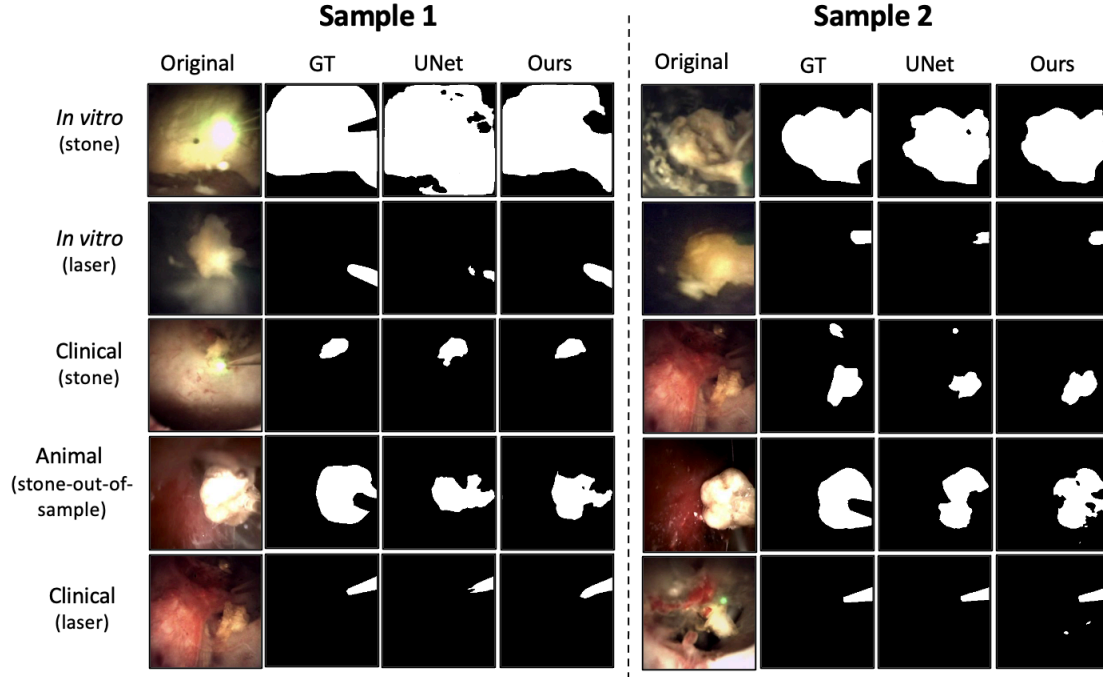


Figure 5.3: Qualitative evaluation of segmentation results obtained from UNet and UNet2VAE trained on *in vitro* and *in vivo* ureteroscopic laser lithotripsy datasets.

network inference on the out-of-sample CVC-ClinicDB dataset and PolypGen-C6 dataset. The proposed model significantly outperforms the UNet model on the Kvasir-SEG test set and generalises well on both out-of-sample datasets compared to the UNet model. It can also be seen from Figure 5.2 that the performance of the proposed model is better than the UNet even when the number of training samples is reduced to half. Figure 5.3 illustrates results on ureteroscopic datasets for two samples for each provided case in the rows. The first two rows show segmentation results of stone and laser classes in the *in vitro* dataset. It can be observed that the proposed method yields closer to ground truth masks compared to the SOTA UNet model. Similar results can be observed in the third and fourth rows for the stone class in the clinical and animal datasets (out-of-sample), respectively. UNet-eVAE provided competitive segmentation performance on clinical samples even for the laser class.

5.2.3 Discussion

This work presents a novel end-to-end trainable method that uses a VAE as an auxiliary network to enhance ROI-specific feature representations that are combined and optimised

DeepResUNet (Baseline)	DSC	IoU	Precision	Recall
DeepResUNet[19]	0.604 ± 0.304	0.496 ± 0.293	0.773 ± 0.292	0.605 ± 0.333
DeepResUNet-eVAE (Ours)	0.689 ± 0.264	0.580 ± 0.276	0.726 ± 0.320	0.766 ± 0.249

Table 5.3: Quantitative comparison of baseline DeepResUNet against DeepResUNet in the proposed configuration(DeepResUNet-eVAE) when trained with 900 samples of Kvasir-SEG

with a UNet for endoscopic image segmentation tasks. Several works in the literature have proposed various shared encoder-based approaches such as Probabilistic UNet [170] and DDANet [165]. As can be observed in Table 5.2, the proposed framework significantly improved the segmentation performance on Kvasir-SEG and ureteroscopy datasets as opposed to state-of-the-art approaches. For example, when trained with 450 samples of Kvasir-SEG, UNet-eVAE provided a DSC improvement of 24.5%, 10.4%, and 7.26% on Kvasir-SEG as opposed to Probabilistic UNet, DDANet, and UNet, respectively. Also, a generalisation improvement can be observed on out-of-sample CVC-ClinicDB and PolypGen-C6 with a DSC boost of 23% and 8.13%, respectively when the network was trained with 450 samples of Kvasir-SEG. Similar improvements can be observed when 900 samples of Kvasir-SEG were used for training. In addition, it can be observed that the performance of UNet-eVAE drops only 8.2% while the performance of DDANet and UNet drop by 16.2% and 14.2%, respectively when the number of Kvasir-SEG training samples were reduced by half. This shows that the proposed UNet-eVAE is robust and can effectively learn to provide reliable segmentation even when the training dataset is small. Also, polyp segmentation results from UNet-eVAE can be seen to have a lower standard deviation across most metrics as compared to state-of-the-art methods. Similarly, in the case of *in vitro* ureteroscopy dataset, UNet-eVAE boosted the DSC of both stone and laser classes. In addition, the proposed method not only improved the DSC of stone class on the clinical test set but also improved the generalisation on out-of-sample animal data as opposed to state-of-the-art UNet. The added strength of the proposed method is that it can boost the performance of any baseline network independent of its architecture. DeepResUNet [122] with our technique yielded a 14% and 17% boost in DSC and IOU,

respectively against its baseline(see Table 5.3) when trained with 900 samples of Kvasir-SEG. It can also be observed in Table 5.2 that the proposed method outperforms SOTA across all metrics except Precision in most cases. However, in most medical imaging scenarios, the cost of False Negatives is high, indicating recall is a better measure than precision[226], which can be observed to be higher for the proposed framework.

5.3 Unsupervised domain adaptation in endoscopy (EndoUDA)

To develop a modality-agnostic approach, an unsupervised domain adaptation (UDA) technique has been proposed that can be trained on an available single WLI modality dataset and can be applied to unseen NBI imagery for both BE and polyp datasets.

Well-established Convolutional Neural Network (CNN) based approaches [40, 41, 227] have been applied to segmentation tasks in GI endoscopy [65, 228, 229]. Wu *et al.* [229] proposed a dual neural network based on UNet with a ResNet backbone for automatic oesophageal lesion segmentation. Guo *et al.* [228] utilised an FCN-based architecture with atrous kernels to segment polyps. However, all these supervised approaches are trained on WLI endoscopy and do not support working in different imaging modalities. To date, the problem of quantifying disease-related changes in NBI, which can produce a very different appearance of the mucosal surface when compared to WLI (see Fig. 5.4) has not been addressed. Training methods for different imaging modalities independently is prohibitive. Firstly, extensive training datasets that cover a similar range of conditions would need to be curated to ensure a comparable performance. Secondly, manual annotations by human experts are required which is not only labour-intensive but also expensive. To tackle these issues domain adaptation methods have been used to adapt trained models to unseen target data for segmentation [230–233]. Although these methods promise to generalise, so far there are only a few examples of applying such methods to medical data [231, 232]. Additionally, the available methods still underperform and have not been adapted for extremely diverse modality cases such as that in GI endoscopy.

This section presents a detailed description of the unsupervised domain adaptation in endoscopy - EndoUDA framework. This is followed by the presentation and discussion of

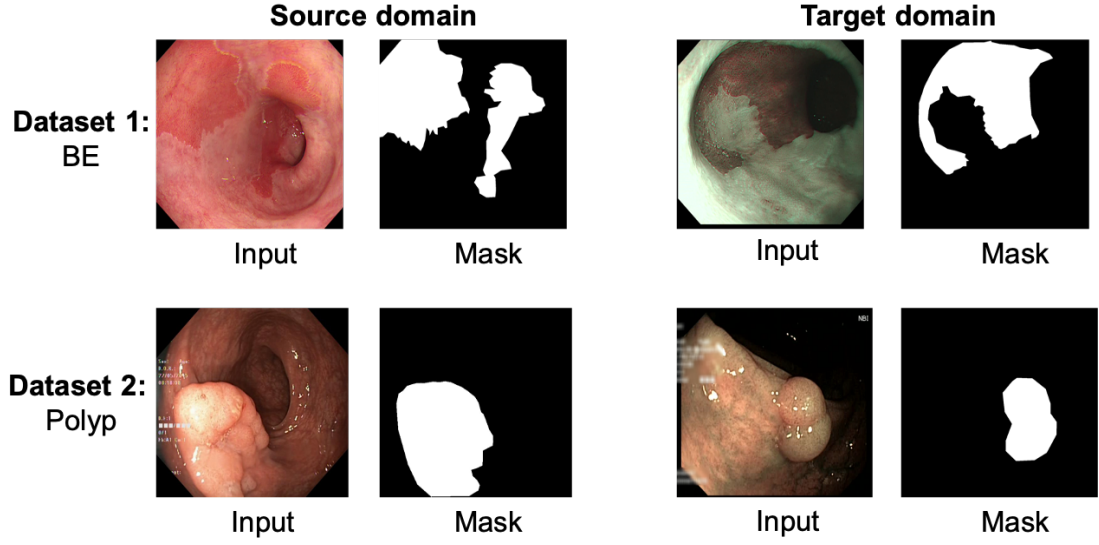


Figure 5.4: Example images and masks of the two GI datasets for which source and target domains are separated based on their modality.

results obtained using the proposed method and state-of-the-art approaches. The figures and text presented in this section have been taken from its published version [234].

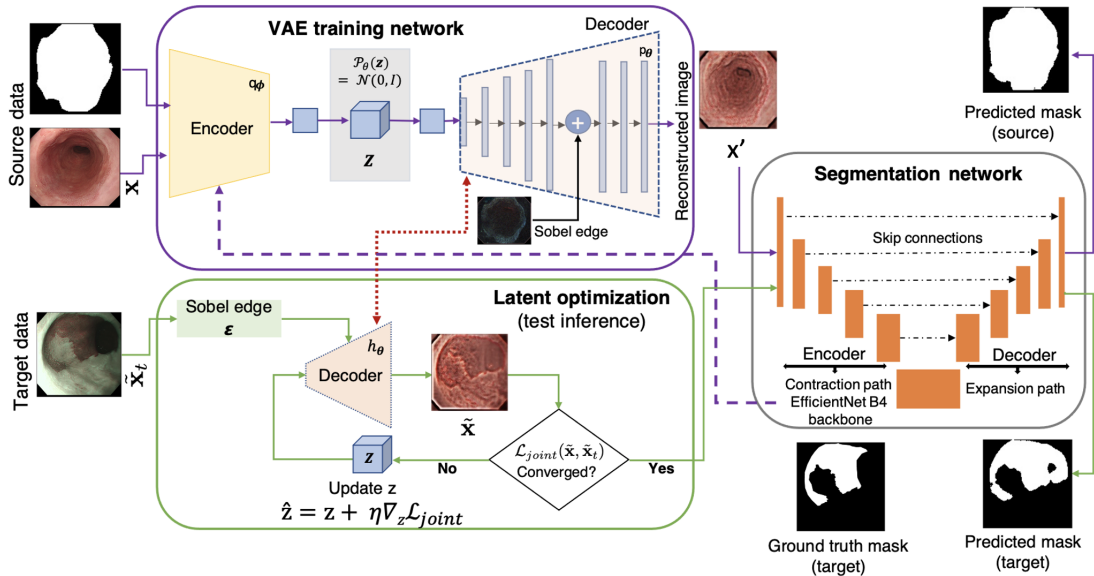


Figure 5.5: Complete architecture of the proposed EndoUDA: a target-independent unsupervised domain adaptation wherein VAE module allows learning the source domain representations while the UNet module leverages the learned encoded features for semantic segmentation. [234]

5.3.1 Method

The proposed EndoUDA consists of components for training and test time inference shown in Figure 5.5. The training network utilises a shared EfficientNet-B4 encoder for variational autoencoder (VAE) and a UNet-based segmentation model. The VAE module allows learning the source domain representations while the UNet module leverages the learned encoded features for semantic segmentation. For inference at test time, learnt latent space embeddings are further optimised to translate to the unseen modality dataset by continuously updating via a joint loss minimisation scheme. Below, details of each of these modules are presented.

5.3.1.1 VAE training on source domain

The loss function of the Variational autoencoder is a combination of negative log-likelihood and a Kulback-Leibler (KL) divergence-based regulariser expressed in Equation 2.4. Compared to the VAE framework in [233], we propose to couple the encoder layer of VAE as the segmentation network backbone. Based on the experiments, EfficientNet which computes the approximate posterior $q_\phi(\mathbf{Z}|\mathbf{X})$ has been chosen. Unlike Pandey *et al.* [233], the originally used perceptual loss module has been removed in the decoder layer used to compute the image likelihood of the observed data $p_\theta(\mathbf{X}|\mathbf{Z})$. A Sobel filter-generated edge map is concatenated with the sixth layer of the decoder network of the VAE through a $\tanh$ nonlinear function (see Fig. 5.5). The reconstruction loss $L_r(\mathbf{X})$ is minimised for the VAE training network given by:

$$L_r(\mathbf{X}) = L(\theta, \phi) = \frac{1}{N} \sum_{i=1}^N \|\mathbf{X}^i - p_\theta(q_\phi(\mathbf{X}^i))\|^2, \quad (5.4)$$

where $\mathbf{X}$ is the source domain image, N is total number of samples, q_ϕ and p_θ are encoder and decoder networks, respectively. Here, the reconstructed image is given by $\mathbf{X}' = p_\theta(q_\phi(\mathbf{X}))$. In addition, as shown in Fig. 5.5, the edges of the input images are extracted and concatenated into the decoder network in VAE through a skip connection to better generalise the source and target domains. Fig. 5.5 shows a schematic diagram of the complete VAE architecture used for training on the source domain.

5.3.1.2 Segmentation module

The segmentation network consists of a UNet with an EfficientNet-B4 [235] backbone. The input from the reconstructed image obtained from the decoder of the VAE is fed to the already-tuned weights of EfficientNet-B4. The segmentation network is adapted by a supervised loss on the source domain given by a classically used binary cross-entropy loss (L_{BCE}) and Dice coefficient loss (L_{DSC}) [222]. L_{DSC} is represented as $L_{DSC} = (2 \sum \mathbf{y} \mathbf{y}^{gt}) / (\sum \mathbf{y}^2 + \sum \mathbf{y}^{gt^2})$, where $\mathbf{y}$ is the predicted probability map and $\mathbf{y}^{gt}$ is the ground truth mask.

5.3.1.3 VAE latent search optimisation via new joint Loss

The trained VAE encoding aims to find the nearest point from the source domain, given a sample from the target domain, through an optimisation process and updates of the latent space representation $\mathbf{Z}$, as can be seen in Fig. 5.5. Once the decoder h_θ of the VAE is trained on the source distribution $p_s(\mathbf{X})$, given an image $\tilde{\mathbf{X}}_t$ from the target distribution, the latent search algorithm finds the nearest point by optimising latent vector over the iterative process. As shown in Fig. 5.4, the WLI images are used in the source domain while the NBI is used in the target domain. For faster optimisation convergence of the latent space encoding with improved generalisation ability of the network, a joint loss function has been proposed that comprises the normalised cross-correlation (NCC) (L_{NCC}) and the structural similarity index measure (SSIM) [236] loss (L_{ssim}) between the image $\tilde{\mathbf{X}}_t$ from target domain and the reconstructed target image $\tilde{\mathbf{X}}$ from h_θ . The loss function for computing NCC is defined as:

$$L_{NCC}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_t) = \frac{1}{2N} \sum \left(\frac{\tilde{\mathbf{X}} - \mu_s}{\sqrt{\sigma_s^2 + \epsilon^2}} - \frac{\tilde{\mathbf{X}}_t - \mu_t}{\sqrt{\sigma_t^2 + \epsilon^2}} \right)^2, \quad (5.5)$$

where $\tilde{\mathbf{X}}$, μ_s and σ_s are reconstructed images from the decoder network, mean and standard deviation of the latent encoded image, respectively. Similarly, $\tilde{\mathbf{X}}_t$, μ_t and σ_t correspond to the target image input, mean and standard deviation of the target image. The SSIM for a pair of images ($\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_t$) and loss function L_{ssim} is defined as:

$$\begin{aligned} SSIM(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_t) &= l(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_t)^A * c(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_t)^B * s(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_t)^C, \\ L_{ssim}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_t) &= 1 - SSIM(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_t), \end{aligned} \quad (5.6)$$

here l , c and s denote luminance, contrast and structure similarities of the given pair of images, respectively with the parameters A , B and $C > 0$. Also, the similarity loss function (L_{sim}) is presented. The final joint loss function is given by:

$$L_{joint} = \Phi L_{NCC} + (1 - \Phi) L_{sim}, \quad (5.7)$$

where $\Phi = 0.75$ is set empirically. After the convergence of (L_{joint}), the optimal latent space representation $\hat{\mathbf{Z}} = \mathbf{Z} + \eta L_{joint}$ is used to generate the closest clone of source domain with a learning rate η which is then used as an input to the segmentation network to predict the target mask.

5.3.2 Experiments and results

This section presents a description of the network implementation and datasets used in this study. This is followed by a benchmark comparison, an ablation study, and an extensive comparison of the proposed approach against state-of-the-art approaches. Finally, the effect of out-of-distribution data in supervised training is also investigated.

5.3.2.1 Implementation Details

RMSProp with a learning rate of 0.0001 was used for the training of the VAE network due to its fast convergence speed while for the segmentation module, Adam optimiser was used with $\beta_1=0.5$, $\beta_2=0.999$, and $lr = 0.001$. The initial learning rate was reduced by a factor of 10 for every validation loss that did not improve in the last 3 epochs and a stopping criterion was set for no improvement in validation loss up to 10 epochs. 100 epochs were used for training the EndoUDA training module with a batch size of 64. For the test module of EndoUDA, a learning rate of $\eta = 0.001$ was set for optimisation of the latent space representation and a fixed 40 iterations with a batch size of 32 was set. The proposed method was implemented using the TensorFlow framework on a machine deployed with NVIDIA RTX 2080Ti graphics card with 12 GB GPU memory. All input images were resized to 128×128 pixels.

Models	BE		Polyp	
	IoU	Dice	IoU	Dice
UNet[40]	0.738	0.846	0.718	0.755
UNet++[237]	0.742	0.859	0.721	0.778
ResUNet[128]	0.789	0.908	0.734	0.794
EfficientUNet[238]	0.841	0.931	0.803	0.896
EfficientUNet++	0.828	0.927	0.767	0.865
DeepLabv3+ [41]	0.802	0.893	0.789	0.869

Table 5.4: Supervised Segmentation analysis: Benchmarking BE and polyp datasets on standard segmentation architectures on 80-20 train-test split. [234]

5.3.2.2 Datasets

Two gastrointestinal (GI) endoscopy datasets have been used for analysing two distinct precancerous anomalies, namely, Barrett’s oesophagus (BE) and polyps. To evaluate the efficacy of the proposed EndoUDA method, clinically acquired white light imaging (WLI) and narrow-band imaging (NBI) modality data for BE and polyps has been used. BE dataset consists of 1182 endoscopy images acquired from 68 unique patients of which 611 WLI images are used as source domain data (train set: 90%, validation set: 10%) and 571 NBI images as target domain data are used for testing. Similarly, for the polyp dataset, the publicly available Hyper-Kvasir dataset [56] was used for training which consists of 1000 images (train set: 80%, validation set: 20%). In addition, we used clinically acquired 42 NBI images of polyps acquired from 20 patients at our local hospital and from collaborators.

5.3.2.3 Benchmark comparison and ablation study

Results of supervised segmentation of BE and polyps in the GI dataset are presented in Table 5.4. Frequently used supervised learning methods have been evaluated with respect to IoU and Dice to establish a baseline. It can be observed that the UNet with an EfficientNet B4 backbone [238] outperformed other baseline methods including DeepLabv3+ [41] by nearly 4% and 2% on both IoU and Dice for BE and polyps, respectively.

Models	BE ($\mu \pm \sigma$)				Polyp ($\mu \pm \sigma$)			
	IoU	Dice	Precision	Recall	IoU	Dice	Precision	Recall
Naive UNet [238]	0.626	0.718	0.654	0.667	0.492	0.564	0.487	0.496
	± 0.02	± 0.015	± 0.018	± 0.021	± 0.018	± 0.012	± 0.011	± 0.011
AdaptSegnet [230]	0.658	0.749	0.721	0.698	0.519	0.577	0.601	0.583
	± 0.019	± 0.016	± 0.012	± 0.022	± 0.017	± 0.015	± 0.016	± 0.012
ADVENT [239]	0.667	0.768	0.739	0.706	0.524	0.591	0.662	0.638
	± 0.012	± 0.014	± 0.015	± 0.012	± 0.011	± 0.019	± 0.021	± 0.014
CellSegUDA [232]	0.673	0.771	0.733	0.691	0.533	0.629	0.688	0.641
	± 0.02	± 0.016	± 0.018	± 0.013	± 0.012	± 0.021	± 0.018	± 0.017
GLSS [233]	0.704	0.815	0.846	0.735	0.558	0.621	0.675	0.649
	± 0.019	± 0.022	± 0.019	± 0.02	± 0.017	± 0.016	± 0.019	± 0.012
EndoUDA	0.733	0.854	0.832	0.784	0.605	0.693	0.722	0.704
	± 0.014	± 0.021	± 0.017	± 0.019	± 0.014	± 0.02	± 0.017	± 0.018

Table 5.5: Empirical results of EndoUDA along with SOTA UDA methods with mean μ and standard deviation σ are provided. All comparisons are provided for source-only trained models and tested on target data. [234]

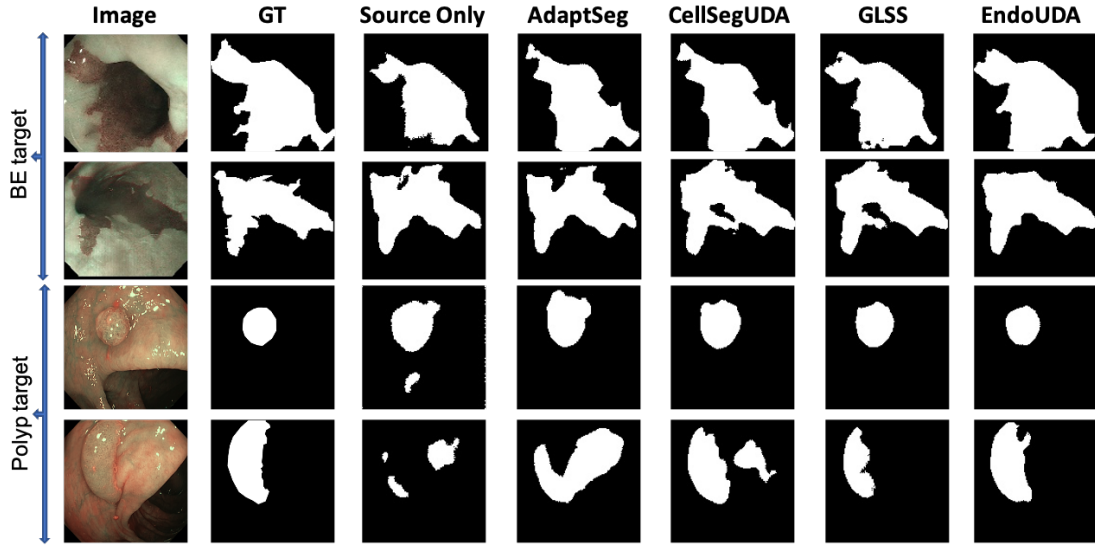


Figure 5.6: Qualitative study: Baseline UDA experiments on BE and polyp images. [234]

5.3.2.4 Comparison with SOTA method

The proposed EndoUDA approach has been compared with four SOTA UDA methods: AdaptSegnet[230], ADVENT[239], CellSegUDA[232], GLSS[233], along with the naive supervised baseline model. The quantitative results are reported in Table 5.5 for four different metrics including intersection-over-union (IoU), Dice, precision and recall metrics. It can be observed that the proposed framework EndoUDA outperformed other SOTA UDA methods and the naive UNet method. An improvement of nearly 4% on

Method	BE_{IoU}	$Polyp_{IoU}$	BE_{Dice}	$Polyp_{Dice}$
UNet (source-only)	0.626	0.492	0.718	0.564
EndoUDA	0.733	0.605	0.854	0.693
UNet (source 100% + target 10%)	0.646	0.536	0.738	0.588
EndoUDA (source 100% + target 10%)	0.744	0.613	0.866	0.706
UNet (source 100% + target 25%)	0.691	0.579	0.761	0.641
EndoUDA (source 100% + target 25%)	0.756	0.624	0.875	0.721
UNet (source 100% + target 50%)	0.738	0.631	0.827	0.734
EndoUDA (source 100% + target 50%)	0.771	0.658	0.889	0.739
UNet (target-trained)	0.841	0.802	0.931	0.896

Table 5.6: IoU and Dice comparisons for out-of-distribution target samples (NBI) in the source (WLI) GI endoscopy datasets [234]

BE and 7% on polyp dataset for with respect to the Dice metric compared to the recent SOTA GLSS method. Similar improvements can be seen for other metrics for both GI datasets and also much higher improvements compared to other SOTA UDA methods. It can be seen in Fig. 5.6 that the predicted masks using the proposed EndoUDA model for BE and polyp images are very close to the ground truth annotations compared to those predicted by other SOTA UDA methods.

5.3.2.5 Effect of out-of-distribution data in supervised training

Further investigation of the performance of EndoUDA compared to the supervised baseline UNet model (with EfficientNet backbone) was performed by training the model with 10%, 25% and 50% of target data samples. It can be observed (see Table 5.6) that the EndoUDA performs consistently better than the baseline approach for a different proportion of target data included in the training. For example, on 10% and 25% of target data mixed in the training sample, EndoUDA still provides nearly a 7-10% improvement on IoU and 8-13% with respect to Dice scores over the baseline method.

5.4 Conclusion

This thesis chapter presented two use cases for VAE-based generative models beyond simple image generation- (1) UNet-eVAE for iterative refinement of segmentation and improved generalisation using VAE embodied learning, (2) EndoUDA approach for

unsupervised domain adaptation in endoscopy. The UNet-eVAE approach uses a VAE as a secondary network to enhance ROI-specific feature representations that are jointly combined and optimised with a UNet network for the image segmentation task. As opposed to shared encoder-based approaches, UNet-eVAE uses the previous epoch's prediction to constrain the VAE to only provide robust fine-grained ROI-specific features that act as attention to the segmentation network. The experimental results on varied endoscopic imaging datasets demonstrate the ability of the proposed UNet-eVAE to generalise on both same-distribution test samples and unseen out-of-sample test datasets. EndoUDA is a novel target-independent unsupervised domain adaptation VAE-based method that comprises a shared encoder for both reconstruction and segmentation and a joint loss function for improved unseen target domain generalisation. The proposed approach has been validated on two GI endoscopy datasets that often require additional modalities for screening and intervention. This proposed modality-agnostic method can be used in clinical GI endoscopy without imposing any constraints on what imaging modality is being used. The qualitative and quantitative results demonstrate the effectiveness and promise of the proposed method compared to both the naive baseline supervised model and state-of-the-art unsupervised domain adaptation segmentation methods. Generative models, therefore, have numerous applications ranging from image synthesis to domain adaption and improved generalisation of tasks such as image segmentation.

6

Turbidity mitigation for improved visualisation

Contents

6.1	Introduction	116
6.2	Implementation of state-of-the-art approaches for image denoising	119
6.2.1	Averaging	120
6.2.2	Traditional dehazing approaches	120
6.2.3	Noise2Noise	122
6.2.4	Deep image prior	122
6.2.5	CycleGAN	124
6.3	Method	124
6.3.1	Network modules	125
6.3.2	Proposed framework	126
6.4	Experiments and Results	127
6.4.1	Datasets	128
6.4.2	Network implementation	128
6.4.3	Evaluation metrics	129
6.4.4	Quantitative results	129
6.4.5	Qualitative results	131
6.4.6	Quality score prediction using a neural network	132
6.5	Discussion	134
6.6	Conclusion	137

The lithotripsy procedure is normally performed with fluid irrigation to keep the surgical cavity dilated and also reduce the extra heat build-up from lithotripsy. This fluid often turns turbid due to the stone debris originating from stone fragmentation and the

intra-operative bleeding that may occur due to the trauma caused to the walls of the cavity. To improve vision quality, the flow of the irrigation fluid is required to be increased. However, this increases the intrarenal pressure which can result in septic complications. Therefore, this chapter presents an attempt towards developing a framework that can effectively disentangle turbidity and provide real-time image restoration of ureteroscopy and laser lithotripsy datasets. The proposed framework uses segmentation as an inductive bias to explicitly disentangle foreground and background space that helps to preserve the foreground information and disentangle turbidity from the background.

Contribution/Originality: I independently designed and implemented the proposed framework in Pytorch with the following novel contributions: (1) the first attempt ever made towards disentangling turbidity and debris for improved visualisation of ureteroscopic laser lithotripsy images (2) proposed a novel end-to-end β -VAE-based framework that uses segmentation as an inductive bias to effectively disentangle turbidity for enhanced visualisation (3) performed an extensive comparison between the proposed approach against state-of-the-art methods for image denoising and enhancement. (4) performed experimental validations on diverse and challenging *in vitro* and *in vivo* (animal) ureteroscopic lithotripsy datasets to demonstrate the effectiveness of the proposed approach. I collected the animal datasets from Boston Scientific. I developed the experimental setup for acquiring *in vitro* (synthetic data) wherein Dr Louise Goldsmith performed laser lithotripsy. I independently annotated the *in vitro* dataset used in this study.

6.1 Introduction

With roughly 300,000 cases performed each year, laser lithotripsy is the most common laser procedure performed in urology [240]. During the procedure, a ureteroscope is inserted through the urinary tract for illumination and imaging, while a laser fibre is inserted through the working channel of the ureteroscope to the location of the stone within the urinary system. Once the location of the stone has been accessed, the laser is activated to fragment the stone into smaller pieces until either they can be extracted using a basket or they are sufficiently small (<2 mm) to eventually pass unrestricted through the urinary tract [241]. Section 2.1.3 presents a more detailed description of the lithotripsy procedure.

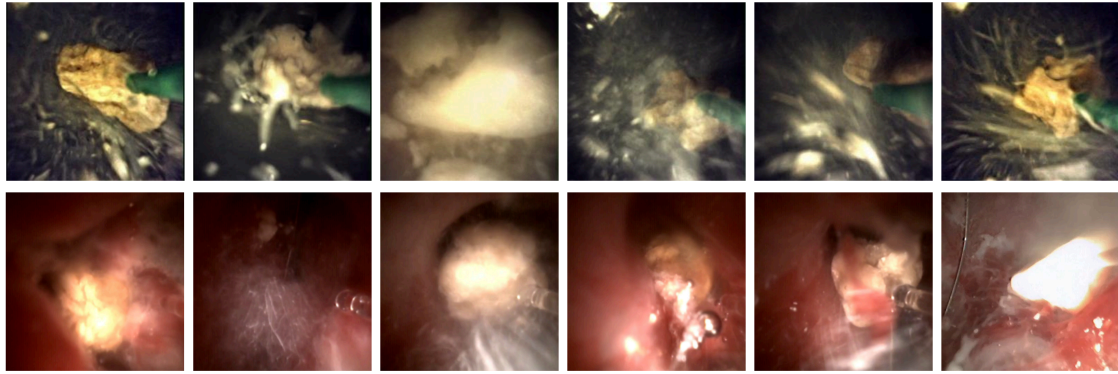


Figure 6.1: Example images for *in vitro* and *in vivo*(animal) ureteroscopic laser lithotripsy images showing turbidity and other artefacts such as blood, saturation, and specularities

Laser lithotripsy can increase intrarenal temperatures, especially at high laser power settings and can lead to thermal ablation of kidney tissue which may result in permanent kidney damage. Therefore, continuous irrigation flow is very important to reduce this heat buildup within the cavity. In addition to cooling, the irrigation flow is necessary for optimal visualisation during the lithotripsy procedure wherein the visualisation is impaired by stone dust and debris [27]. Hence, vision is the key to performing stone fragmentation in an efficient and effective manner [26]. However, increasing the irrigation flow to improve visualisation also increases the intrarenal pressure which can potentially lead to venous, tubular, and lymphatic backflow or septic complications. During ureteroscopic laser lithotripsy, intrarenal pressures have been reported to be as high as 50-350mmHg while the baseline intrarenal pressure is only 10mmHg [27]. Therefore, there is a need for robust computer vision algorithms that can mitigate turbidity and provide optimal vision whilst keeping the irrigation flow low.

A comprehensive review of various hard computing (traditional) and soft computing (neural network-based learning) approaches for denoising and image enhancement has already been presented in Section 2.2.2. Most of the traditional methods for image enhancement work on the basic assumption that an image can be decomposed into several components related to the imaging process and the high-quality image can be obtained by estimating the physical parameters from a single image or by using prior information. However, such methods often lead to colour deviation, saturation or spurious results as it is very difficult to accurately estimate the physical parameters or find a

good reference image [182]. Moreover, most traditional methods have been designed for denoising of natural images only. Supervised learning-based approaches require high-quality noiseless images during training, thereby rendering them inapplicable for most medical imaging applications. Unsupervised learning-based approaches that rely on assumptions about noise models may not be applicable for most medical imaging datasets as they have complicated noise characteristics [191]. GAN-based approaches have shown impressive image restoration results on various datasets but they often amplify noise in dark background areas and may produce halo artefacts [196]. Also, GANs are not suited for representation learning [197, 198]. Variational autoencoder (VAE) [162] has led to a recent wave of generative models as it enables both representation and statistical learning in a single framework [197]. Implicit disentanglement [199–202] using VAEs may provide remarkable results but the obtained representations may not be interpretable [203]. Therefore, explicit disentanglement [198, 204, 205, 209] that involves separation of latent variables into disjoint groups using priors can be useful to obtain interpretable representations and effective disentanglement.

Ureteroscopic laser lithotripsy videos can have turbidity present across the entire image (example images are provided in Fig. 6.1) due to the turbid fluid and stone debris resulting from lithotripsy. Images generated during ureteroscopy and laser lithotripsy procedure can be modelled as:

$$X(x) = \varphi M(C(x)) + (1 - \varphi)T(x) \quad (6.1)$$

where C represents a clean background image with objects such as stones and surgical instruments; T represents particle storm, blood, and turbid fluid; M represents motion artefacts/blur, and φ is a parameter that controls the weight of these terms. The value of φ will decrease as the image becomes turbid. The aim is to disentangle T from images that can act as a virtual flush to provide improved visualisation to the surgical team whilst keeping the irrigation flow low.

To the best of our knowledge, turbidity mitigation and image enhancement in ureteroscopic laser lithotripsy datasets have not been addressed before and is a significantly challenging problem as opposed to the other endoscopy datasets with localised artefacts

(such as specular reflections, bubbles, saturation, etc). To gain a deeper understanding of the nature and complexity of the problem at hand, this study started with the implementation of some state-of-the-art denoising and image enhancement approaches on the ureteroscopy dataset. The motivation behind experimenting with each of these and the corresponding results have been presented in Section 6.2. This is followed by Section 6.3 which presents the proposed end-to-end framework for disentangling turbidity and stone debris from images. Inspired by the work of Nakagawa *et al.* [198], the proposed framework employs segmentation masks as an inductive bias for latent disentanglement of turbidity from challenging ureteroscopic laser lithotripsy datasets. The main contributions made in this work can be summarised as follows:

- A novel end-to-end β -VAE-based framework that uses segmentation as an inductive bias to effectively disentangle turbidity and provide real-time image restoration.
- An extensive comparison between proposed approach against state-of-the-art methods for image denoising and enhancement
- Experimental validations on diverse and challenging *in vitro* and *in vivo* ureteroscopic lithotripsy datasets to demonstrate the effectiveness of our proposed approach.

6.2 Implementation of state-of-the-art approaches for image denoising

Turbidity mitigation for improved visualisation in ureteroscopy and laser lithotripsy is a novel computer vision task that has not been explored. Several state-of-the-art denoising and image enhancement approaches have been proposed in the literature for various datasets. This section presents the results of some of these approaches on our ureteroscopy dataset to help us gain a deeper understanding of the problem at hand.



Figure 6.2: Results for temporal averaging of pixel intensities (across five adjacent frames) shown for three *in vitro* and three animal dataset samples

6.2.1 Averaging

Temporal averaging of frames is useful for image-denoising when the noise can be assumed to be completely random and the images have no movement. The nature of stone debris and turbidity in ureteroscopy images can be assumed to be completely random. Figure 6.2 presents qualitative results of temporal averaging on *in vitro* and animal dataset. It can be seen from Figure 6.2 that the obtained output images are blurred and have lost the fine pixel-level details. This is due to the continuous movement of stone, background and laser fibre across frames. In addition, debris mitigation is not possible using this technique in cases where the debris is present in a pixel location across the entire temporal window (see row 3 and row 6 in Figure 6.2).

6.2.2 Traditional dehazing approaches

Turbidity in ureteroscopy images visually resembles the haze in natural scenes to some extent. Just like the turbidity in lithotripsy images, haze in natural images is also defined as an intrusion element that changes the colour, reduces the contrast, and makes the object



Figure 6.3: Dehazing results using approach (a) presented in [242] and approach (b) presented in [243] implemented on two natural scene images followed by three *in vitro* and three animal dataset ureteroscopy images

features difficult to identify. In the context of haze, $T(x)$ in Equation 6.1 can be thought of as representing the haze that degrades the perceptual image quality. Several methods have been proposed in the literature for the efficient removal of haze from natural scene images [242, 243]. Meng *et al.* proposed a method that uses a combination of constraints on the transmission function and L_1 -norm-based contextual regularisation for image dehazing [242]. Ngo *et al.* proposed a haze removal algorithm with a corresponding hardware implementation [243]. This approach exploited computationally efficient image processing techniques such as detail enhancement, multiple-exposure image fusion, and adaptive tone remapping [243]. Figure 6.3 illustrates the results of the application of this methods [242, 243] on natural scenes and our ureteroscopy dataset. It can be seen that these dehazing approaches perform well on natural images but not on ureteroscopy images. It can be seen that even though it may remove light debris from certain pixels, it cannot improve parts of the image that have heavy debris or other artefacts such as blood. In addition, it can be seen that the background information gets lost and the colour tone of the images gets changed as well which is undesirable. Therefore, it can be concluded that the properties of turbidity in ureteroscopy images are more complex than that of haze in natural images.

6.2.3 Noise2Noise

The Noise2Noise(N2N) algorithm proposed by Lehtinen *et al.* [188] in 2018 showed that it is possible to recover signals under complex corruptions without observing the clean signals and without an explicit characterisation of noise model or other image priors. N2N works by utilising the property of L_2 minimisation wherein the authors stated that the expectation of MSE remains unchanged even if ground truth image/signal is replaced with its zero mean independent and identically distributed Gaussian Noise corrupted realisation. This essentially would imply that the optimal network parameters of a denoising network will not change and provide the same solution as if it was trained on clean data. Even though N2N achieved outstanding performance for various kinds of noise realisation on different datasets, it is not very useful for mitigating turbidity in ureteroscopy datasets due to the lack of a set of turbid images for the same clean image. Therefore, to evaluate the performance of N2N on ureteroscopy datasets, pairs of consecutive images were used assuming that the clean signal underneath is the same. Figure 6.4 illustrates the results of N2N trained on both *in vitro* and *in vivo* datasets by using pairs of consecutive images. It can be seen from Figure 6.4 that the resulting images are blurred and over-smoothed and N2N failed to tackle streaks of debris or other artefacts present in the images. One major reason behind this is that the pairs of consecutive images do not represent the same clean signal underneath as the camera, stone, laser fibre, and background are not static during the lithotripsy procedure. Due to these movements, $M(C(x))$ in Equation 6.1 is not usually the same for a pair of images from the lithotripsy dataset. Also, stone debris/turbidity in images is like an undesirable image artefact that does not essentially possess the characteristics of noise realisations such as Gaussian, Poisson, or Bernoulli.

6.2.4 Deep image prior

As opposed to deep convolutional networks that use a large amount of training data to learn realistic image priors in order to provide excellent results, Deep Image Prior(DIP) [187] showed that the structure of a generator network is sufficient to capture relevant low-level image statistics prior to any learning. The authors of the DIP paper showed that a randomly-initialised neural network can be used as a handcrafted prior with excellent

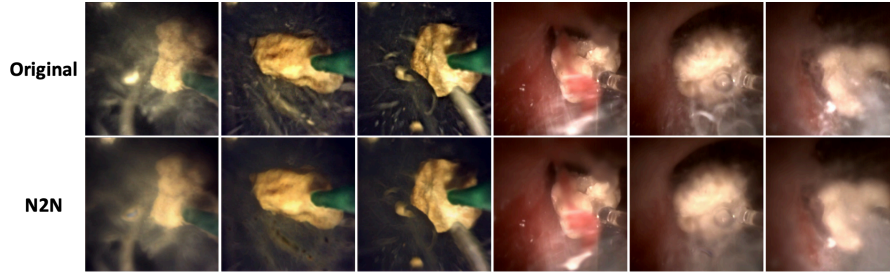


Figure 6.4: Qualitative results of Noise2Noise(N2N) implementation on three *in vitro* and three animal dataset ureteroscopy images

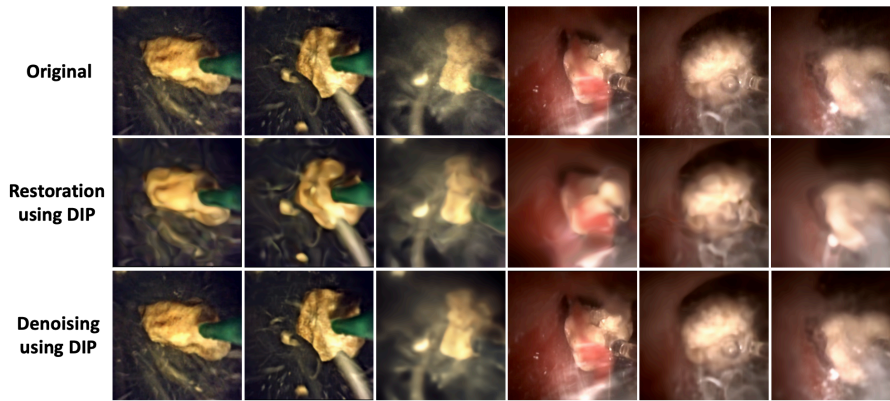


Figure 6.5: Image restoration and denoising using Deep Image Prior on three *in vitro* and three animal dataset ureteroscopy images showing there is no visible change in the image quality of resulting images

results in computer vision problems such as image denoising, super-resolution and inpainting. DIP algorithm bridged the gap between learning-based methods such as DNN and learning-free methods that use handcrafted image priors. Therefore, DIP to our dataset to evaluate its denoising and image restoration performance on our dataset. Image inpainting using DIP is not applicable to our dataset as the stone debris is not usually a localised artefact that can be annotated but is like an additive layer on top of the clean signal. Figure 6.5 shows the denoising results of the DIP algorithm on both *in vitro* and *in vivo* images. One can observe that image restoration using DIP resulted in smoother images while image denoising resulted in no visual change in the quality of the resulting images. This can be because the structure of the generator network cannot capture sufficient details from complex medical datasets that have various image artefacts and weak texture information as opposed to natural scenes.

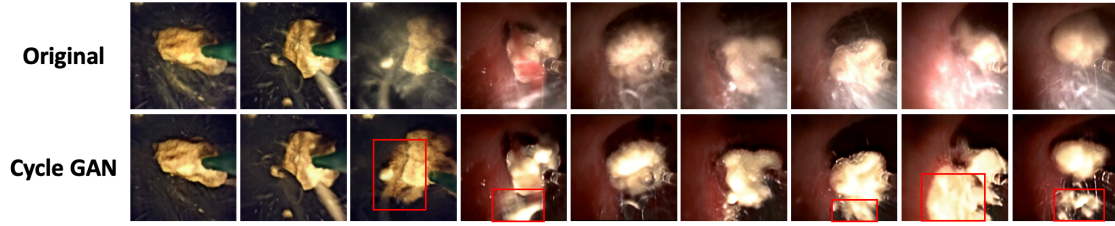


Figure 6.6: Implementation of CycleGAN on our dataset to convert Domain-A (debris) images to Domain-B (clean) images. Failure cases where stone debris or saturated pixels are getting reconstructed as a stone have been highlighted using red rectangles

6.2.5 CycleGAN

CycleGAN is a GAN-based model architecture widely used for unpaired image-to-image translation [194]. It involves the simultaneous training of two generator models and two discriminator models. One generator takes images from one domain and outputs images for the second domain while the other generator takes images from the second domain and outputs images for the first domain. Discriminator models are simultaneously trained to learn to determine if the generated images are plausible. CycleGAN has provided excellent image translation results on a range of applications such as style transfer and object transfiguration. CycleGAN can ideally be also used to convert debris/turbid images into clean images without the use of the corresponding clean image. Figure 6.6 illustrated CycleGAN results of conversion from Domain-A (debris) to Domain-B (clean) images in the case of both *in vitro* and *in vivo* datasets. It can be seen from Figure 6.6 that even though CycleGAN can successfully remove stone debris from some images, it also reconstructs some debris areas as a stone. This is essentially an extremely dangerous result in the field of ureteroscopy as this may cause the surgeon to laser the tissue instead of the stone.

6.3 Method

The proposed framework AE- β VAE-Seg (as shown in Figure 6.7) is designed as an end-to-end CNN framework that optimises a combination of losses from its three sub-networks that work together to enhance the visualisation of lithotripsy images: (1) UNet to induce segmentation bias for explicit separation of foreground and background latent space; (2)

AE (Autoencoder) to help preserve the foreground structure; and (3) β -VAE (Variational Autoencoder) to disentangle turbidity from the background. Details of each network module and its integration into the framework are provided in this section.

6.3.1 Network modules

The proposed framework consists of three sub-networks- UNet, AE and VAE. UNet consists of an encoder and a decoder. The encoder, also called the contracting path, consists of repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with a stride of 2 for downsampling. In the contracting path, the image size gradually reduces and the depth gradually increases as each downsampling step results in double the number of feature channels. The decoder path, also known as the expansive path, involves the application of transposed convolutions along with regular convolutions to gradually increase the size of the image and reduce the depth of the image. To improve spatial information, every step of the decoder path uses skip connections by concatenating the output of the transposed convolution layers with the feature maps from the encoder at the same level where each concatenation is followed by the application of two consecutive regular convolutions [40]. Analogous to UNet, Autoencoder also consists of an encoder and a decoder. The encoder tries to compress the input to a low-dimensional representation and the decoder attempts to reconstruct the input from this compressed representation. Autoencoders can efficiently learn significant features present in input data by trying to minimise the reconstruction error between the input and output image. Variational Autoencoder (VAE) is a specific type of autoencoder that belongs to the category of generative models, similar to GANs (Generative Adversarial Networks). Unlike the convolutional autoencoders that encode the input image as a single point, VAE models encode it as a distribution over the latent space by returning the mean and variance of the latent space distribution. Different samples from this distribution can be fed into the decoder of the VAE models to generate new samples.

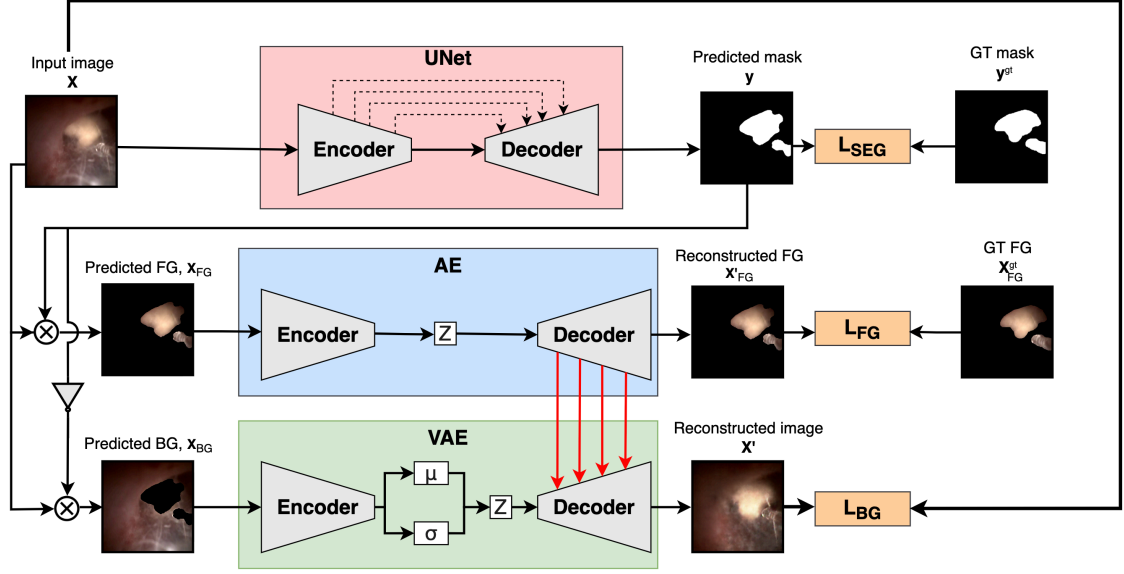


Figure 6.7: Proposed end-to-end framework for enhancing the quality of ureteroscopy images. The framework consists of three sub-networks- UNet, AE and a β -VAE. The predicted foreground and background image obtained with the help of UNet is fed into AE and β -VAE, respectively for explicit disentanglement of the latent space

6.3.2 Proposed framework

The proposed framework, AE- β VAE-Seg is illustrated in Figure 6.7. The input image $\mathbf{X} \in \mathbb{R}^{3 \times h \times w}$ is fed into the UNet model to obtain a segmentation map $\mathbf{y} \in \mathbb{R}^{h \times w}$ that corresponds to a binary classification of each pixel in $\mathbf{X}$. The predicted segmentation map, $\mathbf{y}$ and its inverted version is multiplied respectively with the original input image $\mathbf{X}$ to obtain a foreground (FG) and a background image (BG). This predicted FG image $\mathbf{X}_{FG}$ is fed into an Autoencoder network that is trained to reconstruct the foreground image $\mathbf{X}'_{FG}$. On the other hand, predicted BG image $\mathbf{X}_{BG}$ is fed into a β -VAE model that is tasked with reconstructing the entire input image. To achieve this, feature map representations from the decoder layers of the Autoencoder model are fused into the corresponding decoder layers of the VAE model (indicated using red arrows in Figure 6.7). Z dimension of 10 has been used for the VAE models across all experiments.

For segmentation, a combination of binary cross-entropy loss BCE_{loss1} and dice similarity coefficient loss DSC_{loss2} [222] is used. It is calculated between the predicted

mask $\mathbf{y}$ and its ground truth mask $\mathbf{y}^{gt}$, and can be given by:

$$\mathcal{L}_{SEG} = \underbrace{\frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i^{gt} (\log(\mathbf{y}_i)) + (1 - \mathbf{y}_i^{gt}) (\log(1 - \mathbf{y}_i)))}_{\text{BCE}_{\text{loss1}}} + 1 - \underbrace{\frac{2 * \sum_{i=1}^N (\mathbf{y}_i^{gt} * \mathbf{y}_i)}{\sum_{i=1}^N \mathbf{y}_i^{gt} + \sum_{i=1}^N \mathbf{y}_i}}_{\text{DSC}_{\text{loss2}}}, \quad (6.2)$$

where N is the total number of pixels, $\mathbf{y}^{gt}$ is the label (1 or 0) and $\mathbf{y}$ is the predicted probability. For the reconstruction of the foreground from the autoencoder, binary cross-entropy loss $\text{BCE}_{\text{loss3}}$ is calculated between the reconstructed foreground image from Autoencoder $\mathbf{X}'_{FG}$ and its corresponding ground truth foreground image $\mathbf{X}^{gt}_{FG}$. This can be given by:

$$\mathcal{L}_{FG} = \underbrace{\frac{1}{N} \sum_{i=1}^N (\mathbf{X}^{gt}_{FGi} (\log(\mathbf{X}'_{FGi})) + (1 - \mathbf{X}^{gt}_{FGi}) (\log(1 - \mathbf{X}'_{FGi})))}_{\text{BCE}_{\text{loss3}}} \quad (6.3)$$

where N is the total number of pixels, $\mathbf{X}^{gt}_{FG}$ is the label (range[0,1]) and $\mathbf{X}'_{FG}$ is the predicted probability. For the β -VAE sub-network, a combination of BCE reconstruction loss $\text{BCE}_{\text{loss4}}$ and KL-divergence $\text{KLD}_{\text{loss5}}$ is used. Here, $\text{BCE}_{\text{loss4}}$ is calculated between the reconstructed image from VAE network $\mathbf{X}'$ and the original input image $\mathbf{X}$.

$$\mathcal{L}_{BG} = \underbrace{\frac{1}{N} \sum_{i=1}^N (\mathbf{X}_i (\log(\mathbf{X}'_i)) + (1 - \mathbf{X}_i) (\log(1 - \mathbf{X}'_i)))}_{\text{BCE}_{\text{loss4}}} - 0.5\beta \underbrace{\sum_{k=1}^z (1 + \log \sigma_k^2 - \sigma_k^2 - \mu_k^2)}_{\text{KLD}_{\text{loss5}}}, \quad (6.4)$$

where $\beta = 5$ is used to weight $\text{KLD}_{\text{loss5}}$. The final joint loss ($\mathcal{L}_{SEG} + \mathcal{L}_{FG} + \mathcal{L}_{BG}$) is optimised using Adam optimiser with learning rate of $1e^{-3}$.

6.4 Experiments and Results

This section presents a description of the datasets, network implementation and evaluation metrics used in this study. This is followed by a comparison of the quantitative and qualitative results of the proposed network against state-of-the-art approaches. Whilst highlighting that the well-established image quality metrics are not very suitable for the specific application of turbidity-based image quality prediction, a score prediction based on a neural network has also been presented to further evaluate the effectiveness of the

proposed approach. In Table 6.1, Figure 6.8, Figure 6.9, and Table 6.2, notation Seg in the network name indicates that a segmentation network is used in the framework and network name β VAE-Seg represents our experimental framework wherein the autoencoder model for the foreground has been replaced by a β -VAE.

6.4.1 Datasets

The data used in this study has been acquired using single-use LithoVue™ scope and Lithovue Elite scope (Boston Scientific). The *in vitro* dataset was acquired under controlled settings wherein laser lithotripsy of human kidney stones was performed inside a container with irrigation fluid flowing through it. The *in vitro* dataset consists of 70 training samples, nine validation samples, and nine test samples. The *in vivo* dataset consists of samples from animal studies performed at Boston Scientific sites. It has 364 training images, 46 validation images, and 46 test images. The images were manually annotated into two classes: foreground (stones and instruments) and background to obtain binary segmentation masks. The *in vitro* annotations were performed by me and independently verified by two experts, a senior research associate with over 5 years of experience in endoscopy and a senior urologist. The annotations for the *in vivo* dataset (animal study samples) were provided by Boston Scientific.

6.4.2 Network implementation

The experiments in this work were implemented using the PyTorch framework and trained on NVIDIA Quadro RTX 6000. All image samples were resized to $256 \times 256 \times 3$ pixels and trained using Adam optimiser with a learning rate of $1e^{-3}$ and batch size of 8. In the Adam optimiser, initial decay rates were set to default 0.9 and 0.999 for estimation of the first and second moments of gradient computations, respectively. Validation was performed after each epoch. Early Stopping was used to monitor validation loss and stop the training when no improvement is observed. Training of the segmentation part of the framework is stopped when no further improvement is observed in the prediction of segmentation masks.

6.4.3 Evaluation metrics

Evaluation metrics have been reported separately across the *in vitro* and *in vivo* datasets. Sophisticated metrics such as visual information fidelity (VIF) [244] and relative edge coherence (RECO) [245] have been reported. Most image quality assessment algorithms typically define an image's quality as its fidelity to or resemblance to a "reference" image in a certain perceptual space. Such methods use signal fidelity measures or model features of the human visual system (HVS) in an effort to achieve consistency in quality prediction. VIF is an image quality metric that quantifies the information that could ideally be extracted from the reference image and then quantifies the amount of this image information lost to the distortion using Natural Scene Statistics (NSS), Human visual system (HVS), and an image distortion model. For all practical distortions, VIF lies in the interval [0,1] [244]. RECO is a metric that is based on local analysis with circular harmonic functions that are extracted from Laguerre-Gauss (GL) orthonormal family [245]. A pair of GL coefficients is used to calculate the polar edge coherence. In this approach, any edge degradation is revealed by misalignments or attenuation of higher-order harmonics and these edge disruptions cause a loss in coherence. [245] Due to the lack of ground truth clean images, image quality metrics such as peak signal-to-noise ratio (PSNR) and structural similarity SSIM have not been presented. However, even these well-established metrics [244, 245] have not been specifically developed to evaluate image quality for specific applications such as turbidity mitigation in lithotripsy images. Therefore, to further evaluate the effectiveness of the proposed approach, I have complemented this quantitative analysis by proposing to use a neural network for automated prediction of debris scores (see Section 6.4.6).

6.4.4 Quantitative results

Table 6.1 presents a quantitative analysis of the image quality obtained from state-of-the-art β -VAE and our frameworks (β VAE-Seg and AE- β VAE-Seg) on *in vitro* and *in vivo* ureteroscopic lithotripsy test datasets. It can be seen that our proposed framework AE- β VAE-Seg significantly improved the VIF metric over β -VAE from 0.3439 to 0.7362 in case of *in vitro* and from 0.5234 to 0.6516 in case of *in vivo*. For the RECO metric, it

Dataset	Network	VIF	RECO
In vitro	β -VAE	0.3439 ± 0.19	0.7558 ± 0.48
	β VAE-Seg (Ours)	0.3354 ± 0.19	0.8432 ± 0.51
	AE- β VAE-Seg (Proposed)	0.7362 ± 0.14	0.7585 ± 0.25
In vivo	β -VAE	0.5234 ± 0.14	0.4468 ± 0.28
	β VAE-Seg (Ours)	0.5698 ± 0.12	0.7217 ± 1.54
	Auto β VAE-Seg (Proposed)	0.6516 ± 0.14	0.6763 ± 0.79

Table 6.1: VIF and RECO metric values (mean $\pm$ std) for quantitative assessment of image quality obtained from state-of-the-art β -VAE and our frameworks on *in vitro* and *in vivo* test datasets.

can be seen that our other framework β VAE-Seg improved the image quality by 11.56% and 61.5% over β -VAE in case of *in vitro* and *in vivo* datasets, respectively. It can also be seen that our proposed framework AE- β VAE-Seg provided a 0.36% and 51.36% RECO improvement over β -VAE in case of *in vitro* and *in vivo* datasets, respectively.

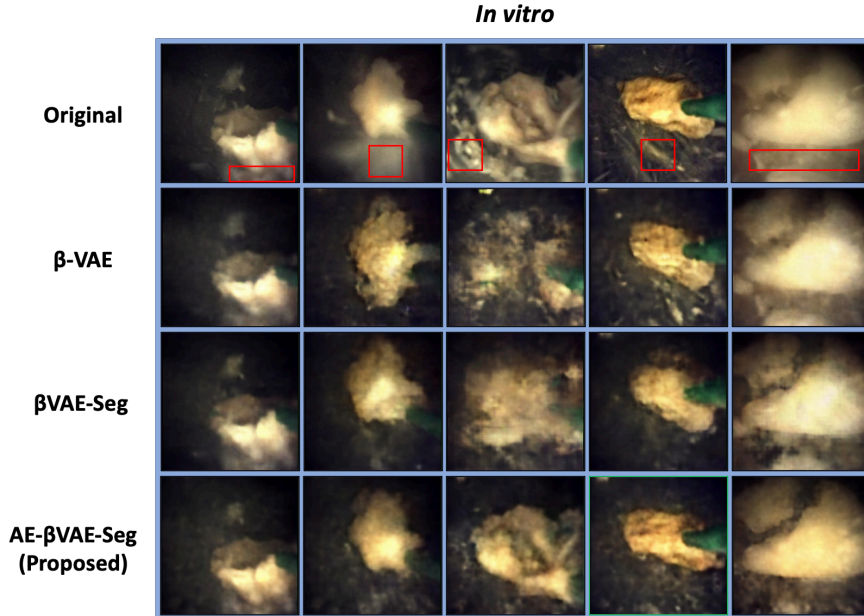


Figure 6.8: Qualitative evaluation of five *in vitro* images (first row) against the resulting images obtained from state-of-the-art β -VAE and our frameworks (shown in subsequent rows). The proposed network can be seen to preserve the foreground structure whilst reducing turbidity (see debris in the regions highlighted using red rectangles (first row) getting reduced using the proposed network (last row)).

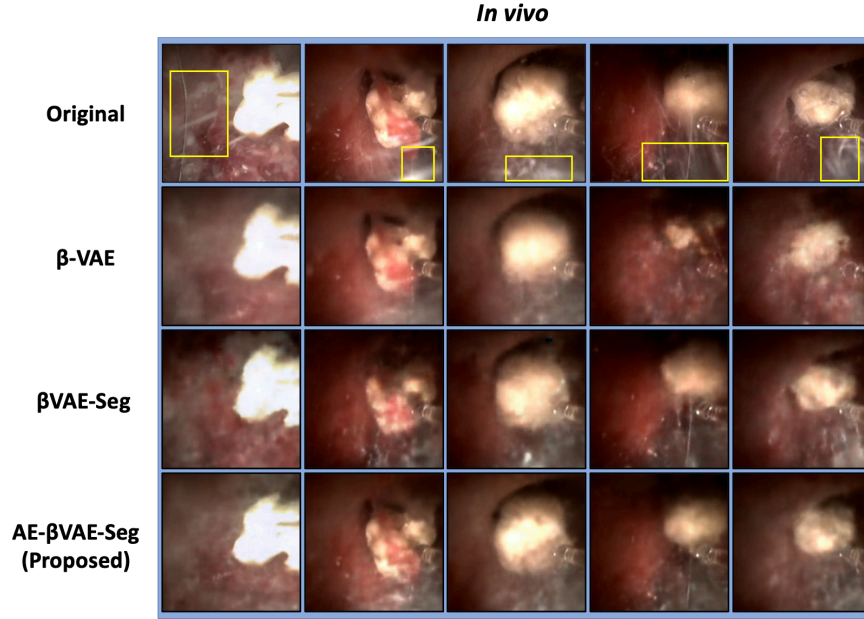


Figure 6.9: Qualitative evaluation of five *in vivo* images (first row) against the resulting images obtained from state-of-the-art β -VAE and our frameworks (shown in subsequent rows). The proposed network can be seen to preserve the foreground structure whilst reducing turbidity (see debris in the regions highlighted using yellow rectangles (first row) getting reduced using the proposed network (last row)).

6.4.5 Qualitative results

Figure 6.8 and Figure 6.9 present the qualitative analysis of the image quality obtained from state-of-the-art β -VAE and our frameworks (β VAE-Seg and AE- β VAE-Seg) on *in vitro* and *in vivo* datasets, respectively. In the case of both *in vitro* and *in vivo*, it can be seen that the state-of-the-art β -VAE completely fails to preserve the structure of the foreground for most images. However, β VAE-Seg that uses segmentation as an inductive bias to explicitly separate the distribution of foreground from the background using two β VAEs performs better than β -VAE in preserving the structure of the stone. This β VAE-Seg framework is outperformed by our proposed framework AE- β VAE-Seg that can be seen to reduce turbidity and debris whilst preserving the foreground information in all the images. This trend of foreground preservation and debris disentanglement also be seen in Figure 6.10 which illustrates the predicted images obtained by interpolating the latent space of 10 dimensions from -3 to 3 for various networks.

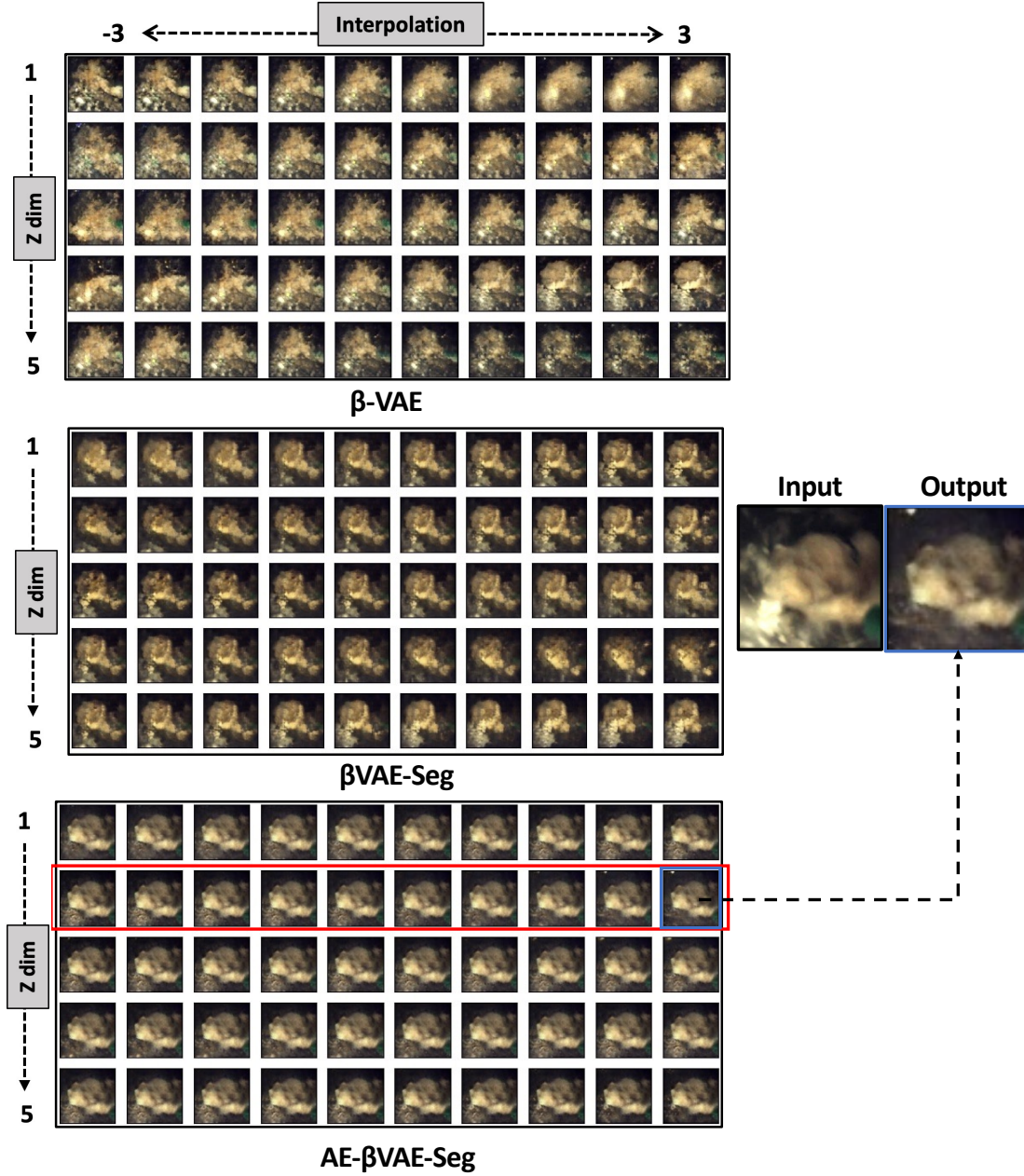


Figure 6.10: Reconstructed images obtained by interpolating the latent space (only shown first five of the ten dimensions) from -3 to 3 for various networks. AE- β VAE-Seg can be seen to preserve foreground information across all latent dimensions with the second latent dimension (highlighted in red) disentangling debris. The right-most image across this second dimension (highlighted in blue) is taken to be the enhanced output image.

6.4.6 Quality score prediction using a neural network

For additional validation of the effectiveness of the proposed method, I annotated a set of *in vitro* images by manually assigning them debris scores based on visual inspection. The debris scores are categorised into three categories: 0-*good*, 1-*medium* and 2-*bad* wherein

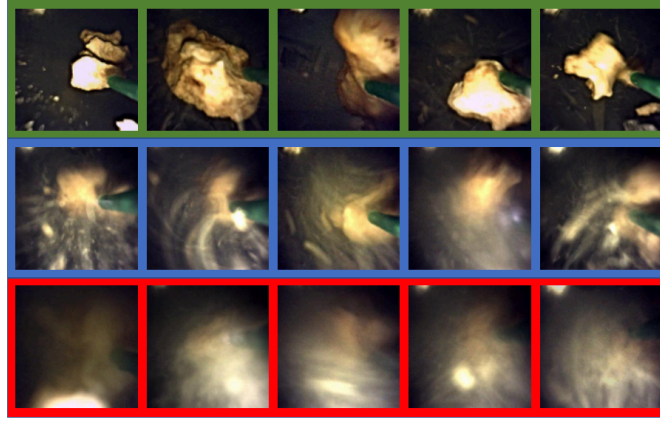


Figure 6.11: Figure illustrating *in vitro* images classified into *good*(green), *medium*(blue), and *bad*(red) categories

good imply that the clinician can continue to laser without any challenges; *medium* implies that the clinician can continue to laser but there is limited visibility(commonly due to stone or blood debris), and *bad* implies that the visibility is extremely poor and that the clinician either needs to wait for the debris to clear out or increase the pressure of the fluid. For a better understanding of the debris scores, sample images classified into *good*(green), *medium*(blue), and *bad*(red) scores have been illustrated in Figure 6.11.

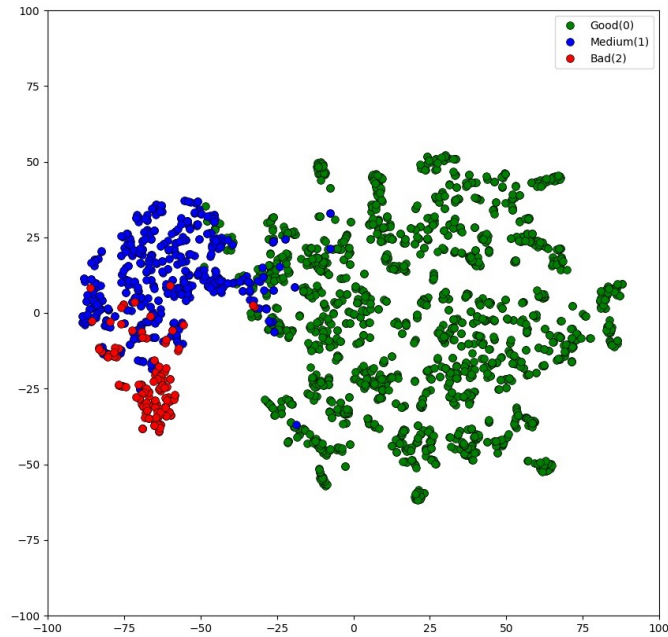


Figure 6.12: t-SNE plot illustrating the distribution of manually annotated debris scores that were used to train the autoencoder-based classification network on *in vitro* dataset

Network	# improved	# degraded
β -VAE	21.05%	13.15%
β VAE-Seg	21.05%	15.78%
AE- β VAE-Seg (Proposed)	31.6%	7.89%

Table 6.2: Percentage of images that showed improvement or degradation out of the 88 *in vitro* images as observed with state-of-the-art β -VAE and our frameworks. Improvement or degradation is calculated based on whether the debris score predicted by the network was decreased or increased, respectively after application of the networks

These labels were then used to train a simple autoencoder network wherein features extracted from the encoder were used for the automated classification of images into their debris scores. To visualise the correctness of the assigned labels, a t-SNE plot has been illustrated in Figure 6.12 showing the distribution of *good*(green), *medium*(blue), and *bad*(red) scores. The training checkpoint obtained from training this prediction network was used to evaluate the image quality of the images in the *in vitro* dataset. Table 6.2 presents the percentage of images that showed improvement or degradation after processing images with state-of-the-art β -VAE and our frameworks across the *in vitro* dataset used in this study. Improvement or degradation is calculated based on whether the debris score predicted by the network was decreased or increased, respectively after the application of the frameworks. It can be seen that our proposed method AE- β VAE-Seg improved the quality of 31.6% *in vitro* images and only degraded the quality of 7.89% images as opposed to β -VAE that only improved 21.05% images and degraded 13.15% images. Manually assigning accurate quality scores to *in vivo* images is a challenge due to the complex background information. Therefore, the validation of the proposed method was only evaluated on *in vitro* dataset.

6.5 Discussion

Mitigation of turbidity and stone debris from ureteroscopic laser lithotripsy images using computer vision algorithms can greatly help the surgical team deliver an effective and efficient treatment by improving the visualisation during the procedure. However, this application of computer vision has not been explored so far. This work presents an attempt to understand the computer vision-based challenges associated with the mitigation of

turbidity and debris from ureteroscopic laser lithotripsy images and proposes an end-to-end framework that can disentangle turbidity and stone debris for improved visualisation.

Section 6.2 presents the results of various state-of-the-art image enhancement approaches such as denoising using temporal averaging, traditional dehazing approaches, Noise2Noise, Deep Image Prior (DIP), and CycleGAN on our dataset. The challenges associated with turbidity mitigation on our dataset include a lack of ground truth clean image, lack of explicit characterisation of the debris (noise) model, and dynamically changing background information due to continuous movement of stone, background and laser fibre across frames. Temporal averaging can potentially work only when some frames across the temporal window have no debris/turbidity and strongly assumes that the temporal change in background information across frames is insignificant, rendering it inapplicable to lithotripsy application (see Figure 6.2). Results of the implementation of traditional dehazing approaches (see Figure 6.3) illustrate that the background information is lost with partial removal of some light debris and the image colour tone is changed. This indicates that the properties of turbidity in ureteroscopy images are more complex than that of haze in natural images. Results of the implementation of Noise2Noise (see Figure 6.4) using consecutive pairs of images shows that it failed to tackle streaks of debris/ heavy turbidity areas and resulted in a blurring of the images. One major reason behind this is that the pairs of consecutive images do not represent the same clean signal underneath as the camera, stone, laser fibre, and background are not static during the lithotripsy procedure. Both image restoration and denoising using Deep Image Prior (DIP) resulted in no visual change in image quality (see Figure 6.5) as its generator network cannot capture the details of weakly textured complex medical datasets as opposed to natural scenes. CycleGAN could translate our images from debris to no debris domain but it provided unacceptable results by reconstructing the stone debris as a stone fragment in a number of cases (see Figure 6.6). This is essentially a dangerous result in the field of lithotripsy as this may cause the surgeon to laser the tissue instead of the stone. Section 6.2 helped to understand the level of challenges and complexities associated with developing a computer vision algorithm that can improve the visualisation of ureteroscopic lithotripsy images.

Considering that the VAEs have been quite successful in representation learning, this work proposes a novel end-to-end β -VAE-based framework that uses segmentation as an inductive bias to disentangle turbidity/debris from the background whilst preserving the foreground information. As can be observed in Table 6.1, Figure 6.8, and Figure 6.9 that our proposed method AE- β VAE-Seg outperformed β -VAE and β VAE-Seg. Here, β VAE-Seg refers to a network where the autoencoder module for foreground information has been replaced by a β VAE, meaning β VAE-Seg has one β -VAE for foreground and one β -VAE for background. It can be observed from Figures 6.8 and 6.9 that the β -VAE fails to preserve the structure of the foreground (stones and laser) for most images. This can also be seen in Figure 6.10 which illustrates the predicted images obtained by interpolating the latent space of 10 dimensions from -3 to 3 for various networks. This essentially implies that the network needs to be constrained to be able to preserve the foreground (stone and laser) while only manipulating the background information to disentangle turbidity. This led to the introduction of a segmentation network into the framework and the usage of two separate β -VAEs to explicitly differentiate the foreground and background distributions. Even though this explicit separation of the foreground and background distributions improved the results as compared to standard β -VAE, it does not guarantee that the foreground information will be preserved along a latent dimension that disentangles debris. This trend can be observed for β VAE-Seg in Figure 6.10, where the structure of the stone is also getting altered along with the debris in most dimensions. Preserving the structure of the stone is the foremost priority in developing an algorithm for turbidity disentanglement. Therefore, the β -VAE module used for modelling the foreground was replaced with an autoencoder to constrain the network to only manipulate the turbidity while preserving the rest of the image details. This is our proposed framework and has been referred to as AE- β VAE-seg. It can be seen in Figure 6.10 that our proposed framework can disentangle debris (see latent dimension 2 highlighted in red) and is able to preserve the foreground information across all latent dimensions.

Table 6.1 demonstrates that the proposed method outperforms the other networks in most cases but not all of them. However, this conclusion is not completely reflected in the qualitative results that clearly show that the proposed method provided a more

enhanced image as compared to β -VAE and β VAE-Seg. This essentially indicates a weakness in the used metrics as these metrics do not include foreground preservation as a factor and have not been specifically developed for such computer vision tasks. Therefore, an autoencoder-based debris score prediction network has been used to evaluate the observed image quality after implementing the presented networks. This has only been performed for the *in vitro* dataset as it was easier to manually assign debris scores to these simple images with no background as opposed to the *in vivo* dataset that has complex background information and other artefacts that make it harder to assign accurate debris scores. Table 6.2 clearly shows that the proposed method improved the quality of 50% more images and degraded the quality of 40% fewer images as compared to the β VAE.

6.6 Conclusion

This chapter presents the first attempt made towards improving the visualisation of ureteroscopic laser lithotripsy images. The proposed framework effectively makes use of segmentation information to explicitly disentangle the foreground and background latent space, thus improving the preservation of the foreground and effective disentanglement of turbidity. The presented results demonstrate that the proposed approach outperforms existing state-of-the-art denoising and image enhancement approaches. Although the proposed approach seems to work well on most images, it has certain limitations as it is unable to completely disentangle the turbidity and debris from images. In addition, using a segmentation network in the framework limits the use of the proposed approach to the availability of ground truth segmentation masks. Future research direction is to develop an unsupervised framework that does not require segmentation masks and can effectively leverage temporal information in videos to provide complete disentanglement of turbidity from images. Such a framework can help the surgical team to perform lithotripsy without having the need to frequently increase the irrigation flow, thereby avoiding septic complications. Additionally, this can also help to shorten the procedure time whilst improving diagnosis and therapy.

7

Conclusion and Future work

Contents

7.1	Summary of contributions	138
7.2	Directions for future work	140
7.3	Conclusion	141

7.1 Summary of contributions

This thesis works towards moving ureteroscopy to the new era of autonomous surgery by making AI an assistive tool for surgeons. The following were achieved throughout the course of this thesis to fulfil the requirements for image segmentation and image enhancement on ureteroscopic laser lithotripsy images:

- **Chapter 1** of this thesis presents an introduction to the ureteroscopic laser lithotripsy procedure and motivates the need for the development of computer vision algorithms to solve the challenges faced by the surgical team during the procedure. This chapter also presents a summary of the ureteroscopy and laser lithotripsy datasets used in this thesis followed by the thesis structure and related contributions.
- **Chapter 2** of this thesis presents the clinical background of ureteroscopic laser lithotripsy followed by a comprehensive review of the existing image segmentation

and image quality enhancement methods recognising the need for the development of computer vision approaches for the ureteroscopic laser lithotripsy application.

- **Chapter 3** of this thesis presents a classical computer vision-based approach for the segmentation of stone fragments in ureteroscopy datasets. The proposed approach is a novel multi-frame total-variational optical-flow-based approach for the segmentation of stone fragments in ureteroscopy images. The proposed method is an extension of the classical two-frame optical flow model and uses robust correlation transform-based data descriptors to provide dense optical flow for the pronounced delineation of stone fragments and can effectively tackle problems related to illumination variability, blur, occlusions and deformations.
- **Chapter 4** of this thesis presents the first-ever deep learning-based approach for multi-class segmentation of ureteroscopic laser lithotripsy datasets. The proposed approach effectively makes use of residual connections and motion information between adjacent frames to produce robust and reliable segmentation of stone fragments and laser fibre in real-time. Extensive comparison with state-of-the-art approaches and extended out-of-sample study has also been performed to exhibit the effectiveness of the proposed approach.
- **Chapter 5** of this thesis presents alternative use cases of VAE-based generative models for tasks beyond simple image generation by presenting a novel UNet-eVAE framework where VAE is used as a secondary network that feeds region-of-interest (ROI) specific fine-grained features into the segmentation network for improved segmentation and generalisation on various endoscopy datasets. This is followed by another VAE-based framework, EndoUDA, for target-independent unsupervised domain adaptation. This approach has the potential to provide improved segmentation on unseen Narrow-Band Imaging (NBI) modality images when trained with White Light Imaging (WLI) modality images for both Barrett's oesophagus (BE) and polyp datasets.

- **Chapter 6** of this thesis marks the first attempt towards understanding the challenging problem of turbidity and debris disentanglement by implementing a number of state-of-the-art image enhancement approaches on ureteroscopic laser lithotripsy images. This is followed by the presentation of a novel end-to-end β -VAE-based framework that uses segmentation as an inductive bias to effectively disentangle turbidity for improved visualisation.

7.2 Directions for future work

This thesis lays the foundation for future research in the field of computer-guided ureteroscopy and laser lithotripsy. The directions for future work have been discussed below:

- **Unsupervised frameworks for segmentation:** The deep learning approaches developed in this thesis for segmentation and image enhancement of ureteroscopic lithotripsy data are supervised frameworks. However, obtaining accurate ground truth annotations is not only an expensive process but also a challenging task due to the occlusions from stone debris and artefacts from the turbid fluid. Therefore, there is a need for developing unsupervised learning-based frameworks that can leverage the intra-frame and temporal inter-frame information to effectively enhance the image quality and also provide robust segmentation of stone fragments and surgical instruments.
- **Field-of-view extension:** The narrow field of view limits the amount of contextual information available to the clinician. This causes difficulties in the determination of the surrounding circumstances of the endoscope. Fragmenting stones using laser results in smaller chunks of stones that wobble around in the surgical cavity. Here, an extended field of view could help the clinician to better track the smaller chunks by keeping them inside the field of view, thereby making the ureteroscopy procedure efficient.

- **Exploitation of spatio-temporal information for turbidity mitigation:** The proposed approach for turbidity mitigation is able to mitigate turbidity to some extent but it still suffers from certain limitations such as its dependence on segmentation masks, loss of fine textural details in images, and partial turbidity mitigation only. Such factors limit its applicability in real clinical scenarios. Further research is required to develop a robust algorithm that is unsupervised and can leverage temporal information of debris across adjacent frames to effectively disentangle turbidity whilst preserving the foreground structure. The development of a robust application-specific image quality metric is also significantly important to evaluate the effectiveness of the developed approaches for turbidity mitigation.
- **Design of comprehensive public datasets:** This is a new field of research and there are no publicly available datasets, making it difficult to benchmark the developed algorithms. Therefore, a large dataset of synthetic and clinical videos should be acquired with sufficient varieties of stones so that the algorithms developed are more robust to real-life clinical scenarios. The availability of large datasets is also helpful for the validation of algorithms before entering the clinical trials phase for deployment in the clinical setting.

7.3 Conclusion

This thesis lays the foundation and paves the way for future research in the field of computer-guided ureteroscopic laser lithotripsy by presenting novel computer vision-based approaches for two important tasks that can help the surgical team perform the procedure more efficiently - segmentation and image enhancement. Segmentation of stone fragments is the primary step for estimation of the stone size which is the most important parameter that the surgeon determines in real-time to decide when the fragmentation procedure can be stopped. In addition, automated segmentation of stones can also help the surgeon in tracking the stone fragments and ensuring that there are no residual fragments left. Segmentation of laser fibre is also quite important considering that it normally has a transparent glass tip that may become less visible during lithotripsy due to the

presence of high turbidity and debris. For optimal visualisation during the procedure, the surgical team is often required to increase the flow of the irrigation in the cavity but this increases the intrarenal pressure and can potentially lead to venous, tubular, and lymphatic backflow or septic complications. Therefore, this thesis has made a contribution towards the development of image enhancement algorithms by proposing a framework that can mitigate turbidity and improve vision whilst keeping the irrigation flow low. Finally, through this thesis, I would like to highlight that the clinical workload of endoscopists can be tackled by the development of medical image analysis tools that can shorten procedure time whilst improving diagnosis and therapy.

References

- [1] Masashi Misawa et al. “Artificial intelligence-assisted polyp detection for colonoscopy: initial experience”. In: *Gastroenterology* 154.8 (2018), pp. 2027–2029.
- [2] Chen-Ming Hsu et al. “Colorectal polyp image detection and classification through grayscale images and deep learning”. In: *Sensors* 21.18 (2021), p. 5995.
- [3] Ji Young Lee et al. “Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets”. In: *Scientific reports* 10.1 (2020), pp. 1–9.
- [4] Albert J de Groof et al. “Deep-learning system detects neoplasia in patients with Barrett’s esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking”. In: *Gastroenterology* 158.4 (2020), pp. 915–929.
- [5] Dehua Tang et al. “Development and validation of a real-time artificial intelligence-assisted system for detecting early gastric cancer: A multicentre retrospective diagnostic study”. In: *EBioMedicine* 62 (2020), p. 103146.
- [6] Tilahun Alelign et al. “Kidney stone disease: an update on current concepts”. In: *Advances in urology* 2018 (2018).
- [7] Atul Sohgaure et al. “A review on epidemiology and etiology of renal stone”. In: *Am. J. Drug Discov. Dev* 7.2 (2017), pp. 54–62.
- [8] Nicole L Miller et al. “Management of kidney stones”. In: *Bmj* 334.7591 (2007), pp. 468–472.
- [9] Pramod P Reddy et al. “Ureteroscopy: The standard of care in the management of upper tract urolithiasis in children”. In: *Indian journal of urology: IJU: journal of the Urological Society of India* 26.4 (2010), p. 555.
- [10] *Kidney stones*. June 2022. URL: <https://www.mayoclinic.org/diseases-conditions/kidney-stones/symptoms-causes/syc-20355755>.
- [11] Ali H Aldoukhi et al. “Holmium laser lithotripsy in the new stone age: dust or bust?” In: *Frontiers in surgery* 4 (2017), p. 57.
- [12] Vincent De Coninck et al. “Complications of ureteroscopy: a complete overview”. In: *World journal of urology* (2019), pp. 1–20.
- [13] Anna E Wright et al. “Ureteroscopy and stones: Current status and future expectations”. In: *World journal of nephrology* 3.4 (2014), p. 243.
- [14] Elaine M Worcester et al. “Calcium kidney stones”. In: *New England Journal of Medicine* 363.10 (2010), pp. 954–963.
- [15] *Kidney Stones Pictures: Visuals, symptoms, and treatment*. URL: <https://www.medicalnewstoday.com/articles/kidney-stones-pictures#types>.

- [16] Kristian M Black et al. “Deep learning computer vision algorithm for detecting kidney stone composition”. In: *BJU international* 125.6 (2020), pp. 920–924.
- [17] Wayne Brisbane et al. “An overview of kidney stone imaging techniques”. In: *Nature Reviews Urology* 13.11 (2016), p. 654.
- [18] Benoît Rosa et al. “An algorithm for calculi segmentation on ureteroscopic images”. In: *International journal of computer assisted radiology and surgery* 6.2 (2011), pp. 237–246.
- [19] *Ureteroscopy and holmium laser lithotripsy for kidney stones*. URL: <https://royanurology.com/services/ureteroscopy-and-holmium-laser-lithotripsy-for-kidney-stones/>.
- [20] *LithoVue™ single-use Digital Flexible Ureteroscope*. URL: <https://www.bostonscientific.com/en-US/products/Ureteroscopes/LithoVue.html>.
- [21] Petrisor Geavlete et al. “Pushing the boundaries of ureteroscopy: current status and future perspectives”. In: *Nature Reviews Urology* 11.7 (2014), p. 373.
- [22] Demetrius H Bagley et al. “Flexible ureteropyeloscopy: diagnosis and treatment in the upper urinary tract”. In: *The Journal of urology* 138.2 (1987), pp. 280–285.
- [23] Salvatore Buttice et al. “LithoVue™: A new single-use digital flexible ureteroscope”. In: *Central European Journal of Urology* (2016).
- [24] Steeve Doizi et al. “Dusting technique for lithotripsy: what does it mean?” In: *Nature Reviews Urology* 15.11 (2018), pp. 653–654.
- [25] Brian R Matлага et al. “Ureteroscopic laser lithotripsy: a review of dusting vs fragmentation with extraction”. In: *Journal of endourology* 32.1 (2018), pp. 1–6.
- [26] Nicholas J. Rukin et al. “Tips and tricks of ureteroscopy: Consensus statement. part II. Advanced ureteroscopy”. In: *Central European Journal of Urology* (2016).
- [27] Søren Kissow Lildal et al. “Consultation on kidney stones, Copenhagen 2019: aspects of intracorporeal lithotripsy in flexible ureterorenoscopy”. In: *World journal of urology* 39.6 (2021), pp. 1673–1682.
- [28] Bowen Cheng et al. “Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 12475–12485.
- [29] Prajawal Sinha et al. “Medical image segmentation: hard and soft computing approaches”. In: *SN Applied Sciences* 2.2 (2020), pp. 1–8.
- [30] Nobuyuki Otsu. “A threshold selection method from gray-level histograms”. In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66.
- [31] Richard Nock et al. “Statistical region merging”. In: *IEEE Transactions on pattern analysis and machine intelligence* 26.11 (2004), pp. 1452–1458.
- [32] B Karthicsonia et al. “Edge based segmentation in medical images”. In: *International Journal of Engineering and Advanced Technology (IJEAT) ISSN* (2019), pp. 2249–8958.
- [33] Laurent Najman et al. “Watershed of a continuous function”. In: *Signal Processing* 38.1 (1994), pp. 99–112.

- [34] Nameirakpam Dhanachandra et al. “Image segmentation using K-means clustering algorithm and subtractive clustering algorithm”. In: *Procedia Computer Science* 54 (2015), pp. 764–771.
- [35] Shervin Minaee et al. “Image segmentation using deep learning: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* (2021).
- [36] Jonathan Long et al. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.
- [37] Yading Yuan et al. “Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance”. In: *IEEE transactions on medical imaging* 36.9 (2017), pp. 1876–1886.
- [38] Hyeonwoo Noh et al. “Learning deconvolution network for semantic segmentation”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1520–1528.
- [39] Alex Kendall et al. “Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding”. In: Jan. 2017.
- [40] Olaf Ronneberger et al. “UNet: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.
- [41] Liang-Chieh Chen et al. “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
- [42] Maoke Yang et al. “Denseaspp for semantic segmentation in street scenes”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3684–3692.
- [43] Hengshuang Zhao et al. “Pyramid scene parsing network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.
- [44] Golnaz Ghiasi et al. “Laplacian pyramid reconstruction and refinement for semantic segmentation”. In: *European conference on computer vision*. Springer. 2016, pp. 519–534.
- [45] Francesco Visin et al. “Reseg: A recurrent neural network-based model for semantic segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 41–48.
- [46] Xiaodan Liang et al. “Semantic object parsing with graph lstm”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 125–143.
- [47] Liang-Chieh Chen et al. “Attention to scale: Scale-aware semantic image segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3640–3649.
- [48] Jun Fu et al. “Dual attention network for scene segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3146–3154.
- [49] Nasim Souly et al. “Semi supervised semantic segmentation using generative adversarial network”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5688–5696.

- [50] Yuan Xue et al. “SegAN: adversarial network with multi-scale L1 loss for medical image segmentation”. In: *Neuroinformatics* 16.3 (2018), pp. 383–392.
- [51] Shuai Zheng et al. “Conditional random fields as recurrent neural networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1529–1537.
- [52] Ziwei Liu et al. “Semantic image segmentation via deep parsing network”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1377–1385.
- [53] Sharib Ali et al. “A deep learning framework for quality assessment and restoration in video endoscopy”. In: *Medical image analysis* 68 (2021), p. 101900.
- [54] Debesh Jha et al. “Real-time polyp detection, localization and segmentation in colonoscopy using deep learning”. In: *IEEE Access* 9 (2021), pp. 40496–40510.
- [55] John R Van Sörnsen De Koste et al. “Renal mobility during uncoached quiet respiration: An analysis of 4DCT scans”. In: *International Journal of Radiation Oncology* Biology* Physics* 64.3 (2006), pp. 799–803.
- [56] Hanna Borgli et al. “HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy”. In: *Scientific data* 7.1 (2020), pp. 1–14.
- [57] Eva Tuba et al. “An algorithm for automated segmentation for bleeding detection in endoscopic images”. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 4579–4586.
- [58] VB Prasath. “Polyp detection and segmentation from video capsule endoscopy: A review”. In: *Journal of Imaging* 3.1 (2017), p. 1.
- [59] Yixuan Yuan et al. “Saliency based ulcer detection for wireless capsule endoscopy diagnosis”. In: *IEEE transactions on medical imaging* 34.10 (2015), pp. 2046–2057.
- [60] Mahdi Alizadeh et al. “Segmentation of small bowel tumors in wireless capsule endoscopy using level set method”. In: *2014 IEEE 27th International Symposium on Computer-Based Medical Systems*. IEEE. 2014, pp. 562–563.
- [61] Xiao Jia et al. “A study on automated segmentation of blood regions in wireless capsule endoscopy images using fully convolutional networks”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 179–182.
- [62] Tonmoy Ghosh et al. “Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images”. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 3034–3038.
- [63] Jaeyong Kang et al. “Ensemble of instance segmentation models for polyp segmentation in colonoscopy images”. In: *IEEE Access* 7 (2019), pp. 26440–26447.
- [64] Hemin Ali Qadir et al. “Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor cnn always perform better?” In: *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*. IEEE. 2019, pp. 1–6.
- [65] Sharib Ali et al. “Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy”. In: *Medical Image Analysis* 70 (2021), p. 102002.

- [66] Shuai Wang et al. “Multi-scale Context-guided Deep Network for Automated Lesion Segmentation with Endoscopy Images of Gastrointestinal Tract”. In: *IEEE Journal of Biomedical and Health Informatics* (2020).
- [67] Kun Yang et al. “Automatic polyp detection and segmentation using shuffle efficient channel attention network”. In: *Alexandria Engineering Journal* 61.1 (2022), pp. 917–926. URL: <https://www.sciencedirect.com/science/article/pii/S1110016821003148>.
- [68] LI Wuyang et al. “Joint polyp detection and segmentation with heterogeneous endoscopic data”. In: *3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021): co-located with the 17th IEEE International Symposium on Biomedical Imaging (ISBI 2021)*. CEUR-WS Team. 2021, pp. 69–79.
- [69] Nilar Thein et al. “A comparison of three preprocessing techniques for kidney stone segmentation in CT scan images”. In: *2018 11th Biomedical Engineering International Conference (BMEiCON)*. IEEE. 2018, pp. 1–5.
- [70] Avinash R Kambadakone et al. “New and evolving concepts in the imaging and management of urolithiasis: urologists’ perspective”. In: *Radiographics* 30.3 (2010), pp. 603–623.
- [71] Prema T Akkasaligar et al. “Segmentation of Kidney Stones in Medical Ultrasound Images”. In: *International Conference on Recent Trends in Image Processing and Pattern Recognition*. Springer. 2018, pp. 200–208.
- [72] PR Tamilselvi et al. “Segmentation of Calculi from Ultrasound Kidney Images by Region Indicator with Contour Segmentation Method”. In: *global journal of computer science and technology* (2012).
- [73] PR Tamilselvi et al. “A modified watershed segmentation method to segment renal calculi in ultrasound kidney images”. In: *International Journal of Intelligent Information Technologies (IJIT)* 8.1 (2012), pp. 46–61.
- [74] PR Tamiselvi. “Segmentation of renal calculi using squared euclidean distance method”. In: *international journal of scientific engineering and technology* 2.7 (2013), pp. 651–655.
- [75] Nilar Thein et al. “An image preprocessing method for kidney stone segmentation in CT scan images”. In: *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*. IEEE. 2018, pp. 147–150.
- [76] Prema T Akkasaligar et al. “Kidney stone detection in computed tomography images”. In: *2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon)*. IEEE. 2017, pp. 353–356.
- [77] Jianfei Liu et al. “Computer-aided detection of renal calculi from noncontrast CT images using TV-flow and MSER features”. In: *Medical physics* 42.1 (2015), pp. 144–153.
- [78] Martin Långkvist et al. “Computer aided detection of ureteral stones in thin slice computed tomography volumes using convolutional neural networks”. In: *Computers in biology and medicine* 97 (2018), pp. 153–160.
- [79] P Pallavi et al. “Automatic segmentation of kidney stones using active contour method by machine learning”. In: *AIP Conference Proceedings*. Vol. 2405. 1. AIP Publishing LLC. 2022, p. 020001.

- [80] Soumya Gupta et al. “Motion induced segmentation of stone fragments in ureteroscopy video”. In: *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 11315. International Society for Optics and Photonics. 2020, p. 1131514.
- [81] Soumya Gupta et al. “MI-UNet: Improved Segmentation in Ureteroscopy”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 212–216.
- [82] Soumya Gupta et al. “Multi-class motion-based semantic segmentation for ureteroscopy and laser lithotripsy”. In: *Computerized Medical Imaging and Graphics* 101 (2022), p. 102112. URL: <https://www.sciencedirect.com/science/article/pii/S0895611122000829>.
- [83] S. Galic et al. “Spatio-temporal image segmentation using optical flow and clustering algorithm”. In: *IWISPA 2000. Proceedings of the First International Workshop on Image and Signal Processing and Analysis, in conjunction with 22nd International Conference on Information Technology Interfaces. (IEEE (2000))*, pp. 63–68.
- [84] Tvrtko Macan et al. “Hybrid optical flow and segmentation technique for LV motion detection”. In: *Medical Imaging 2001: Physiology and Function from Multidimensional Images*. Ed. by Chin-Tu Chen et al. Vol. 4321. International Society for Optics and Photonics. SPIE, 2001, pp. 475–482. URL: <https://doi.org/10.1117/12.428175>.
- [85] Valérie Duay et al. “Dense Deformation Field Estimation for Atlas Registration using the Active Contour Framework”. In: *European Signal Processing Conference* (Jan. 2006).
- [86] Andrés Bruhn et al. “Combining the Advantages of Local and Global Optic Flow Methods”. In: *Pattern Recognition*. Ed. by Luc Van Gool. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 454–462.
- [87] Bruce D Lucas et al. *An iterative image registration technique with an application to stereo vision*. Vol. 81. Vancouver, 1981.
- [88] J. L. Barron et al. “Performance of optical flow techniques”. In: *International Journal of Computer Vision* 12.1 (1994), pp. 43–77. URL: <https://doi.org/10.1007/BF01420984>.
- [89] Berthold KP Horn et al. “Determining optical flow”. In: *Techniques and Applications of Image Understanding*. Vol. 281. International Society for Optics and Photonics. 1981, pp. 319–331.
- [90] Simon Baker et al. “A database and evaluation methodology for optical flow”. In: *International journal of computer vision* 92.1 (2011), pp. 1–31.
- [91] Thomas Brox et al. “High Accuracy Optical Flow Estimation Based on a Theory for Warping”. In: *ECCV*. 2004.
- [92] Li Xu et al. *Motion Detail Preserving Optical Flow Estimation*. 2012.
- [93] L. Mei et al. “Illumination-Invariance Optical Flow Estimation Using Weighted Regularization Transform”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2019), pp. 1–1.
- [94] C. Zach et al. “A Duality Based Approach for Realtime TV-L1 Optical Flow”. In: *Proceedings of the 29th DAGM Conference on Pattern Recognition*. Heidelberg, Germany: Springer-Verlag, 2007, pp. 214–223. URL: <http://dl.acm.org/citation.cfm?id=1771530.1771554>.

- [95] Ce Liu et al. “Sift flow: Dense correspondence across scenes and its applications”. In: *IEEE transactions on pattern analysis and machine intelligence* 33.5 (2010), pp. 978–994.
- [96] Thomas Brox et al. “Large displacement optical flow: descriptor matching in variational motion estimation”. In: *IEEE transactions on pattern analysis and machine intelligence* 33.3 (2010), pp. 500–513.
- [97] M. Drulea et al. “Motion Estimation Using the Correlation Transform”. In: *IEEE Transactions on Image Processing* 22.8 (Aug. 2013), pp. 3260–3270.
- [98] Zhuoyuan Chen et al. “Large displacement optical flow from nearest neighbor fields”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 2443–2450.
- [99] Oliver Demetz et al. *The Complete Rank Transform: A Tool for Accurate and Morphologically Invariant Matching of Structures*. 2013.
- [100] Mattias P. Heinrich et al. “MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration”. In: *Medical Image Analysis* 16.7 (2012), pp. 1423–1435.
- [101] Bartłomiej W Papież et al. “Motion correction of intravital microscopy of preclinical lung tumour imaging using multichannel structural image descriptor”. In: *Biomedical Image Registration: 6th International Workshop, WBIR 2014, London, UK, July 7-8, 2014. Proceedings* 6. Springer. 2014, pp. 164–173.
- [102] Sharib Ali et al. “Comparison of Self-similarity Measures for Multi-modal Non-rigid Registration of 3D-PLI Brain Images”. In: *Bildverarbeitung für die Medizin 2018*. 2018, pp. 49–54.
- [103] Mattias P Heinrich et al. “Non-local shape descriptor: A new similarity metric for deformable multi-modal registration”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011: 14th International Conference, Toronto, Canada, September 18-22, 2011, Proceedings, Part II* 14. Springer. 2011, pp. 541–548.
- [104] Mattias Paul Heinrich et al. “Towards realtime multimodal fusion for image-guided interventions using self-similarities”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I* 16. Springer. 2013, pp. 187–194.
- [105] Joachim Weickert et al. “A survey on variational optic flow methods for small displacements”. In: *Mathematical models for registration and applications to medical imaging*. Springer, 2006, pp. 103–136.
- [106] Luis Miguel Álvarez León et al. “A PDE model for computing the optical flow”. In: (1999).
- [107] Hans-Hellmut Nagel et al. “An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5 (1986), pp. 565–593.
- [108] Christoph Schnörr et al. “A nonlinear regularization approach to early vision”. In: *Biological cybernetics* 72.2 (1994), pp. 141–149.
- [109] Manuel Werlberger et al. “Motion estimation with non-local total variation regularization”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 2464–2471.

- [110] Jiangjian Xiao et al. “Bilateral filtering-based optical flow estimation with occlusion detection”. In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part I* 9. Springer. 2006, pp. 211–224.
- [111] Bartłomiej W Papież et al. “Complex lung motion estimation via adaptive bilateral filtering of the deformation field”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22–26, 2013, Proceedings, Part III* 16. Springer. 2013, pp. 25–32.
- [112] Bartłomiej W Papież et al. “An implicit sliding-motion preserving regularisation via bilateral filtering for deformable image registration”. In: *Medical image analysis* 18.8 (2014), pp. 1299–1311.
- [113] Md. Zahangir Alom et al. “Recurrent Residual Convolutional Neural Network based on UNet (R2UNet) for Medical Image Segmentation”. In: (Feb. 2018).
- [114] F. Yu et al. “Multi-Scale Context Aggregation by Dilated Convolutions”. In: *CoRR* abs/1511.07122 (2016).
- [115] Ozan Oktay et al. “Attention UNet: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999* (2018).
- [116] Z. Zeng et al. “RIC-Unet: An Improved Neural Network Based on Unet for Nuclei Segmentation in Histology Images”. In: *IEEE Access* (2019).
- [117] Liansheng Wang et al. “IDDF2018-ABS-0261 Polyp detection using an unet based model”. In: *Gut* 67 (2018).
- [118] Özgün Çiçek et al. “3D UNet: Learning Dense Volumetric Segmentation from Sparse Annotation”. In: *CoRR* abs/1606.06650 (2016).
- [119] Max-Heinrich Laves et al. “A Dataset of Laryngeal Endoscopic Images with Comparative Study on Convolution Neural Network Based Semantic Segmentation”. In: *CoRR* abs/1807.06081 (2018). URL: <http://arxiv.org/abs/1807.06081>.
- [120] Fausto Milletari et al. “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571.
- [121] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [122] Zhengxin Zhang et al. “Road extraction by deep residual UNet”. In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.
- [123] Gal Peretz et al. *brain-tumor-segmentation*. Available at <https://github.com/galprz/brain-tumor-segmentation>. 2019. URL: <https://github.com/galprz/brain-tumor-segmentation>.
- [124] Fabian Isensee et al. “OR-UNet: an Optimized Robust Residual UNet for Instrument Segmentation in Endoscopic Images”. In: *arXiv preprint arXiv:2004.12668* (2020).
- [125] Ming Liang et al. “Recurrent convolutional neural network for object recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3367–3375.

- [126] Ryuhei Hamaguchi et al. “Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery”. In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 1442–1450.
- [127] Shengyuan Piao et al. “Accuracy improvement of unet based on dilated convolution”. In: *Journal of Physics: Conference Series*. Vol. 1345. IOP Publishing. 2019, p. 052066.
- [128] Foivos I Diakogiannis et al. “Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 162 (2020), pp. 94–114.
- [129] Quang Nguyen et al. “Colorectal segmentation using multiple encoder-decoder network in colonoscopy images”. In: *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE. 2018, pp. 208–211.
- [130] Xinzi Sun et al. “Colorectal Polyp Segmentation by UNet with Dilation Convolution”. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE. 2019, pp. 851–858.
- [131] Peter Anderson et al. “Bottom-up and top-down attention for image captioning and visual question answering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6077–6086.
- [132] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.
- [133] Fei Wang et al. “Residual attention network for image classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3156–3164.
- [134] LeeJunHyun. *Image-Segmentation*. Available at https://github.com/LeeJunHyun/Image_Segmentation. 2019. URL: https://github.com/LeeJunHyun/Image_Segmentation.
- [135] Debesh Jha et al. “Resunet++: An advanced architecture for medical image segmentation”. In: *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE. 2019, pp. 225–2255.
- [136] Yabo Fu et al. “Deep learning in medical image registration: a review”. In: *Physics in Medicine & Biology* (2020).
- [137] Chen Qin et al. “Joint learning of motion estimation and segmentation for cardiac MR image sequences”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 472–480.
- [138] Dwarikanath Mahapatra et al. “Joint registration and segmentation of xray images using generative adversarial networks”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2018, pp. 73–80.
- [139] J. Chen et al. “WLD: A Robust Local Image Descriptor”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), pp. 1705–1720.
- [140] Oliver Demetz et al. “The Complete Rank Transform: A Tool for Accurate and Morphologically Invariant Matching of Structures.” In: *BMVC*. 2013.
- [141] Philippe Weinzaepfel et al. “DeepFlow: Large displacement optical flow with deep matching”. In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 1385–1392.

- [142] Jean-Philippe Thirion. “Image matching as a diffusion process: an analogy with Maxwell’s demons”. In: (1998).
- [143] Tom Vercauteren et al. “Diffeomorphic demons: Efficient non-parametric image registration”. In: *NeuroImage* 45.1 (2009), S61–S72.
- [144] Roger P Woods et al. “Automated image registration: I. General methods and intrasubject, intramodality validation”. In: *Journal of computer assisted tomography* 22.1 (1998), pp. 139–152.
- [145] Dinggang Shen et al. “HAMMER: hierarchical attribute matching mechanism for elastic registration”. In: *IEEE transactions on medical imaging* 21.11 (2002), pp. 1421–1439.
- [146] J Andersson et al. “FNIRT–FMRIB’s non-linear image registration tool”. In: *Human Brain Mapping* 2008 (2008).
- [147] He Wang et al. “Validation of an accelerated ‘demons’ algorithm for deformable image registration in radiation therapy”. In: *Physics in Medicine & Biology* 50.12 (2005), p. 2887.
- [148] Peter Rogelj et al. “Symmetric image registration”. In: *Medical image analysis* 10.3 (2006), pp. 484–493.
- [149] Brian B Avants et al. “Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain”. In: *Medical image analysis* 12.1 (2008), pp. 26–41.
- [150] Brian B Avants et al. “A reproducible evaluation of ANTs similarity metric performance in brain image registration”. In: *Neuroimage* 54.3 (2011), pp. 2033–2044.
- [151] Jun Zhang. “Inverse-consistent deep networks for unsupervised deformable image registration”. In: *arXiv preprint arXiv:1809.03443* (2018).
- [152] Alexey Dosovitskiy et al. “Flownet: Learning optical flow with convolutional networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2758–2766.
- [153] Sergey Zagoruyko et al. “Learning to compare image patches via convolutional neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4353–4361.
- [154] Jure Zbontar et al. “Computing the stereo matching cost with a convolutional neural network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1592–1599.
- [155] Xiaohuan Cao et al. “Deformable image registration using a cue-aware deep regression network”. In: *IEEE Transactions on Biomedical Engineering* 65.9 (2018), pp. 1900–1911.
- [156] Xiao Yang et al. “Quicksilver: Fast predictive image registration—a deep learning approach”. In: *NeuroImage* 158 (2017), pp. 378–396.
- [157] Siyuan Shan et al. “Unsupervised end-to-end learning for deformable medical image registration”. In: *arXiv preprint arXiv:1711.08608* (2017).
- [158] Guha Balakrishnan et al. “Voxelmorph: a learning framework for deformable medical image registration”. In: *IEEE transactions on medical imaging* 38.8 (2019), pp. 1788–1800.
- [159] Bob D de Vos et al. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical image analysis* 52 (2019), pp. 128–143.

- [160] Sharib Ali et al. “Conv2Warp: An unsupervised deformable image registration with continuous convolution and warping”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2019, pp. 489–497.
- [161] Geoffrey E Hinton et al. “Reducing the dimensionality of data with neural networks”. In: *science* 313.5786 (2006), pp. 504–507.
- [162] Diederik P Kingma et al. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).
- [163] Michael Crawshaw. “Multi-task learning with deep neural networks: A survey”. In: *arXiv preprint arXiv:2009.09796* (2020).
- [164] Veronica Corona et al. “Enhancing joint reconstruction and segmentation with non-convex Bregman iteration”. In: *Inverse Problems* 35.5 (2019), p. 055001.
- [165] Nikhil Kumar Tomar et al. “DDANet: Dual decoder attention network for automatic polyp segmentation”. In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 307–314.
- [166] Adrian Galdran et al. “Double Encoder-Decoder Networks for Gastrointestinal Polyp Segmentation”. In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 293–307.
- [167] Deng-Ping Fan et al. “Pranet: Parallel reverse attention network for polyp segmentation”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2020, pp. 263–273.
- [168] Michael Yeung et al. “Focus UNet: A novel dual attention-gated CNN for polyp segmentation during colonoscopy”. In: *Computers in Biology and Medicine* 137 (2021), p. 104815.
- [169] Yizhe Zhu et al. “S3vae: Self-supervised sequential vae for representation disentanglement and data generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6538–6547.
- [170] Simon Kohl et al. “A probabilistic UNet for segmentation of ambiguous images”. In: *Advances in neural information processing systems* 31 (2018).
- [171] Ke Li et al. “3D UNet Brain Tumor Segmentation Using VAE Skip Connection”. In: *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*. IEEE. 2020, pp. 97–101.
- [172] Julien Mairal et al. “Non-local sparse models for image restoration”. In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 2272–2279.
- [173] Leonid I Rudin et al. “Nonlinear total variation based noise removal algorithms”. In: *Physica D: nonlinear phenomena* 60.1-4 (1992), pp. 259–268.
- [174] A. Buades et al. “A non-local algorithm for image denoising”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 2. 2005, 60–65 vol. 2.
- [175] Kostadin Dabov et al. “Image denoising by sparse 3-D transform-domain collaborative filtering”. In: *IEEE Transactions on image processing* 16.8 (2007), pp. 2080–2095.
- [176] Revital Ironi et al. “Colorization by Example.” In: *Rendering techniques* 29 (2005), pp. 201–210.

- [177] Tomihisa Welsh et al. “Transferring color to greyscale images”. In: *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. 2002, pp. 277–280.
- [178] Dana Berman et al. “Non-local image dehazing”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1674–1682.
- [179] Ye Xin et al. “Specular reflection image enhancement based on a dark channel prior”. In: *IEEE Photonics Journal* 13.1 (2021), pp. 1–11.
- [180] Xiang-Yu Kong et al. “Low-light image enhancement via poisson noise aware retinex model”. In: *IEEE Signal Processing Letters* 28 (2021), pp. 1540–1544.
- [181] Mading Li et al. “Structure-revealing low-light image enhancement via robust retinex model”. In: *IEEE Transactions on Image Processing* 27.6 (2018), pp. 2828–2841.
- [182] Wenjie Li et al. “Endoscopy image enhancement method by generalized imaging defect models based adversarial training”. In: *Physics in Medicine & Biology* 67.9 (2022), p. 095016.
- [183] Kai Zhang et al. “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising”. In: *IEEE transactions on image processing* 26.7 (2017), pp. 3142–3155.
- [184] Martin Weigert et al. “Content-aware image restoration: pushing the limits of fluorescence microscopy”. In: *Nature methods* 15.12 (2018), pp. 1090–1097.
- [185] Chongyi Li et al. “LightenNet: A convolutional neural network for weakly illuminated image enhancement”. In: *Pattern recognition letters* 104 (2018), pp. 15–22.
- [186] Bolun Cai et al. “Dehazenet: An end-to-end system for single image haze removal”. In: *IEEE Transactions on Image Processing* 25.11 (2016), pp. 5187–5198.
- [187] Dmitry Ulyanov et al. “Deep image prior”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9446–9454.
- [188] Jaakko Lehtinen et al. “Noise2Noise: Learning image restoration without clean data”. In: *arXiv preprint arXiv:1803.04189* (2018).
- [189] Zhengwei Wang et al. “Generative adversarial networks: A survey and taxonomy”. In: *arXiv preprint arXiv:1906.01529* 2 (2019).
- [190] Alexander Krull et al. “Noise2void-learning denoising from single noisy images”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2129–2137.
- [191] Dufan Wu et al. “Consensus neural network for medical imaging denoising with only noisy training samples”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 741–749.
- [192] Long Chen et al. “De-smokeGCN: generative cooperative networks for joint surgical smoke detection and removal”. In: *IEEE transactions on medical imaging* 39.5 (2019), pp. 1615–1625.
- [193] Yifan Jiang et al. “Enlightengan: Deep light enhancement without paired supervision”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 2340–2349.
- [194] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

- [195] Qijing You et al. “Fundus image enhancement method based on CycleGAN”. In: *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2019, pp. 4500–4503.
- [196] Yuhui Ma et al. “Cycle structure and illumination constrained GAN for medical image enhancement”. In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2020, pp. 667–677.
- [197] Zheng Ding et al. “Guided variational autoencoder for disentanglement learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7920–7929.
- [198] Nao Nakagawa et al. “Disentangled Representation Learning in Real-World Image Datasets via Image Segmentation Prior”. In: *IEEE Access* 9 (2021), pp. 110880–110888.
- [199] Irina Higgins et al. “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework”. In: *ICLR*. 2017.
- [200] Christopher P Burgess et al. “Understanding disentangling in β -VAE”. In: *arXiv preprint arXiv:1804.03599* (2018).
- [201] Ricky TQ Chen et al. “Isolating sources of disentanglement in variational autoencoders”. In: *Advances in neural information processing systems* 31 (2018).
- [202] Abhishek Kumar et al. “Variational inference of disentangled latent concepts from unlabeled observations”. In: *arXiv preprint arXiv:1711.00848* (2017).
- [203] Francesco Locatello et al. “Challenging common assumptions in the unsupervised learning of disentangled representations”. In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.
- [204] Tristan Bepler et al. “Explicitly disentangling image content from translation and rotation with spatial-VAE”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [205] Nicki Skafted Detlefsen et al. “Explicit Disentanglement of Appearance and Perspective in Generative Models”. In: *arXiv e-prints* (2019), arXiv–1906.
- [206] Volker Blanz et al. “A morphable model for the synthesis of 3D faces”. In: *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 1999, pp. 187–194.
- [207] Yu Deng et al. “Disentangled and controllable face image generation via 3d imitative-contrastive learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5154–5163.
- [208] Klaus Greff et al. “Multi-object representation learning with iterative variational inference”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2424–2433.
- [209] M. Awiszus et al. “Learning Disentangled Representations via Independent Subspaces”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Los Alamitos, CA, USA: IEEE Computer Society, 2019, pp. 560–568. URL: <https://doi.ieeecomputersociety.org/10.1109/ICCVW.2019.00069>.
- [210] PR Tamilselvi et al. “An efficient segmentation of calculi from US renal calculi images using ANFIS system”. In: *European Journal of Scientific Research* 55.2 (2011), pp. 323–333.

- [211] Jyoti Verma et al. “Analysis and identification of kidney stone using Kth nearest neighbour (KNN) and support vector machine (SVM) classification techniques”. In: *Pattern Recognition and Image Analysis* 27 (2017), pp. 574–580.
- [212] John R. Van Sörnsen De Koste et al. “Renal mobility during uncoached quiet respiration: An analysis of 4DCT scans”. In: *International Journal of Radiation Oncology Biology Physics* (2006).
- [213] Sharib Ali et al. “Illumination invariant optical flow using neighborhood descriptors”. In: *Computer Vision and Image Understanding* 145 (Dec. 2015).
- [214] Antonin Chambolle et al. “A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging”. In: *Journal of Mathematical Imaging and Vision* 40.1 (May 2011), pp. 120–145. URL: <https://doi.org/10.1007/s10851-010-0251-1>.
- [215] R. Rockafellar. “Monotone Operators and the Proximal Point Algorithm”. In: *SIAM Journal on Control and Optimization* 14.5 (1976), pp. 877–898. eprint: <https://doi.org/10.1137/0314056>. URL: <https://doi.org/10.1137/0314056>.
- [216] Abhishek Dutta et al. “The VIA Annotation Software for Images, Audio and Video”. In: *Proceedings of the 27th ACM International Conference on Multimedia*. Nice, France, 2019. URL: <https://doi.org/10.1145/3343031.3350535>.
- [217] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.
- [218] Alexey Bokhovkin et al. “Boundary loss for remote sensing imagery semantic segmentation”. In: *International Symposium on Neural Networks*. Springer. 2019, pp. 388–401.
- [219] Sharib Ali et al. “An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy”. In: *Scientific reports* 10.1 (2020), pp. 1–15.
- [220] Freddie Bray et al. “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: a cancer journal for clinicians* 68.6 (2018), pp. 394–424.
- [221] Venkataraman Subramanian et al. “Advanced endoscopic imaging: a review of commercially available technologies”. In: *Clinical Gastroenterology and Hepatology* 12.3 (2014), pp. 368–376.
- [222] Fausto Milletari et al. “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571.
- [223] Debesh Jha et al. “Kvasir-seg: A segmented polyp dataset”. In: *International Conference on Multimedia Modeling*. Springer. 2020, pp. 451–462.
- [224] Jorge Bernal et al. “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians”. In: *Computerized Medical Imaging and Graphics* 43 (2015), pp. 99–111.
- [225] Sharib Ali et al. “PolypGen: A multi-center polyp detection and segmentation dataset for generalisability assessment”. In: *CoRR* abs/2106.04463 (2021). arXiv: 2106.04463. URL: <https://arxiv.org/abs/2106.04463>.

- [226] Seyed Raein Hashemi et al. “Asymmetric Loss Functions and Deep Densely-Connected Networks for Highly-Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection”. In: *IEEE Access* 7 (2019), pp. 1721–1735.
- [227] J. Long et al. “Fully convolutional networks for semantic segmentation”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3431–3440.
- [228] Yunbo Guo et al. “Polyp Segmentation with Fully Convolutional Deep Neural Networks—Extended Evaluation Study”. In: *Journal of Imaging* 6.7 (2020), p. 69.
- [229] Zhan Wu et al. “ELNet: Automatic classification and segmentation for esophageal lesions using convolutional neural network”. In: *Medical Image Analysis* 67 (2021), p. 101838.
- [230] Y. Tsai et al. “Learning to Adapt Structured Output Space for Semantic Segmentation”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7472–7481.
- [231] Nanqing Dong et al. “Unsupervised Domain Adaptation for Automatic Estimation of Cardiothoracic Ratio”. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2018, pp. 544–552.
- [232] Mohammad Minhazul Haq et al. “Adversarial Domain Adaptation for Cell Segmentation”. In: *Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL)*. 2020, pp. 277–287.
- [233] Prashant Pandey et al. “Unsupervised Domain Adaptation for Semantic Segmentation of NIR Images Through Generative Latent Search”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020, pp. 413–429.
- [234] Numan Celik et al. “Endouda: a modality independent segmentation approach for endoscopy imaging”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 303–312.
- [235] Mingxing Tan et al. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of International Conference on Machine Learning (ICML)*. 2019, pp. 6105–6114.
- [236] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.
- [237] Zongwei Zhou et al. “Unet++: Redesigning skip connections to exploit multiscale features in image segmentation”. In: *IEEE transactions on medical imaging* 39.6 (2019), pp. 1856–1867.
- [238] B. Baheti et al. “Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment”. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 1473–1481.
- [239] Tuan-Hung Vu et al. “ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2512–2521.
- [240] Charles D Scales et al. “Comparative effectiveness of shock wave lithotripsy and ureteroscopy for treating patients with kidney stones”. In: *JAMA surgery* 149.7 (2014), pp. 648–653.
- [241] Nathaniel M Fried. “Recent advances in infrared laser lithotripsy”. In: *Biomedical optics express* 9.9 (2018), pp. 4552–4568.

- [242] Gaofeng Meng et al. “Efficient image dehazing with boundary constraint and contextual regularization”. In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 617–624.
- [243] Dat Ngo et al. “Single image haze removal from image enhancement perspective for real-time vision-based systems”. In: *Sensors* 20.18 (2020), p. 5170.
- [244] H.R. Sheikh et al. “Image information and visual quality”. In: *IEEE Transactions on Image Processing* 15.2 (2006), pp. 430–444.
- [245] Vittorio Baroncini et al. “The polar edge coherence: a quasi blind metric for video quality assessment”. In: *2009 17th European Signal Processing Conference*. IEEE. 2009, pp. 564–568.