

# Sli2Vol: Annotate a 3D Volume from a Single Slice with Self-Supervised Learning

Pak-Hei Yeung<sup>1</sup>, Ana I.L. Namburete<sup>1+</sup>, and Weidi Xie<sup>1,2+</sup>

<sup>1</sup> Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, United Kingdom

`pak.yeung@pmb.ox.ac.uk`, `ana.namburete@eng.ox.ac.uk`

<sup>2</sup> Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford, United Kingdom  
`weidi@robots.ox.ac.uk`

[https://pakheiyung.github.io/Sli2Vol\\_wp/](https://pakheiyung.github.io/Sli2Vol_wp/)

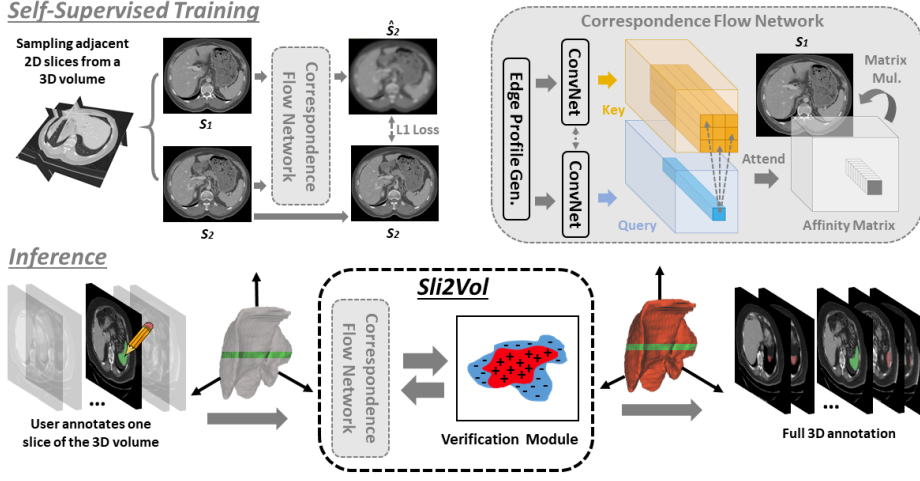
**Abstract.** The objective of this work is to segment any *arbitrary* structures of interest (SOI) in 3D volumes by only annotating a *single* slice, (*i.e.* semi-automatic 3D segmentation). We show that high accuracy can be achieved by simply propagating the 2D slice segmentation with an affinity matrix between consecutive slices, which can be learnt in a self-supervised manner, namely slice reconstruction. Specifically, we compare our proposed framework, termed as **Sli2Vol**, with supervised approaches and two other unsupervised/ self-supervised slice registration approaches, on 8 public datasets (both CT and MRI scans), spanning 9 different SOIs. Without any parameter-tuning, the same model achieves superior performance with Dice scores (0-100 scale) of over 80 for most of the benchmarks, including the ones that are unseen during training. Our results show *generalizability* of the proposed approach across data from different machines and with different SOIs: a major use case of semi-automatic segmentation methods where fully supervised approaches would normally struggle.

**Keywords:** Self-supervised learning · Semi-automatic segmentation.

## 1 Introduction

Image segmentation is arguably one of the most important tasks in medical image analysis, as it identifies the structure of interest (SOI) with arbitrary shape (*i.e.* pixel level predictions), encompassing rich information, such as the position and size. In recent years, the development and application of different deep convolutional neural networks (ConvNet), for example U-Net [19], have significantly boosted the accuracy of computer-aided medical image segmentation.

Training fully automatic segmentation models comes with several limitations: *firstly*, annotations for the training volumes are usually a costly process to acquire; *secondly*, once domain shift appears, (*i.e.* from differences in scanner, acquisition protocol or the SOI varies during inference), the model may suffer



**Fig. 1.** Pipeline of our proposed framework. During *self-supervised training*, pair of adjacent slices sampled from 3D volumes are used to train a correspondence flow network. Provided with the 2D mask of a single slice of a volume, the trained network with the verification module can be used to propagate the initial annotation to the whole volume during *inference*.

a catastrophic drop in performance, requiring new annotations and additional fine-tuning. These factors have limited the use of the automatic segmentation approaches to applications with inter-vendor and inter-operator variance.

As an alternative, semi-automatic approaches are able to operate interactively with the end users: this is the scenario considered in this paper. Specifically, the goal is to segment any *arbitrary* SOIs in 3D volumes by only annotating a *single* slice within the volume, which may facilitate more flexible analysis of *arbitrary* SOIs with the desired generalizability (*e.g.* inter-scanner variability), and significantly reduce the annotating cost for fully supervised learning.

Similar tools have been developed with level set or random forest methods, which show excellent performance as reported in [5, 7, 17, 26, 27]. However, implementation of specific regularization and heavy parameter-tuning are usually required for different SOIs, limiting its use in practice. On the other hand, related work in medical image registration explores the use of pixelwise correspondence from optical flow [10, 14] or unsupervised approaches [3, 8, 18], which in principle could be harnessed for the propagation of a 2D mask between slices within a volume. However, they are prone to error drift, *i.e.* error accumulation, introduced by inter-slice propagation of registration errors. In this work, we aim to overcome these limitations.

Here, we focus on the task of propagating the 2D slice segmentation through the entire 3D volume by matching correspondences between consecutive slices. Our work makes the following contributions: *firstly*, we explore mask propa-

gation approaches based on unsupervised/self-supervised registration of slices, namely, naïve optical flow [6] and VoxelMorph [3], and our proposed self-supervised approach, called **Sli2Vol**, which is based on learning to match slices’ correspondences [15, 16] and using a newly proposed edge profile for information bottleneck. **Sli2Vol** is able to propagate the mask at a speed of 2.27 slices per second in inference. *Secondly*, to alleviate the error accumulation in mask propagation, we propose and exploit a simple verification module for refining the mask during inference time. *Thirdly*, we benchmark **Sli2Vol** on 8 public CT and MRI datasets [12, 22, 23, 25], spanning 9 anatomical structures. Without any parameter-tuning, a *single Sli2Vol* model achieves Dice scores (0-100 scale) above 80 for most of the benchmarks, which outperforms other supervised and unsupervised approaches for all datasets in cross-domain evaluation. To the best of our knowledge, this is the first study to undertake cross-domain evaluation on such large-scale and diverse benchmarks for semi-automatic segmentation approaches, which shifts the focus to *generalizability* across different devices, clinical sites and anatomical SOIs.

## 2 Methods

In Section 2.1, we first formulate the problem setting in this paper, namely semi-automatic segmentation for 3D volume with *single* slice annotation. Next, we introduce the training stage of our proposed approach, **Sli2Vol**, in Section 2.2 and our proposed edge profile generator in Section 2.3. This is followed by the computations for inference (2.4), including our proposed verification module (2.5).

### 2.1 Problem Setup

In general, given a 3D volume, denoted by  $\mathbf{V} \in \mathcal{R}^{H \times W \times D}$ , where  $H$ ,  $W$  and  $D$  are the height, width and depth of the volume, respectively, our goal is to segment the SOI in the volume based on a user-provided 2D segmentation mask for the *single* slice, *i.e.*  $\mathbf{M}_i \in \mathcal{R}^{H \times W \times 1}$  with 1’s indicating the SOI, and 0’s as background. The outputs will be a set of masks for an individual slice, *i.e.*  $\{\mathbf{M}_1, \mathbf{M}_2, \dots, \mathbf{M}_D\}$ .

Inspired by [15, 16], we formulate this problem as learning feature representations that establish robust pixelwise correspondences between adjacent slices in a 3D volume, which results in a set of affinity matrices,  $\mathbf{A}_{i \rightarrow i+1}$ , for propagating the 2D mask between consecutive slices by *weighting and copying*. Model training follows a self-supervised learning scheme, where raw data is used, and only one slice annotation is required during inference time.

### 2.2 Self-Supervised Training of Sli2Vol

In this section, we detail the self-supervised approach for learning the dense correspondences. Conceptually, the idea is to task a deep network for slice reconstruction by *weighting and copying* pixels from its neighboring slice. The

affinity matrices used for weighting are acquired as a by-product, and can be directly used for mask propagation during inference.

During training, a pair of adjacent slices,  $\{\mathbf{S}_1, \mathbf{S}_2\}$ ,  $\mathbf{S}_i \in \mathcal{R}^{H \times W \times 1}$ , are sampled from a training volume, and then fed to a ConvNet, parametrized by  $\psi(\cdot, \theta)$  (as shown in the upper part of Fig. 1):

$$[\mathbf{k}_1, \mathbf{q}_2] = [\psi(g(\mathbf{S}_1); \theta), \psi(g(\mathbf{S}_2); \theta)] \quad (1)$$

where  $g(\cdot)$  denotes an *edge profile generator* (details in Section 2.3) and  $\mathbf{k}_1, \mathbf{q}_2 \in \mathcal{R}^{H \times W \times c}$  refer to the feature representation ( $c$  channels) computed from corresponding slices, termed as *key* and *query* respectively (Fig. 1). The difference in notation (*i.e.*  $\mathbf{q}$  and  $\mathbf{k}$ ) is just for emphasizing their functional difference.

Reshaping  $\mathbf{k}_1$  and  $\mathbf{q}_2$  to  $\mathcal{R}^{HW \times c}$ , an affinity matrix,  $\mathbf{A}_{1 \rightarrow 2} \in \mathcal{R}^{HW \times \delta}$ , is computed to represent the feature similarity between the two slices (Fig. 1):

$$\mathbf{A}_{1 \rightarrow 2}(u, v) = \frac{\exp\langle \mathbf{q}_2(u, :), \mathbf{k}_1(v, :) \rangle}{\sum_{\lambda \in \Omega} \exp\langle \mathbf{q}_2(u, :), \mathbf{k}_1(\lambda, :) \rangle} \quad (2)$$

where  $\langle \cdot, \cdot \rangle$  is the dot product between two vectors and  $\Omega$  is the window surrounding pixel  $v$  (*i.e.* in  $\mathcal{R}^{H \times W}$  space) for computing local attention, with  $n(\Omega) = \delta$ .

**Loss Function.** During training,  $\mathbf{A}_{1 \rightarrow 2}$  is used to *weight and copy* pixels from  $\mathbf{S}_1$  (*i.e.* reshape to  $\mathcal{R}^{HW \times 1}$ ) to reconstruct  $\mathbf{S}_2$ , denoted as  $\hat{\mathbf{S}}_2$ , by:

$$\hat{\mathbf{S}}_2(u, 1) = \sum_v^{\Omega} \mathbf{A}_{1 \rightarrow 2}(u, v) \mathbf{S}_1(v, 1). \quad (3)$$

We apply mean absolute error (MAE) between  $\mathbf{S}_2$  and  $\hat{\mathbf{S}}_2$  as the training loss.

### 2.3 Edge Profile Generator

Essentially, the basic assumption of the above-mentioned idea is that, to better reconstruct  $\mathbf{S}_2$  via copying pixel from  $\mathbf{S}_1$ , the model must learn to establish reliable correspondences between the two slices. However, naïvely training the model may actually incur trivial solutions, for example, the model can perfectly solve the reconstruction task by simply matching the *pixel intensity* of  $\mathbf{S}_1$  and  $\mathbf{S}_2$ .

In Lai *et al.* [15, 16], the authors show that input color channel (*i.e.* *RGB* or *Lab*) dropout is an effective information bottleneck, which breaks the correlation between the color channels and forces the model to learn more robust correspondences. However, this is usually not feasible in medical images, as only single input channel is available in most of the modalities.

We propose to use a *profile of edges* as an *information bottleneck* to avoid trivial solution. Specifically, for each pixel, we convert its intensity value to a normalized edge histogram, by computing the derivatives along  $d$  different directions at  $s$  different scales, *i.e.*  $g(\mathbf{S}_i) \in \mathcal{R}^{H \times W \times (d \times s)}$ , followed by a *softmax*

normalization through all the derivatives. Intuitively,  $g(\cdot)$  explicitly represents the edge distributions centered each pixel of the slice  $\mathbf{S}_i$ , and force the model to pay more attentions to the edges during reconstruction. Experimental results in Section 4.2 verify the essence of this design in improving the model performance.

## 2.4 Inference

Given a volume,  $\mathbf{V}$  and an initial mask at the  $i$ -th slice,  $\mathbf{M}_i$ , the affinity matrix,  $\mathbf{A}_{i \rightarrow i+1}$ , output from  $\psi(\cdot, \theta)$  is used to propagate  $\mathbf{M}_i$  iteratively to the whole  $\mathbf{V}$ .

In detail, two consecutive slices,  $\{\mathbf{S}_i, \mathbf{S}_{i+1}\}$ , are sampled from the volume  $\mathbf{V}$  and fed into  $\psi(g(\cdot), \theta)$  to get  $\mathbf{A}_{i \rightarrow i+1}$ , which is then used to propagate  $\mathbf{M}_i$ , using Eq. 3, ending up with  $\hat{\mathbf{M}}_{i+1}$ . This set of computations (Fig. 2 in the Supplementary Materials in Section 6) is then repeated for the next two consecutive slices,  $\{\mathbf{S}_{i+1}, \mathbf{S}_{i+2}\}$ , in either direction, until the whole volume is covered.

## 2.5 Verification Module

In practice, we find that directly using  $\hat{\mathbf{M}}_{i+1}$  for further propagation will potentially accumulate the prediction error after each iteration. To alleviate this drifting issue, and further boost the performance, we propose a simple verification module to correct the mask after each iteration of mask propagation.

Specifically, two regions, namely positive ( $\mathbf{P} \in \mathcal{R}^{H \times W}$ ) and negative ( $\mathbf{N} \in \mathcal{R}^{H \times W}$ ) regions, are constructed.  $\mathbf{P}$  refers to the delineated SOI in  $\mathbf{M}_i$ , and  $\mathbf{N}$  is identified by subtracting  $\mathbf{P}$  from its own morphologically dilated version. Intuitively, the negative region denotes the thin and non-overlapping region surrounding  $\mathbf{P}$  (Fig. 2 in the Supplementary Materials in Section 6). We maintain the *mean intensity value* within each region:

$$p = \frac{1}{|P_i|} \langle P_i, S_i \rangle \quad n = \frac{1}{|N_i|} \langle N_i, S_i \rangle$$

where  $\langle \cdot, \cdot \rangle$  denotes Frobenius inner product,  $p$  and  $n$  refer to the positive and negative query values respectively.

During inference time, assuming  $\hat{\mathbf{M}}_{i+1}$  is the predicted mask from the propagation, each of the proposed foreground pixels  $u$  in  $\mathbf{S}_{i+1}$ , is then compared to  $p$  and  $n$  and being re-classified according to its distance to the two values by:

$$\mathbf{M}_{i+1}^u = \begin{cases} 1, & \text{if } \hat{\mathbf{M}}_{i+1}^u = 1 \text{ and } \sqrt{(\mathbf{S}_{i+1}^u - p)^2} < \sqrt{(\mathbf{S}_{i+1}^u - n)^2} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

This set of computations is then repeated for the next round of propagation, where  $p$  and  $n$  are updated using the corrected mask,  $\mathbf{M}_{i+1}$ , and  $\mathbf{S}_{i+1}$ .

## 3 Experimental Setup

We benchmark our framework, **Sli2Vol**, on 8 different public datasets, spanning 9 different SOIs, and compare with a variety of fully supervised and semi-automatic approaches, using standard Dice coefficient (in a 0-100 scale) as the

evaluation metrics. In Section 3.1, we introduce the datasets used in this paper. In Section 3.2, we summarize the experiments conducted for this study.

### 3.1 Dataset

Four training and eight testing datasets are involved. For **chest and abdominal CT**, a *single* model is trained on 3 unannotated dataset (*i.e.* C4KC-KiTS [9], CT-LN [21] and CT-Pancreas [20]) and tested on 7 other datasets (*i.e.* Sliver07 [25], CHAOS [12], 3Dircadb-01, 02 [23], and Decath-Spleen, Liver and Pancreas [22]).

For **cardiac MRI**, models are trained on the 2D video dataset from Kaggle [1], and tested on a 3D volume dataset (*i.e.* Decath-Heart [22]), which manifests large domain shift. Further details of the datasets are provided in Table 2 in the Supplementary Materials in Section 6.

### 3.2 Baseline Comparison

**Sli2Vol** and a set of baseline approaches are tested, with their implementation details summarized in Table 3 in the Supplementary Materials in Section 6.

*First*, we experiment with two approaches trained on fully annotated 3D data. **Fully Supervised (FS) - Same Domain** refers to the scenario where the training and testing data come from the *same* benchmark dataset. Results from both state-of-the-art methods [2, 11, 13, 24] and 3D UNets trained by us are reported. On the other hand, **FS - Different Domain** aims to evaluate the generalizability of FS approaches when training and testing data come from *different* domains. Therefore, we train the 3D UNet (same architecture and settings as the **FS - Same Domain** for fair comparison) on a source dataset, and test it on another benchmark of the same task.

*Second*, we consider the case where only a single slice is annotated in each testing volume to train a 2D UNet (**FS - Single Slice**). For example, in *Sliver07*, the model trained on 20 slice annotations (single slice from each volume), is tested on the same set of 20 volumes. This approach utilizes the same amount of manual annotation as **Sli2Vol**, so as to investigate if a model trained on single slice annotations is sufficient to generalize to the whole volume.

*Third*, approaches based on registration of consecutive slices, namely **Optical Flow** [4, 6], **VoxelMorph2D (VM) - UNet** and **VM - ResNet18Stride1**, are tested. The two VMs utilize a UNet backbone as proposed originally in [3] as well as the same backbone (*i.e.* ResNet18Stride1) as **Sli2Vol**, respectively.

For **Sli2Vol**, **FS - Single Slice**, **Optical Flow** and **VM**, we randomly pick one of the  $\pm 3$  slices around the slice with the largest groundtruth annotation as the initial mask. This simulates the process of a user sliding through the whole volume and roughly identifying the slice with the largest SOI to annotate, which is achievable in reality.

Modality		MRI		Abdominal and Chest CT												
Training Dataset (for row e to j)		Kaggle		C4KC-KiTS, CT-LN and CT-Pancreas												
Testing Dataset		Decath-Heart		Decath-Liver		CHAOS	Decath-Pancreas		3D-IRCADb-01 and 3D-IRCADb-02							
ROI		Left	Right	Liver	Liver	Spleen	Pancreas	Heart	Gall-bladder	Kidney	Surrenal-gland	Liver	Lung	Pancreas	Spleen	Mean Results
Number of Volumes		20	20	20	131	41	281	3	8	17	11	22	12	4	7	
Automatic (Trained with Fully Annotated Data)																
(a) Fully Supervised-same domain		92.7 [11]	94.8 [2]	97.8 [13]	95.4 [11]	96.0 [11]	79.3 [11]	-	-	-	-	96.5 [24]	-	-	-	-
(b) Fully Supervised-different domain		-	74.8 ±13.2	76.5 ±8.8	56.0 ±23.6	-	-	-	-	-	-	-	-	-	-	-
Semi-automatic																
(c) Fully Supervised-single slice		62.5 ±5.2	86.9 ±4.1	84.3 ±4.1	85.0 ±5.5	74.4 ±12.0	49.9 ±13.4	25.6 ±6.5	47.9 ±15.5	57.9 ±21.1	30.8 ±15.6	80.3 ±13.8	81.0 ±10.8	20.4 ±7.9	58.6 ±4.7	60.4
(d) Optical Flow		51.1 ±7.4	65.2 ±8.8	72.0 ±9.9	47.0 ±15.9	72.9 ±14.5	25.1 ±8.2	32.2 ±11.6	24.6 ±12.4	73.6 ±14.6	22.1 ±12.9	68.4 ±9.4	33.6 ±18.0	21.9 ±12.6	70.8 ±17.5	48.6
(e) VoxelMorph2D-UNet		42.9 ±5.0	57.2 ±9.8	66.5 ±10.5	38.5 ±12.5	61.5 ±19.5	21.4 ±6.7	20.3 ±6.5	20.2 ±12.2	70.1 ±18.6	41.1 ±15.3	60.5 ±9.7	38.7 ±21.2	28.3 ±11.0	54.1 ±12.4	44.4
(f) VoxelMorph2D-ResNet18NoStride		45.7 ±4.1	61.2 ±8.5	68.4 ±9.8	42.2 ±12.4	58.3 ±17.3	23.5 ±7.8	22.1 ±6.7	21.8 ±13.1	77.8 ±18.4	48.4 ±15.3	60.6 ±10.4	36.5 ±20.0	32.3 ±13.3	60.0 ±12.1	47.5
Sli2Vol Ablation Studies																
(g) Correspondence Flow Network		62.4 ±9.2	75.0 ±6.5	78.9 ±7.9	66.0 ±13.1	81.1 ±13.9	43.9 ±12.9	55.4 ±24.3	62.4 ±20.7	86.0 ±19.0	45.9 ±18.6	75.0 ±8.6	45.2 ±25.4	44.3 ±17.2	81.8 ±19.6	64.5
(h) Network + Edge Profile		56.8 ±8.4	74.8 ±7.4	77.8 ±8.4	64.4 ±14.1	83.6 ±13.2	48.9 ±11.2	49.4 ±12.3	68.5 ±13.8	86.8 ±15.7	58.3 ±16.6	73.9 ±8.5	48.8 ±26.4	53.9 ±7.1	85.8 ±13.0	66.6
(i) Network + Verif. Module		80.8 ±5.0	81.1 ±5.0	83.4 ±6.3	72.0 ±8.9	79.1 ±17.3	37.3 ±13.6	50.9 ±11.6	70.7 ±12.7	83.3 ±21.4	47.5 ±20.8	78.8 ±6.9	79.8 ±29.3	45.2 ±10.5	74.5 ±23.7	68.9
(j) Network + Verif. Module + Edge Profile		80.4 ±4.5	91.3 ±3.2	91.0 ±2.9	86.8 ±7.2	88.4 ±10.9	54.2 ±10.0	75.9 ±10.9	68.9 ±9.9	91.4 ±4.8	88.2 ±13.5	81.4 ±3.0	81.4 ±28.5	58.2 ±4.6	90.2 ±9.5	78.2

**Table 1.** Results (mean Dice scores  $\pm$  standard deviation) of different approaches on different datasets and SOIs. Higher value represents better performance. In **row a**, results from both state-of-the-art methods [2, 11, 13] and 3D UNets trained by us (values in the bracket) are reported. Results in **row a** and **b** are only partially available in literature and they are reported just for demonstrating the approximated upper bound and limitation of fully supervised approaches, which are not meant to be directly compared to our proposed approach.

## 4 Results and Discussion

The results of all the experiments are presented in Table 1, with qualitative examples shown in Fig. 3 in the Supplementary Materials in Section 6. In Section 4.1, we explore the performance change of automatic approaches in the presence of domain shift, which leads to the analysis of the results of **Sli2Vol** in Section 4.2.

### 4.1 Automatic Approaches

As expected, although the state-of-the-art performance is achieved by the **FS - Same Domain (row a)**, a significant performance drop (*i.e.* over 20 Dice) can be observed (by comparing **row b** and the values in the brackets in **row a**) for cross-domain (*i.e.* same SOI, different benchmarks) evaluation (**row b**).

Such variation of performance may be partially minimized by increasing the amount and diversity of training data, better design of training augmentation, and application of domain adaptation techniques. However, these may not always be practical in real-world scenarios, due to the high cost of data annotation and frequent domain shifts, for example variation of scanners and acquisition protocols in different clinical sites.

### 4.2 Semi-automatic Approaches

**Sli2Vol**, by contrast, does not need any annotated data for training, but only annotation of a single slice during inference to indicate the SOI to be segmented.

**Single Slice Annotation.** With the same amount of annotation, **Sli2Vol (row j)** clearly outperforms other baseline approaches (**row c - f**) on all benchmarks significantly ( $p < 0.05$ , t-test), with an average Dice score margin of over 18.

**Propagation-Based Methods.** Higher Dice score shown in **row g** over **row d - f** suggests that solely self-supervised correspondence matching may incur less severe error drift and, hence, be more suitable than **Optical Flow** and **VM** for mask propagation within a volume. Comparison of results in **row e, f** and **g** further verifies that the backbone architecture is not the determining factor for the superior performance achieved by **Sli2Vol**. Our proposed edge profile (**row h**) is shown to be a more effective bottleneck than using the original slice as input (**row g**) and it further boosts the marginal benefit of the verification module, which is manifested by comparing the performance gain from **row g** to **i** and that from **row h** to **j**

**Self-Supervised Learning.** Remarkably, **Sli2Vol** trained with self-supervised learning is agnostic to SOIs and domains. As for abdominal and chest CT, a *single Sli2Vol* model without any fine-tuning achieves a mean Dice score of

78.0 when testing on 7 datasets spanning 8 anatomical structures. As for the cardiac MRI experiments with large training-testing domain shift, **Sli2Vol** still performs reasonably well with a Dice score of 80.4 (**row j**). Under this scenario, **Sli2Vol** outperforms the fully supervised approaches significantly ( $p < 0.05$ , t-test), by more than 20 Dice scores (**row j** vs. **row b**), and the annotation efforts are much lower, *i.e.* only a single slice per volume.

## 5 Conclusion

In summary, we investigate on semi-automatic 3D segmentations, where any *arbitrary* SOIs in 3D volumes are segmented by only annotating a single slice. The proposed architecture, **Sli2Vol**, is trained with self-supervised learning to output affinity matrices between consecutive slices through correspondence matching, which are then used to propagate the segmentation through the volume. Benchmarking on 8 datasets with 9 different SOIs, **Sli2Vol** shows superior generalizability and accuracy as compared to other baseline approaches, agnostic to the SOI. We envision to provide end users with more flexibility to segment and analyze different SOIs with our proposed framework, which shows great potential to be further developed as a general interactive segmentation tool in our future works, to facilitate the community to study various anatomical structures, and minimize the cost of annotating large dataset.

**Acknowledgments.** PH. Yeung is grateful for support from the RC Lee Centenary Scholarship. A. Namburete is funded by the UK Royal Academy of Engineering under its Engineering for Development Research Fellowship scheme. W. Xie is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Programme Grant Seebibyte (EP/M013774/1) and Grant Visual AI (EP/T028572/1). We thank Madeleine Wyburd and Nicola Dinsdale for their valuable suggestions and comments about the work.

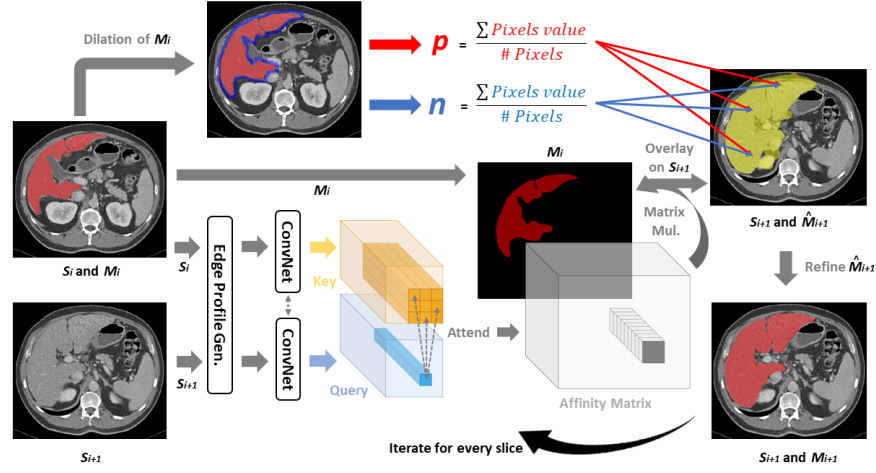
## References

1. Data science bowl cardiac challenge data, <https://www.kaggle.com/c/second-annual-data-science-bowl>
2. Ahmad, M., Ai, D., Xie, G., Qadri, S.F., Song, H., Huang, Y., Wang, Y., Yang, J.: Deep belief network modeling for automatic liver segmentation. *IEEE Access* **7**, 20585–20595 (2019)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging* **38**(8), 1788–1800 (2019)
4. Bradski, G.: The OpenCV Library. Dr. Dobb’s Journal of Software Tools (2000)
5. Dawant, B.M., Li, R., Lennon, B., Li, S.: Semi-automatic segmentation of the liver and its evaluation on the miccai 2007 grand challenge data set. *3D Segmentation in The Clinic: A Grand Challenge* pp. 215–221 (2007)

6. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian Conference on Image Analysis. pp. 363–370. Springer (2003)
7. Foruzan, A.H., Chen, Y.W.: Improved segmentation of low-contrast lesions using sigmoid edge model. *International Journal of Computer Assisted Radiology and Surgery* **11**(7), 1267–1283 (2016)
8. Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A.: Mrf-based deformable registration and ventilation estimation of lung ct. *IEEE Transactions on Medical Imaging* **32**(7), 1239–1248 (2013)
9. Heller, N., Sathianathen, N., Kalapara, A., et al.: C4kc kits challenge kidney tumor segmentation dataset (2019). <https://doi.org/10.7937/TCIA.2019.IX49E8NX>, <https://wiki.cancerimagingarchive.net/x/UwakAw>
10. Hermann, S., Werner, R.: High accuracy optical flow for 3d medical image registration using the census cost function. In: Pacific-Rim Symposium on Image and Video Technology. pp. 23–35. Springer (2013)
11. Isensee, F., Petersen, J., Klein, A., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486* (2018)
12. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis* **69**, 101950 (2021)
13. Kavur, A.E., Gezer, N.S., Barış, M., et al.: CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation (Jan 2020), <https://arxiv.org/abs/2001.06535>
14. Keeling, S.L., Ring, W.: Medical image registration and interpolation by optical flow with maximal rigidity. *Journal of Mathematical Imaging and Vision* **23**(1), 47–65 (2005)
15. Lai, Z., Lu, E., Xie, W.: Mast: A memory-augmented self-supervised tracker. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6479–6488 (2020)
16. Lai, Z., Xie, W.: Self-supervised learning for video correspondence flow. In: British Machine Vision Conference (2019)
17. Li, C., Wang, X., Eberl, S., Fulham, M., Yin, Y., Chen, J., Feng, D.D.: A likelihood and local constraint level set model for liver tumor segmentation from ct volumes. *IEEE Transactions on Biomedical Engineering* **60**(10), 2967–2977 (2013)
18. Mocanu, S., Moody, A.R., Khademi, A.: Flowreg: Fast deformable unsupervised medical image registration using optical flow. *arXiv preprint arXiv:2101.09639* (2021)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
20. Roth, H., Farag, A., Turkbey, E.B., Lu, L., Liu, J., Summers, R.M.: Data from pancreas-ct (2016). <https://doi.org/10.7937/K9/TCIA.2016.TNB1KQBU>, <https://wiki.cancerimagingarchive.net/x/eIlXAQ>
21. Roth, H., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: A new 2.5 d representation for lymph node detection in ct (2015). <https://doi.org/10.7937/K9/TCIA.2015.AQIHCNM>, <https://wiki.cancerimagingarchive.net/x/OgAtAQ>
22. Simpson, A.L., Antonelli, M., Bakas, S., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019)

23. Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., Osswald, A., Bouhadjar, M., Marescaux, J.: 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. IRCAD, Strasbourg, France, Tech. Rep (2010)
24. Tran, S.T., Cheng, C.H., Liu, D.G.: A multiple layer u-net, un-net, for liver and liver tumor segmentation in ct. *IEEE Access* (2020)
25. Van Ginneken, B., Heimann, T., Styner, M.: 3d segmentation in the clinic: A grand challenge. In: *MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge*. vol. 1, pp. 7–15 (2007)
26. Wang, G., Zuluaga, M.A., Pratt, R., Aertsen, M., David, A.L., Deprest, J., Vercauteren, T., Ourselin, S.: Slic-seg: slice-by-slice segmentation propagation of the placenta in fetal mri using one-plane scribbles and online learning. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 29–37. Springer (2015)
27. Zheng, Z., Zhang, X., Xu, H., Liang, W., Zheng, S., Shi, Y.: A unified level set framework combining hybrid algorithms for liver and liver tumor segmentation in ct images. *BioMed Research International* **2018** (2018)

## 6 Supplementary Materials



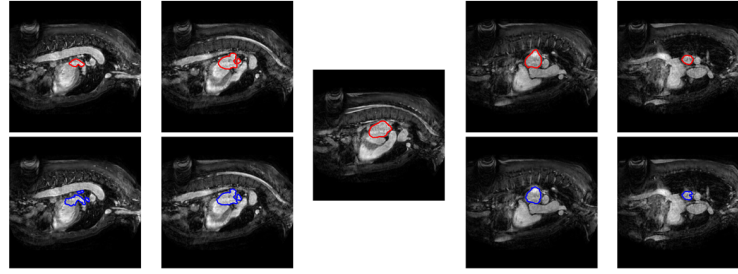
**Fig. 2.** Computation of each iteration of Sli2Vol during *inference*.  $\{S_i, S_{i+1}\}$ , sampled from  $V$  are fed into the trained correspondence flow network to obtain the affinity matrix to propagate  $M_i$  to  $\hat{M}_{i+1}$ .  $\hat{M}_{i+1}$  is then refined by  $p$  and  $n$ , obtained by  $M_i$  and  $S_i$ , to get the final mask,  $M_{i+1}$ .

Modality	Abdominal and Chest CT										Cardiac MRI	
Task	Training			Testing					Training	Testing		
Name	C4KC-KiTS	CT-LN	CT-Pancreas	Sliver07	CHAOS	Decath-Liver	Decath-Spleen	Decath-Pancreas	3DircaDb -01	3DircaDb -02	Kaggle	Decath-Heart
Type	3D Volumes											
SOI	-	-	-	Liver	Liver	Spleen	Liver	Pancreas	Multiple	Multiple	-	Left atrium
Number	310	86	82	20	20	41	131	281	20	2	14370	20
Scanner	Multiple	Multiple	Philips & Siemens MDCT	Multiple	Philips Secura Philips Mx8000 Toshiba AquilionOne	NA	Multiple	NA	NA	NA	NA	Philips 1.5T Achieva
Resolution (xy) (mm)	Varying	Varying	Varying	0.55-0.8	0.7-0.8	Varying	0.5-1.0	Varying	Varying	Varying	Varying	1.25
Resolution (z) (mm)	Varying	Varying	1.5-2.5	1.0-3.0	3.0-3.2	2.5-5.0	0.45-6.0	2.5	Varying	Varying	-	2.7
Details	[9]	[21]	[20]	[25]	[12]	[22]	[22]	[22]	[23]	[23]	[1]	[22]

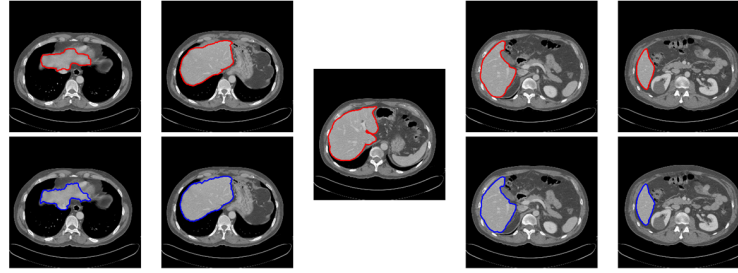
Table 2. Summarization of different datasets used for our experiments.

<i>Approaches</i>	Fully Supervised - Same Domain	Fully Supervised - Different Domain	Fully Supervised - Single Slice	Optical Flow	VoxelMorph2D - UNet	VoxelMorph2D - ResNet18Stride1	Sli2Vol
<i>Backbone Architecture</i>	Varying	- 3D Unet - 16 filters at first level	- 2D Unet - 64 filters at first level	conventional off-the-shelf optical flow algorithm [6]	- 2D Unet - 64 filters at first level	- ResNet18 without max pooling and stride at every layer equals 1 - 16 filters at first level	
<i>Training hyper-parameter</i>	Varying	- Batch size of 1 - Learning rate (lr) of 0.0001 - lr halved when errors plateaued - ADAM optimization	- Batch size of 10	-	- Batch size of 10 - Learning rate (lr) of 0.0001 - lr halved every epoch - ADAM optimization		
<i>Input dim.</i>	Varying	(128, 128, 128)			(256, 256)		
<i>Remarks</i>	Results from [2, 11, 13, 24]	Results from model trained by ourselves	-	Hyperparameters from OpenCV [4]: pyr_scale = 0.5 level = 3 winsize = 7	-	-	Other hyperparameters: d = 8 s = 3 w = 15x15

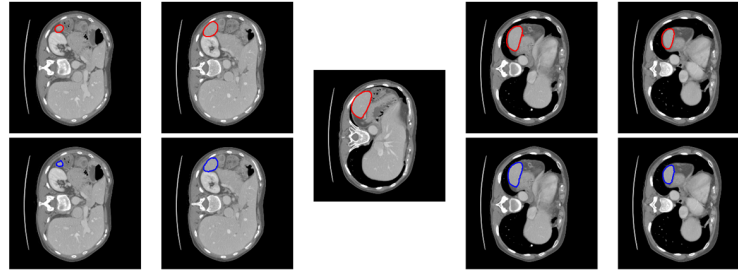
Table 3. Implementation details of Sli2Vol and other baseline approaches.



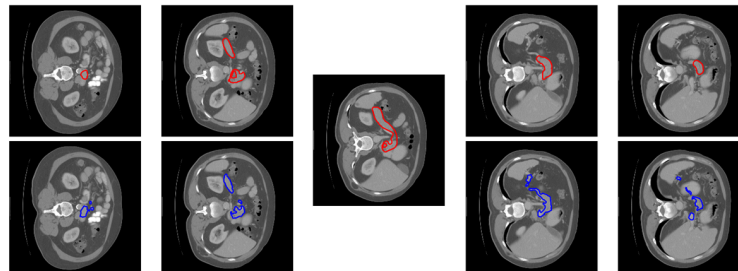
*MRI – Left Atrium*



*CT – Liver*



*CT – Spleen*



*CT – Pancreas*

**Fig. 3.** Examples of segmentation result generated by **Sli2Vol**. The middle slice is the initial annotation. **Red** contours represent groundtruth segmentation while **blue** contours represent segmentation generated by **Sli2Vol**.