Applied Soft Computing 114 (2022) 108074

Contents lists available at ScienceDirect

Applied Soft Computing

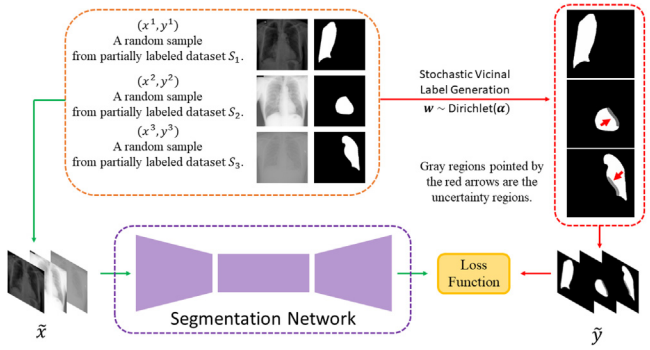
journal homepage: www.elsevier.com/locate/asoc

Towards robust partially supervised multi-structure medical image segmentation on small-scale data

Nanqing Dong [a], Michael Kampffmeyer [b,*], Xiaodan Liang [c], Min Xu [d], Irina Voiculescu [a], Eric Xing [e]

[a] Department of Computer Science, University of Oxford, UK
[b] Machine Learning Group, UiT The Arctic University of Norway, Norway
[c] School of Intelligent Systems Engineering, Sun Yat-sen University, China
[d] Computational Biology Department, Carnegie Mellon University, USA
[e] Machine Learning Department, Carnegie Mellon University, USA

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 May 2021
Received in revised form 24 August 2021
Accepted 24 October 2021
Available online 20 November 2021

Keywords:
Deep learning
Partially supervised learning
Data scarcity
Medical image segmentation

ABSTRACT

The data-driven nature of deep learning (DL) models for semantic segmentation requires a large number of pixel-level annotations. However, large-scale and fully labeled medical datasets are often unavailable for practical tasks. Recently, partially supervised methods have been proposed to utilize images with incomplete labels in the medical domain. To bridge the methodological gaps in partially supervised learning (PSL) under data scarcity, we propose *Vicinal Labels Under Uncertainty* (VLUU), a simple yet efficient framework utilizing the human structure similarity for partially supervised medical image segmentation. Motivated by multi-task learning and vicinal risk minimization, VLUU transforms the partially supervised problem into a fully supervised problem by generating vicinal labels. We systematically evaluate VLUU under the challenges of small-scale data, dataset shift, and class imbalance on two commonly used segmentation datasets for the tasks of chest organ segmentation and optic disc-and-cup segmentation. The experimental results show that VLUU can consistently outperform previous partially supervised models in these settings. Our research suggests a new research direction in label-efficient deep learning with partial supervision.

* Corresponding author.
E-mail addresses: nanqing.dong@cs.ox.ac.uk (N. Dong), michael.c.kampffmeyer@uit.no (M. Kampffmeyer).

https://doi.org/10.1016/j.asoc.2021.108074

1. Introduction

Convolutional Neural Networks (CNNs) have been a game-changer for the task of semantic segmentation [1–3], as they can learn pixel-level mappings from the image space to the label space via end-to-end training. To learn these complex mappings, state-of-the-art CNNs usually leverage a large number of parameters and require the availability of large-scale fully labeled datasets, which are often unavailable for real-life tasks. In the medical domain, where annotations require substantial efforts from clinical experts, obtaining these datasets can be challenging. This has led to an increasing interest in learning from partially labeled data, when fully labeled data is not available. *Partially supervised learning* (PSL) is still an open research question in medical image segmentation [4–8]. From the perspective of multi-task learning (MTL) [9], a semantic segmentation task can be decomposed into multiple sub-tasks corresponding to each semantic class of interest, which provides the theoretical foundations of learning from partial ground truth. Given a medical image segmentation task with multiple classes of interest, it is common to collect and merge several *available*, smaller but relevant datasets into a larger dataset under the challenges of small-scale data, dataset shift, and class imbalance. These smaller datasets were originally labeled for sub-tasks, such that only the objects related to the specific sub-task are annotated, while other objects are merged into the background. In other words, the training images do not have complete annotations for all classes of interest but are *partially labeled*. For example, in the task of abdominal organ segmentation, a pancreas dataset and a liver dataset might be available separately, where only the pancreas and the liver are labeled, respectively.

A key challenge, leading to poor segmentation performance when considering multiple partially labeled datasets, is that the semantic classes of one dataset could be categorized as the background for another dataset that was annotated for a different purpose. Traditional semantic segmentation models [1–3] can therefore not be directly applied and trained end-to-end in a supervised fashion. Further, given the small amount of partially labeled data, deep learning (DL) models are prone to overfitting.

Recent studies in PSL [4–6,10,7,8] all assume that, for each class of interest, enough training examples are accessible. However, considering the data scarcity in most practical medical tasks, usually, only few training examples might be available, making previous approaches impractical.

To bridge the methodological gaps when only small-scale partially labeled data is available, we propose a simple yet efficient framework *Vicinal Labels Under Uncertainty* (VLUU) by exploring the statistical similarity of human structures (e.g. shape, size, location) among different patients. See Fig. 1 for an illustration of such a similarity. The proposed framework is motivated by vicinal risk minimization (VRM) [11], where the fully labeled vicinal examples are generated by linearly combining randomly sampled partial labels with a weight randomly sampled from a Dirichlet distribution. These vicinal examples allow us to transform the partially supervised problem into a fully supervised one. That is to say, we can utilize any existing supervised segmentation networks and loss functions to solve partially supervised problems. The generated vicinal labels contain uncertainty regions where classes of interest could potentially overlap. We utilize these uncertainties in the training process to improve the robustness of DL models.

Recent studies have shown that VRM can consistently improve the performance of CNNs for image classification tasks [12,13]. However, there is a lack of definition of VRM for dense prediction tasks with incomplete labels, e.g. [12,13] cannot be directly applied on partially supervised semantic segmentation tasks. Instead, we revisit VRM, a long-ignored but particularly efficient approach, to tackle this problem. Specifically, by defining a generic vicinity distribution, VLUU learns a mapping from a sequence of images to a vicinal label which is generated by statistically *mixing up* the corresponding partial labels of the input images.

We perform the first systematic study of partially supervised methods under data scarcity challenges, such as small-scale data, *domain shift* or *dataset shift* [17], and class imbalance, on two representative medical image segmentation tasks, namely chest organ segmentation and optic disc-and-cup segmentation. The experiments show that VLUU is more robust than previous partially supervised methods under these settings. The proposed framework has five advantages over previous methods: (1) it is easy to implement without relying on complex loss functions, network architectures, and optimization procedures; (2) it can be trained end-to-end in supervised settings with common segmentation networks and loss functions; (3) it does not require any fully labeled images in the training data; (4) it can efficiently reduce the risk of overfitting for small-scale data; and (5) it can be easily extended to adversarial training.

Our main contributions can be summarized as follows:

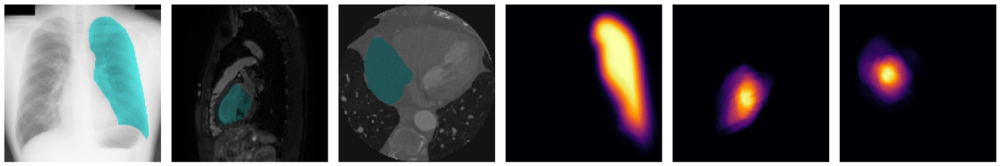
1. We propose a simple yet robust framework for partially supervised medical image segmentation, which is robust when there is only limited partially labeled data.
2. We provide theoretical interpretations for the proposed framework based on vicinal risk minimization and multi-task learning.
3. We systematically evaluate the robustness of partially supervised methods and show that the proposed framework can outperform state-of-the-art partially supervised methods under various data scarcity challenges.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature. Sections 3 and 4 describe the proposed framework and its properties. Section 5 describes the proposed benchmark task and provides experimental results and analysis. Section 6 summarizes this work.

2. Related works

2.1. Semi-supervised learning

In machine learning, semi-supervised learning (SSL) falls between supervised learning (SL), where only fully labeled training data are available, and unsupervised learning (UL), where no labels are available. In semi-supervised learning, the training set consists of both labeled and unlabeled data. The robust state-of-the-art semi-supervised methods include label propagation (LP) [18], graph neural networks [19,20], and cross consistency training [21]. Most semi-supervised methods cannot be applied to PSL problems directly as they are required to minimize a supervised loss, however, among these seminal SSL methods, LP [22] can be applied to tackle partially labeled data directly. With LP, pseudo-labels are generated based on prior information (partially labeled data). Then, the pseudo-labels are fine-tuned iteratively toward convergence [23]. LP is computationally expensive and the quality of the pseudo-labels is highly dependent on the number of training data. [6] has demonstrated that LP is a powerful solution to PSL with fully labeled datasets as prior. As a robust method tested by time, LP is a strong baseline in this work.



(a) (b) (c) (d) (e) (f)

Fig. 1. Annotated examples of different type of medical images (first row): (a) a posteroanterior X-ray image with the ground truth annotation of the left lung; (b) a sagittal MRI image with the ground truth annotation of the left ventricle; (c) an axial CT image with the ground truth annotation of the right atrium. The label distributions (normalized density heatmap) of the corresponding organs in public datasets (second row): (d) the left lungs in the JSRT dataset [14]; (e) the left ventricles in the MRI-WHS dataset [15]; (f) the right atriums in the CT-WHS dataset [16].

2.2. Partially supervised learning

Closely related to SSL, partially supervised learning (PSL), or the partial labels problem, describes the situation where each example has an incomplete label (e.g. only one semantic class is annotated out of a few classes of interest). Concretely, given a collection of multiple small partially labeled datasets, each dataset may only contain annotations for a *proper subset* of classes of interest and these subsets are disjoint. In such a case, the images in the collection are partially labeled. A more rigorous formulation of the problem is presented in Section 3.2.

PSL is a topic of active research as the perfect fully labeled training datasets tend to be only available for specific research tasks. In recent studies, several methods have been proposed to address semantic segmentation with partial labels from different aspects. [24] treats a grid of image patches as nodes and uses conditional random fields to propagate information. However, as a result, the predicted segmentation masks will be unnatural due to the patch-wise prediction. In DL, a common approach is to treat the missing labels as the background. This approach can be viewed as a naive form of *noisy labels* [25] and only works when the pixels of missing classes take up a much smaller portion of the images, compared with the pixels of the background. For benchmark datasets in computer vision such as PASCAL VOC [26] and MS COCO [27], there are only a few classes present in each image or the objects can be very small. Thus, merging unlabeled pixels into the background might be an efficient solution for these datasets. In contrast, for common medical datasets, multiple classes can be present in each image and the objects of interest (e.g. organs) may take up the majority of the pixels. Another common approach in DL is to ignore the cross entropy of the missing classes during backpropagation [4,5]. The limitation of this approach is that abandoning the pixel information of missing classes means that the learners (CNNs) will receive much less supervision during the learning process, both from the image space and the label space. A direct result is that the learner cannot discriminate the classes of interest against the background. Recently, PaNN [6] proposes a complex Expectation–Maximization (EM) algorithm with a primal–dual optimization procedure. However, PaNN requires the availability of fully labeled images as prior, which is often unavailable. To address general semantic segmentation [26,28,10] proposes to use a complex encoder–decoder architecture to condition the partial information within the CNN, which requires a large dataset to comply with the large number of parameters. PIPO-FAN [7] proposes a complex pyramid feature fusion mechanism and a target adaptive loss (TAL). Unlike the other methods, PIPO-FAN has a demanding requirement in the training process, i.e. the examples with the same partial labels must be trained together. It is worth mentioning that TAL also treats the missing labels as the background. Recently, a state-of-the-art work [8] tackles PSL by proposing a marginal loss and an exclusion loss, which are designed for partially supervised medical image segmentation. From the perspective of DL, [8] tries to address PSL at the last step of feed-forward propagation, while this work addresses PSL at the data preparation step, which is before the feed-forward propagation process. To sum up, all of these methods are only applicable when substantial partially labeled images or fully labeled images are available. In addition, previous studies do not consider the practical situations such as dataset shift and class imbalance. A detailed empirical analysis is provided in Section 5.1.

2.3. Multi-task learning

By leveraging task-specific information, multi-task learning (MTL) [9] can improve the model generalization when the tasks of interest are somewhat related. In the era of DL, we aim to use a neural network (NN) to map the input to the output, given a task. In contrast to single-task learning, where each task is handled by an independent NN, MTL can reduce the memory footprint, increase overall inference speed, and improve the model performance. When the associated tasks contain complementary information, MTL can regularize each single task. For dense prediction tasks, a good example is semantic segmentation, where we always assume that the classes of interest are mutually exclusive. Depending on the data modality, task affinity [29] between sub-tasks and task fusion strategy, there are various types of MTL. We depict several common MTL workflows that are related to our work in Fig. 2. Semantic segmentation falls into the category Fig. 2(d). As pointed out by [30], pixel-level tasks in visual understanding often have similar characteristics, which can be potentially used to boost the performance by MTL. We argue that PSL problems can be reformulated as MTL problems by utilizing human structure similarity.

3. Method

3.1. Preliminaries

In SL, given a training dataset $S = \{X, Y\}$ with images $X = \{x_i\}_{i=1}^n$ and ground truth labels $Y = \{y_i\}_{i=1}^n$, the empirical risk is defined as

$$\mathcal{R}(h) = \frac{1}{n}\sum_{i}^{n} L(h(x_i), y_i), \tag{1}$$

where $L(\cdot,\cdot)$ is the loss function and $h \in \mathcal{H}$ is the hypothesis. In this work, we assume that L and h are universal as they can be any loss function and model in a standard supervised setting. For example, for a popular choice of semantic segmentation, L could be the cross entropy and h could be a CNN. The minimization of the empirical risk $\mathcal{R}(h)$ is also known as Empirical Risk Minimization (ERM) in statistical learning literature [31].

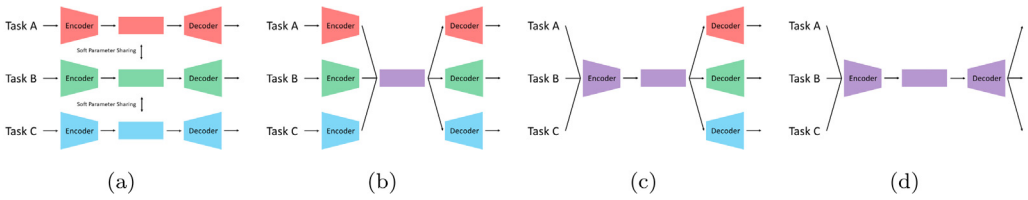


Fig. 2. Common MTL workflows for dense prediction tasks. The data modalities of the input are different: (a) The different tasks have separate networks, which are linked through *soft parameter sharing*. Note, without soft parameter sharing, (a) depicts the standard multiple single-task learning. (b) The different tasks have independent encoders and decoders but share the same network backbone (in purple), which is also known as *hard parameter sharing*. The data modalities of the input are identical: (c) Each task has independent output, which requires an independent decoder. (d) The tasks can share the same decoder.

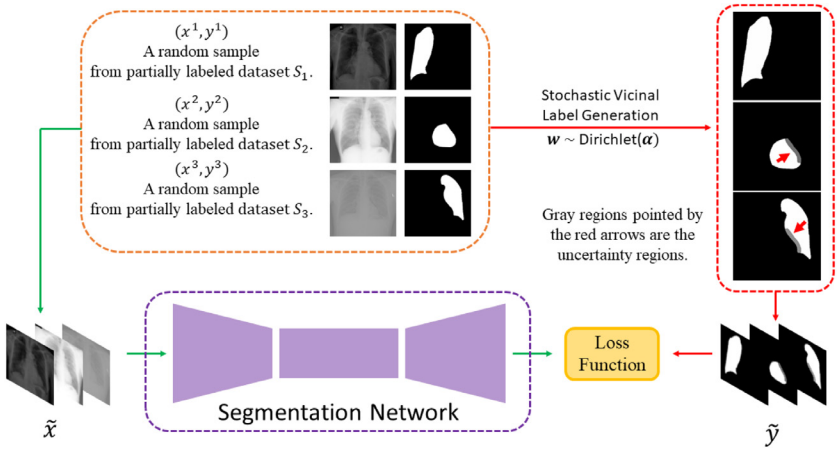


Fig. 3. Illustration of the standard training pipeline. Here, we use the chest organ segmentation task as an example. Assume there are three classes of interest, which are left lung, heart, and right lung. And there are three corresponding partially labeled sub-datasets, denoted as S_1, S_2 and S_3. $\{(x^1, y^1), (x^2, y^2), (x^3, y^3)\}$ are randomly sampled from S_1, S_2 and S_3, respectively. The vicinal example pair $(\tilde{x}, \tilde{y})$ is generated by Eqs. (2) and (3) with $K = 3$. The segmentation network could be any standard segmentation network such as FCN [1] or U-Net [2]. For simplicity, the background mask is not shown in the figure and we use grayscale images to visualize the vicinal labels.

3.2. Problem formulation

Assume there are $K > 1$ mutually exclusive semantic classes of interest present in the same image, i.e. there is no hierarchical relationship between classes and all classes are present. In this work, we focus on the challenging situation that each image is annotated for only one semantic class. For partially labeled images, we can always split S into K sub-datasets where each sub-dataset contains label information of only one class. Here, the K datasets are mutually exclusive in terms of both images and classes. Mathematically, we have $S = \bigcup_j^K S_j$, where $S_j = \{X_j, Y_j\}$ denotes the partially labeled dataset with label information of semantic class j. In S_j, $X_j = \{x_i^j\}_{i=1}^{n_j}$ is the image set of the images with label information of the semantic class j and $Y_j = \{y_i^j\}_{i=1}^{n_j}$ contains the corresponding partial labels. In addition, we define $S_j \subset \mathcal{D}_j$, where $\mathcal{D}_j$ denotes the source domain for S_j, and we define $d(\mathcal{D}_{j_1}, \mathcal{D}_{j_2}) \neq 0 \; \forall j_1 \neq j_2$, where $d(\cdot, \cdot)$ measures the distributional discrepancy between two distribution. That is to say, dataset shift exists. As a comparison, previous studies usually fail to validate this assumption when using one fully labeled dataset to simulate the partially labeled datasets.

Note, the problem formulation here describes the most general case as all other cases are trivial extensions. For example, when an image has annotations for more than one semantic class, duplicate image copies could exist in multiple datasets and the above mathematical formulation still holds.

3.3. Vicinal labels under uncertainty

In a fully supervised setting, introducing statistical randomness [11] and using the convex combination of the training data [12,13] are two efficient methods to improve the robustness of DL models. However, as none of these methods can address the missing class information, they have been ignored in multi-class semantic segmentation with partial supervision for a long time. In this work, we integrate and extend these two simple ideas. Instead of designing complex networks [10,7] or loss functions [8], we utilize the partial labels in a multi-task fashion. A naive solution is to decompose the partially supervised multi-class segmentation task into multiple binary segmentation tasks. As both the input and the output share the same characteristics, we want to use a shared encoder and decoder, similar to Fig. 2(d). However, unlike semantic segmentation, where there is only a single image as input and the corresponding label is based on the same image, we now have images and labels from different partially labeled datasets. We propose to *fuse* the tasks based on the human structure similarity.

Let x be a 2D medical image with size $H \times W$, represented by a 2D array, which has been pre-processed via instance normalization and optional spatial alignment. So y is the corresponding partial label with one semantic class annotated, represented by a 3D array $(H \times W \times (K+1))$, where the last dimension corresponds to the semantic classes. For each pixel in x, the corresponding element in y is a $(K+1)$-element one-hot vector for the background and K semantic classes. For simplicity, we use $y[k]$ to denote the

binary label map for class $k \leq K$ ($k = 0$ denotes the background), which is the $(k+1)$th semantic channel of y. Let (x^j, y^j) be a random sample from S_j, and so $\{(x^j, y^j)\}_{j=1}^K$ is a K-element tuple of such samples. We define

$$\tilde{x} = concat(\{x^j\}_{j=1}^K) \tag{2}$$

$$\tilde{y} = \begin{cases} \frac{w_k y^k[k]}{\sum_{j=1}^K w_j y^j[j] + \epsilon} & k > 0 \\ 1 - \sum_{j=1}^K \tilde{y}[j] & k = 0, \end{cases} \tag{3}$$

where *concat* is the *concatenate* operation that concatenate $\{x^j\}_{j=1}^K$ along a new dimension. We have $\boldsymbol{w} = (w_1, \ldots, w_K) \sim$ Dirichlet($\boldsymbol{\alpha}$) with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in (0, \infty)^K$ and $\epsilon > 0$ is a small number to ensure numeric stability, e.g. $\epsilon = 10^{-3}$. Without prior information over the true label distributions, we setup $\boldsymbol{\alpha}$ as a constant vector, i.e. $\alpha_k = \alpha \ \forall 1 \leq k \leq K$. Given $(\tilde{x}, \tilde{y})$, we transform a partially supervised problem into a fully supervised one and we can utilize any existing supervised segmentation network and loss function. See Fig. 3 for the illustration of the training pipeline. In each class channel of the vicinal label, the continuous probabilities are transformed into grayscale pixels for visualization. There are two origins of uncertainty for generating the vicinal labels when there is an overlap between partial labels. First, the sampling of input images is stochastic. Second, $\boldsymbol{w}$ is randomly sampled from a Dirichlet distribution (e.g. $\boldsymbol{w} = (0.33, 0.41, 0.26)$ used in Fig. 3). See the upper right corner in Fig. 3 for visual examples intuitively, where y_2 and y_3 have an overlapping region.

3.3.1. Theoretical interpretation

The proposed solution can be interpreted from two aspects, namely vicinal risk minimization (VRM) [11] and MTL respectively. In VRM, a vicinity distribution $\mathcal{V}$ is defined as the probability distribution for the virtual image-label pair (also known as vicinal example) $(\tilde{x}, \tilde{y})$ in the vicinity of (x, y). The vicinal risk is defined as

$$\mathcal{R}_{\mathcal{V}}(h) = \frac{1}{n} \sum_{i}^{n} L(h(\tilde{x}_i), \tilde{y}_i). \tag{4}$$

Eq. (3) factually defines a non-parametric anatomical prior for the label distribution. In state-of-the-art VRM works for image classification [12,13], the vicinal image is usually defined as the convex combination of real images, where the parameters for the convex combination are sampled from statistical distributions. As a comparison, we utilize a CNN (h in Eq. (4)) to learn this parametric convex combination jointly with semantic segmentation. Eq. (2) and the CNN jointly play the role of $\tilde{x}$ in Eq. (4). By combining Eq. (2) and (3), we inexplicitly define a generic $\mathcal{V}$.

On the other hand, given K sub-tasks, we are using a CNN to learn a $K \mapsto K$ task mapping. Eq. (3) is a task-fusion process that fuses different but related task knowledge. We want to maximally share the network architecture from a MTL perspective. To achieve this, the novelty here is that we utilize the human structure similarity to *mix up* the partial labels. Meanwhile, the uncertainty regions in the vicinal labels, caused by the stochastic convex combination of partial labels, can reduce the risk of overfitting and improve the robustness when the training data is small.

3.3.2. Extension to adversarial training

Compared with previous works in PSL [4–6,10,7,8], VLUU can be potentially further improved through adversarial training. Adversarial training was first proposed by [32] and several breakthroughs have been made through adversarial training in medical image segmentation [33–36]. However, adversarial training for semantic segmentation is ill-defined when the ground truth labels are missing [37]. As VLUU can transform the partially supervised problem into a fully supervised one, it is natural to consider incorporating VLUU and adversarial training. Note, having complete labels during training gives VLUU unparalleled advantages in utilizing some well-known properties of adversarial training, which is difficult for most partially supervised methods.

In standard adversarial training, the segmentation network and the discriminator play a zero-sum game. The discriminator is trained to discriminate the prediction masks produced by the segmentation network from the ground truth masks. Meanwhile, the segmentation network is trained to confuse the discriminator by producing realistic prediction masks. Adversarial training benefits from the human structure similarity as it makes the unknown true label distributions easier to be caught by the discriminator than for general objects [38]. In other words, there is smaller instance-wise variation in the size, shape, and location of human organs (or structures), as shown in Fig. 1, than for general objects.

Assume the segmentation network is parameterized by f_θ and the discriminator is parameterized by g_ϕ. Given ϕ fixed, θ is updated by minimizing

$$\mathcal{L}_\theta = \mathcal{L}_{seg}(f_\theta(\tilde{x}), \tilde{y}) - \lambda \log g_\phi(f_\theta(\tilde{x})), \tag{5}$$

where $\mathcal{L}_{seg}$ is the multi-class cross-entropy loss for standard supervised semantic segmentation and λ controls the weight of the adversarial loss. Given θ fixed, ϕ is updated by minimizing

$$\mathcal{L}_\phi = -\log g_\phi(\tilde{y}) - \log(1 - g_\phi(f_\theta(\tilde{x}))). \tag{6}$$

See Fig. 4 for the illustration of adversarial training with the vicinal examples. We denote VLUU with adversarial training as VLUU-ADV.

Further, *continuous* vicinal labels have a built-in advantage in stabilizing adversarial training. They alleviate the problem that there commonly is a clear discrepancy between the *discrete* distribution of the ground truth and the *continuous* distribution of the pixel-wise predictions, which can be easily caught by the discriminator [37] and destabilize training, leading to oscillating parameters [39]. Last but not least, with adversarial training, VLUU can further utilize unlabeled data in addition to the partially labeled data. For the interested readers, the problem formulation and application of adversarial training for SSL can be found in [40].

4. Theoretical analysis

In this section, we will discuss the theoretical advantages and limitations of the proposed framework.

4.1. Enlarged sample space

One of the main challenges for DL is overfitting caused by data scarcity. In this work, there are two aspects of data scarcity: (1) each image has an incomplete label, and (2) each S_i has only a small number of images. For (1), Eqs. (3) and (2) generate fully labeled vicinal example pairs, thus traditional end-to-end training techniques in supervised learning can finally be applied.

For (2), with limited training data, state-of-the-art CNN architectures can easily overfit to the training data. Let us first isolate the randomness effect caused by the Dirichlet distribution by setting $w_i = \frac{1}{K}$. The proposed framework enlarges the sample space from $\sum_i n_i$ partially labeled examples to $\prod_i n_i$ fully labeled example pairs. In fact, given $\{(x_i, y_i)\}_{i=1}^K$, Dirichlet($\boldsymbol{\alpha}$) can theoretically generate an infinite number of $\tilde{y}$ determined by $\boldsymbol{w}$. We efficiently mitigate the overfitting problem by enlarging the sample space of $\tilde{S}$.

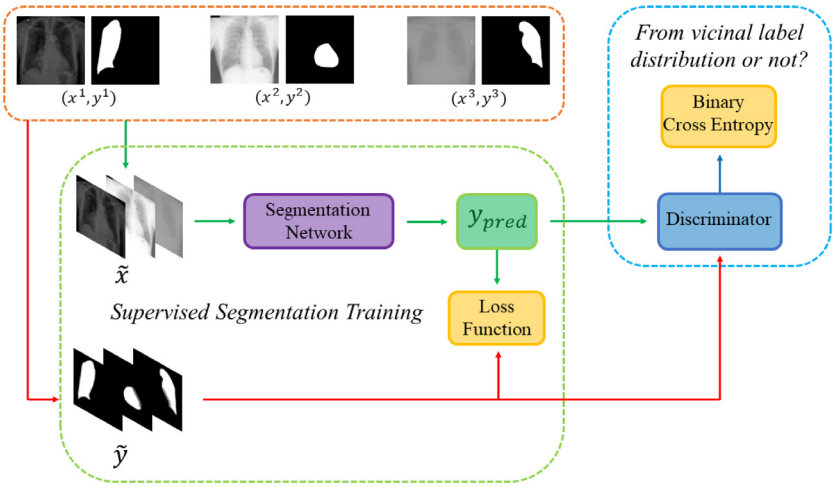


Fig. 4. Illustration of adversarial training pipeline. $(\tilde{x}, \tilde{y})$ is generated by Eqs. (2) and (3). Same as Fig. 3, the background mask is not shown in the figure and we use grayscale images to visualize the vicinal labels. The segmentation network is trained with $(\tilde{x}, \tilde{y})$ in a supervised fashion. y_{pred} is the output of the segmentation network, which is the concatenation of $(K+1)$ probability maps. An auxiliary discriminator is trained to identify whether y_{pred} is sampled from the vicinal distribution, i.e. discriminate y_{pred} against $\tilde{y}$. The segmentation network and the discriminator are trained alternatively. See Eqs. (5) and (6) for details.

4.2. Label smoothing

In semantic segmentation tasks, labels usually follow a discrete distribution, while Eq. (3) defines a continuous distribution. Even though the application of continuous label distributions is rare in semantic segmentation, they have led to recent breakthroughs in image classification [41,12]. We expect Eq. (3) can improve the robustness of the model as suggested by recent theoretical analysis of continuous label distributions [42].

4.3. Computational cost

The training process of the proposed framework is almost identical to the training process for a fully supervised task, i.e. given a segmentation network, there is no additional optimization cost such as multi-stage training [6]. Similarly, the proposed method utilizes the same memory footprint in terms of CNN weights. As a comparison, a semi-supervised method such as label propagation and knowledge transfer will require the training of multiple segmentation networks to generate pseudo-labels. For the proposed method, the major overheads arising from the data generation process are the random sampling and the element-wise operations on low-dimensional arrays, which are negligible compared to the backpropagation cost. Eqs. (3) and (2) can be easily implemented by any scientific computing frameworks supporting broadcasting, such as NumPy, PyTorch, and TensorFlow.

4.4. Limitations

The main purpose of the proposed framework is to train DL-based segmentation models with partial labels in an efficient way. As discussed in Section 3.2, the design of Eqs. (3) and (2) makes a strong assumption that all classes of interest are present in each image and there is no hierarchical relationship between the semantic classes, i.e. the classes of interest are mutually exclusive, e.g. organs in the same body part or sub-structures under the same structure. The situation where the semantic classes have a hierarchical structure, e.g. liver and liver tumor, is beyond the scope of discussion.

Note, the proposed framework is designed for DL tasks on only a few images without complete annotations. When fully labeled data is available, state-of-the-art supervised and semi-supervised methods have obvious advantages over the proposed framework. However, the proposed framework fills the gap when supervised and semi-supervised methods fail.

5. Empirical analysis

The purposes of the experimental design are threefold. First, there is no known empirical study of PSL with limited data. We want to investigate the impact of limited partial labels on DL. Second, we want to systematically evaluate the robustness of the representative partially supervised methods in a controlled environment. Third, we want to demonstrate the effectiveness of VLUU in situations where only a few partially labeled images are available. Thus, the choice of the network backbone or loss function is *independent* of the proposed learning framework. In addition, the simulated experiments are solely to demonstrate the challenges of data scarcity in a controllable environment. We consider two medical image segmentation tasks, chest organ segmentation and optic disc-and-cup segmentation.

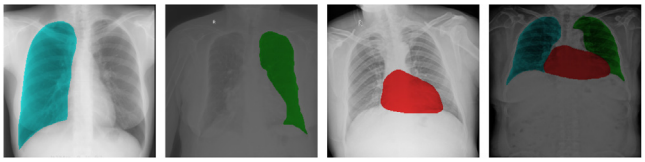
5.1. Chest organ segmentation

The task of chest organ segmentation is a simple benchmark task in medical image segmentation. In this task, we consider three semantic classes, namely *left lung*, *right lung*, and *heart*. We can easily control the environment to get an insight into the impact of the limited partial labels on various representative partially supervised methods and the efficiency of VLUU. Without specification, the experimental comparison is conducted in such a way that different models use the same network backbone, loss function, training strategy, and the set of hyperparameters.

5.1.1. Datasets

We use two public datasets to simulate the realistic situations that each partially labeled dataset is annotated for a different semantic class and is collected from an independent source. Unlike [8], which only consider partially labeled datasets, we use two fully labeled datasets to better understand the influence of partial labels.

The **JSRT** dataset, released by the Japanese Society of Radiological Technology (JSRT), is a benchmark dataset for chest organ segmentation [14]. JSRT contains 247 grayscale CXRs with pixel-wise annotations of lungs and hearts. Each CXR has a size of 2048×2048.



(a) L (b) R (c) H (d) T

Fig. 5. Visual comparison of three partially labeled training sets and one fully labeled test set with corresponding annotations.

The **Wingspan** dataset was collected by Wingspan Technology for the study of transfer learning and unsupervised domain adaptation in chest organ segmentation [40]. Wingspan contains 221 grayscale CXRs with pixel-wise annotations of lungs and hearts. The CXRs were collected from 6 hospitals with different imaging protocols. Wingspan expresses a large variety in the data modalities including brightness, contrast, position, and size.

We use three partially labeled datasets as the training set and one fully labeled as the test set, where the four datasets are collected from four different sources. We choose this setup to simulate the practical scenarios where dataset shift exists, which is a challenging situation for DL models. We use the JSRT dataset as the left lung dataset, denoted as L. We use a subset of the Wingspan dataset containing 18 CXRs as the right lung dataset, denoted as R. We use another subset of the Wingspan dataset containing 18 CXRs as the right lung dataset, denoted as H. We use the rest of the Wingspan dataset as the fully labeled test set, which contains 185 CXRs, and denote it as T. The visual comparison of the data modalities of the four sets can be viewed in Fig. 5. Note, all four sets are collected from 4 different sources (hospitals with different imaging protocols).

5.1.2. Baseline models

For a fair comparison, we use the same segmentation network for all methods, which is a FCN [1] with a ResNet18 [43] backbone. Considering the data scarcity situation, we choose ResNet-FCN as it can both achieve promising results on chest organ segmentation tasks [40] and avoid overfitting. We choose the following representative approaches as the baseline models.

Fully Supervised Learning Approach To illustrate the effect of limited partial labels on DL models, we consider two practical approaches in computer vision that are commonly used during large-scale training. As discussed in Section 2.2, two methods can be used to train end-to-end methods in a supervised fashion. The first one is to categorize the uncertain (missing) classes as the background in the training, which can be considered as a naive solution with *noisy labels*. We denote the first baseline as MBG because we mix uncertain pixels with the background pixels. The second baseline is to ignore the cross-entropy of the missing classes during the backpropagation. This method is motivated by the nature of multi-task learning for neural networks. We denote this method as IMBP. It is worth mentioning that MBG and IMBP further motivate many recently proposed methods for PSL [4,5,7].

Semi-Supervised Learning Approach We adopt a strong SSL baseline, label propagation (LP) [18], to solve PSL problem. LP is not an end-to-end method as there are multiple training stages. It first generates *noisy* pseudo-labels for the unlabeled classes based on the partially labeled data. Then the pseudo-labels and ground truth labels are trained together to make the final prediction. However, the quality of the noisy pseudo-labels is highly dependent on the quality of the partially labeled examples and noisy labels might harm the later fine-tuning stage. In this work, we use K independent binary segmentation networks to generate the initial pseudo-labels.

Multi-Task Learning Approach A classical way to address MTL problems is to fuse knowledge extracted from each individual sub-task [44], which is also known as knowledge transfer (KT) in the *transfer learning* literature. We train K binary segmentation networks with a shared ResNet feature extractor but independent deconvolutional layers. We alternatively optimize K binary segmentation networks on the corresponding K partially labeled datasets. The final prediction masks is generated by fusing K binary prediction masks. For each pixel, if all classes of interest have probabilities less than the threshold 0.5, we treat it as the background. Otherwise, the pixel is categorized as the class with the highest probability.

Partially Supervised Learning Approach We consider the state-of-the-art partially supervised method *exclusion loss* (EL) [8], which is designed for the same problem formulation in Section 3.2. EL has shown superior performance over recent partially supervised methods, such as PaNN [6] and PIPO-FAN [7], in all aspects. Unlike EL, recent partially supervised methods rely on either large training data [4,5,10,7] or fully labeled data as a prior [6], which are not applicable for some situations. Similar to our approach, EL can be applied to any existing segmentation networks. So they can be compared with VLUU in a fair setting.

5.1.3. Implementation

The image size is fixed to be 256×256. We pre-process the raw images by instance normalization. Given an image x, we obtain the normalized image $\hat{x}$ by $\hat{x}^{ij} = \frac{x^{ij}-\mu(x)}{\sigma(x)}$, where (i, j) is the position of the pixel in a 256×256 image, and μ and σ are the mean and standard deviation of the pixels of x. In this study, we do not apply other pre-processing techniques as there is no obvious difference in the relative position of objects in each image and the proposed framework is robust against slight misalignment. In practice, when partially labeled datasets are acquired from different imaging protocols, pre-processing techniques such as registration, resizing, and cropping are necessary. There are no fully labeled images in the training set and we consider the setting where each training image only has an annotation of one semantic class, as described in Section 3.2.

All experiments are implemented in PyTorch on an NVIDIA Tesla V100. For a fair comparison, all the networks are initialized with the **same** random seed and trained from scratch. We use a standard multi-class cross-entropy as the loss function for all the experiments. The batch size is 8. The models are trained to converge with an Adam [45] optimizer and a fixed learning rate of 10^{-3}. The performance metric in this study is the mean Intersection-Over-Union (mIOU) between the prediction masks and ground truth masks over the three classes of interest. For VLUU, we set $\alpha = 0.1$.

5.1.4. Comparison under small-scale data

Because the partially labeled datasets are collected from different sources, we will focus on the challenges of data scarcity and class imbalance. As we want to examine how the size of the partially labeled datasets affects the DL models, we only include n examples of each partially labeled dataset for a quantitative comparison. We provide the performance of the segmentation networks trained on the same training data but with complete annotations as an *Oracle* to provide a reference for the performance. The results are shown in Table 1. Supervised methods fail to address the partial labels due to overfitting. As shown in Fig. 6, MBG tends to predict every pixel as the background while IMBP fails to identify the background, which follows the discussion in Section 2.2. LP, KT, and EL mitigate the partial labels problem from different perspectives and achieve much better performance than supervised methods. However, these seminal methods suffer from the limited training data and multi-source

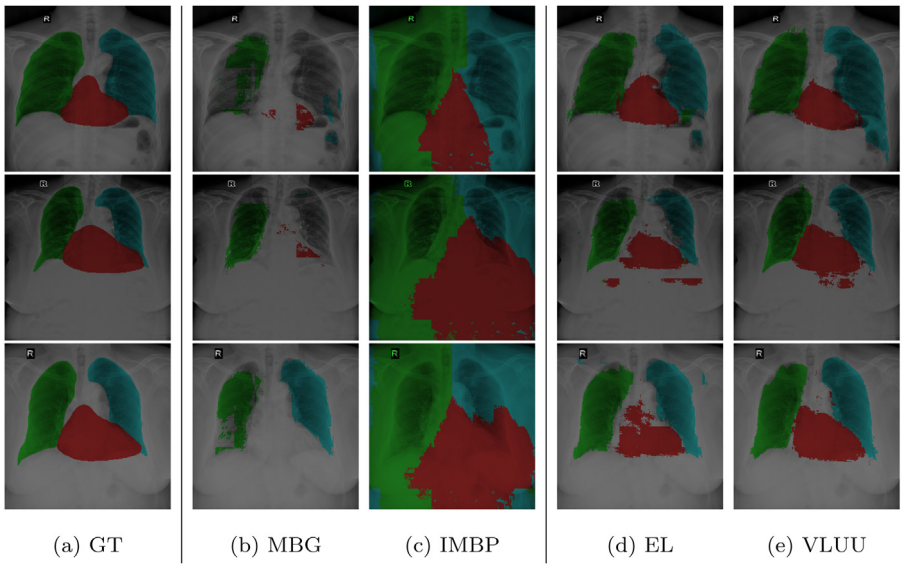


Fig. 6. Qualitative comparison of end-to-end methods on partially supervised chest organ segmentation with $n = 15$. GT denotes the ground truth. The segmentation network is ResNet-FCN. n denotes the number of images in each partially labeled dataset. Traditional training strategies for supervised learning, such as (b) MBG and (c) IMBP, fail for PSL. Compared with (d) EL, (e) VLUU generates more realistic organ masks.

Table 1
Quantitative comparison (mIOU) on partially supervised chest organ segmentation with small-scale data. The segmentation network is ResNet-FCN. n denotes the number of images in each partially labeled dataset.

Method	Type	$n = 5$	$n = 10$	$n = 15$
MBG	SL	0.3187	0.3221	0.2715
IMBP [4]	SL	0.2715	0.3161	0.3218
LP [18]	SSL	0.5821	0.7444	0.7588
KT [44]	MTL	0.6478	0.6686	0.7071
EL [8]	PSL	0.6306	0.6591	0.7506
VLUU	PSL	**0.7063**	**0.7462**	**0.7615**
Oracle	SL	0.7860	0.8395	0.8487

Table 2
Quantitative comparison (mIOU) of methods on chest organ segmentation with class imbalance. The segmentation network is ResNet-FCN. η denotes the ratio of the number of images in the dataset L or R to the number of images in the dataset H.

Method	Type	$\eta = 1$	$\eta = 2$	$\eta = 3$
MBG	SL	0.3187	0.3633	0.3433
IMBP [4]	SL	0.2715	0.3052	0.3029
LP [18]	SSL	0.5821	0.6344	0.6555
KT [44]	MTL	0.6478	0.6511	0.6446
EL [8]	PSL	0.6306	0.7263	0.7347
VLUU	PSL	**0.7063**	**0.7268**	**0.7365**
Oracle	SL	0.7860	0.8208	0.8340

domain shift. Among the baseline methods, LP is the most computationally expensive method as it requires considerably more training time and memory footprint than all other methods. In addition, LP is more sensitive to the size of the training set. In practice, semi-supervised models expect a large set of unlabeled data, which is not aligned with the problem formulation in this work. Compared with semi-supervised methods, MTL methods usually consume a much smaller memory footprint depending on the number of shared layers. By comparing KT and VLUU, we can see that VLUU has more shared neural architectures than KT, which can reduce the memory footprint and substantially improve the model performance. As the state-of-the-art partially supervised method, EL purely relies on using a modified loss function to extract knowledge from the training. When there is not enough training data, EL performs worse than KT and VLUU. In contrast to the baseline methods, VLUU achieves the best performance on small-scale data. Without acquiring any new supervision, VLUU incorporating a coarse anatomical knowledge by uniquely utilizing human structure similarity.

It is worth mentioning that, MBG, IMBP, EL, and VLUU are end-to-end methods, i.e. they do not require any auxiliary NNs or multi-stage training procedures. We provide the qualitative comparison of end-to-end methods in Fig. 6. VLUU tends to output more realistic masks than the STOA method EL in terms of the location and shape.

5.1.5. Comparison under class imbalance

Considering the availability of the medical data and the difficulty of annotating certain organs or structures, we simulate the class imbalance situations in PSL. Here, we use η to control the class imbalance. As the heart is more difficult to annotate than the two lungs [33], we set the partially labeled dataset for the heart (H) to have $n = 5$ and the partially labeled datasets for the two lungs (L and R) to both have ηn examples. The results are shown in Table 2. Compared with Table 1, the class imbalance does have a severe negative impact on the baseline methods MBG, IMBP, and KT, as more training data could even decrease the performance. While LP, EL, and VLUU could benefit from more training data, LP achieves much lower performance than EL and VLUU. VLUU can generally achieve comparable performance with EL while outperforming EL by a large margin with small n. Compared with the baseline methods, VLUU mitigates the class imbalance by utilizing human structure similarity to generate a balanced vicinal label distribution.

5.1.6. Ablation studies

Impact of Network Complexity Under the data scarcity challenge, the complexity of the segmentation network will usually play an important role. The network complexity is determined by the number of parameters and the network architecture. For supervised tasks, U-Net should outperform ResNet-FCN because U-Net has more parameters than ResNet-FCN[1] and a better network architecture design for medical image segmentation tasks [2].

[1] U-Net has 38.8M parameters and FCN with a ResNet18 backbone has 13.3M parameters.

Table 3
The impact of network complexity on VLUU with ResNet-FCN as the segmentation network. n denotes the number of images in each partially labeled dataset.

Network	$n = 5$	$n = 10$	$n = 15$
FCN [1]	**0.7063**	**0.7462**	0.7615
U-Net [2]	0.5411	0.7261	**0.7799**

Table 4
Robustness of VLUU under different random initiations. The performance (mean mIOU ± standard deviation) of VLUU is more stable than the performance of EL.

Method	$n = 5$	$n = 10$	$n = 15$
EL [8]	0.6313 ± 0.1997	0.2587 ± 0.3966	0.7506 ± 0.1576
VLUU	0.7058 ± 0.1226	0.7399 ± 0.1200	0.7609 ± 0.1036

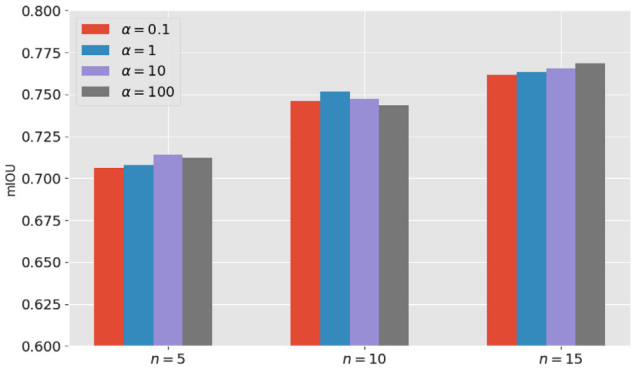


Fig. 7. Sensitivity of α to n. Overall, VLUU is robust for various α.

Clearly, there is a trade-off in the network selection between the network complexity and network performance when the partially labeled datasets are small. Here, we evaluate VLUU with both FCN and U-Net, and results are shown in Table 3. We hypothesize that complex networks have a negative impact on VLUU when there is only limited data. Given a small amount of training data, complex networks could have both performance gain due to more parameters and delicate architectures, and performance drop due to overfitting, depending on the amount of training data.

Sensitivity to $\boldsymbol{\alpha}$ The performance of a ResNet-FCN trained by VLUU with different α is shown in Fig. 7. Overall, VLUU is not sensitive to α as there are only small differences between the performance for different α values. Note, Dirichlet($\boldsymbol{\alpha}$) is asymptotically close to a uniform distribution when $\alpha \to \infty$, i.e. $w_i = \frac{1}{K}$. In addition, there is a trade-off in selecting the optimal α. Small α indicates a larger variation in the label distribution, which means larger uncertainty. So, for tasks such as chest organ segmentation where the organs have relatively fixed locations and similar shapes, a large α might help. However, a small α should be more robust as it introduces more uncertainty when K is large. In this work, we use $\alpha = 0.1$ for consistency.

Effect of Random Initiation To examine the sensitivity of the proposed framework to the effect of random initiation, we repeat the experiments in Table 1 for EL and VLUU for 5 times each. This time, the backbone network is randomly initiated at each time. Unlike the results in Table 1, which are the highest mIOU, we report the mean and standard deviation of mIOUs in Table 4. Compared with the loss-based partially supervised method EL, the label-based partially supervised method VLUU is more robust with smaller standard deviation.

Adversarial Training For VLUU-ADV, we use a standard ResNet binary classifier as the discriminator as we use a ResNet-FCN as the segmentation network. In fact, the choice of the discriminator is a research question in its own right [38]. [46] shows that having the same backbones for the segmentation network and the discriminator can increase the stability of adversarial training. We follow the training scheme in Section 3.3.2, where the adversarial loss [37] in Eq. (5) is weighted by $\lambda = 0.001$. We report the results of VLUU and VLUU-ADV in Table 5, where VLUU-ADV shows slightly better results than VLUU. We conclude that ADV can be used as an add-on module for VLUU with *appropriate* α and *delicate* design of the network architecture for the discriminator.

Table 5
Quantitative comparison (mIOU) between VLUU and VLUU-ADV with ResNet-FCN as the segmentation network. n denote the number of images in each partially labeled dataset.

Method	$n = 5$	$n = 10$	$n = 15$
VLUU	0.7063	**0.7462**	0.7615
VLUU-ADV	**0.7171**	0.7412	**0.7630**

5.2. Optic disc-and-cup segmentation

In addition to chest organ segmentation, another task where all classes of interests are present in each image is the optic disc-and-cup segmentation. As an important step of early screening of glaucoma, optic disc-and-cup segmentation on the fundus images localizes the optic disc-and-cup for the analysis of the optical nerve head [47]. An increase in the optic cup-to-disc ratio could be an indicator of the presence of glaucoma [48]. The annotation of the optic disc is more difficult than that of the optic cup. In addition, the optic disc and optic cup have a unique geometric property that the optic cup is always enclosed by the optic disc. That is to say, if we want to annotate the optic disc, we have to annotate the optic cup first. Although this is not the standard problem formulation, VLUU can be applied to this situation directly as discussed in Section 3.2.

5.2.1. Datasets

We use the REFUGE dataset[2] to simulate the experiments for optic disc-and-cup segmentation. As there are two classes of interest, there should be at least two partially labeled datasets. However, as explained before, it is less practical to have a partially labeled dataset for optic disc. Instead, we have one larger partially labeled dataset for optic cup (denoted as P) and one smaller fully labeled dataset (denoted as F) as the training set. This motivation behind is twofold. First, the annotation of optic cup requires less human effort and is much cheaper to acquire than the annotation of optic disc. Second, we want to introduce the class imbalance. As REFUGE is collected from multiple sources, we create two sub-datasets from two sources to simulate the dataset shift in the training set. We use the validation set of REFUGE as the test set (denoted as T), which contains 400 fundus images.

As REFUGE is collected from multiple sources, the fundus images have various image size. The images are pre-processed by registration, cropping, and resizing to have a fixed resolution of 256×256. So the pre-processed images contain the whole region of the optical nerve head. See Fig. 8 for examples of the training set and the test set.

5.2.2. Implementation

Based on the results in the previous section, we only compare EL and VLUU, as EL and VLUU consistently outperform other methods. In addition, we use a new baseline PaNN [6]. PaNN requires that there is a small fully labeled dataset in the training set to learn the prior, which fits our task setup in Section 5.2.1

[2] https://refuge.grand-challenge.org.

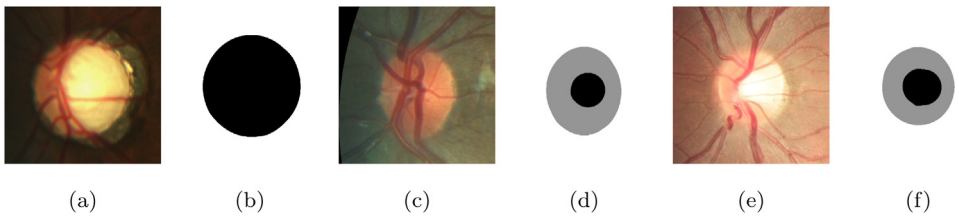


Fig. 8. Visual comparison of the fundus images in the training set and the test set. The training set consists of a partially labeled dataset for optic cup only and a fully labeled dataset for both optic disc and optic cup. (a) A fundus image from the dataset P. (b) The corresponding ground truth mask of (a) with the optic cup annotated as black. (c) A fundus image from the dataset F. (d) The corresponding ground truth mask of (c) with the optic disc annotated as gray and the optic cup annotated as black. (e) A fundus image from the test set T. (f) The corresponding ground truth mask of (e). Note, there are clear dataset shifts among the three datasets.

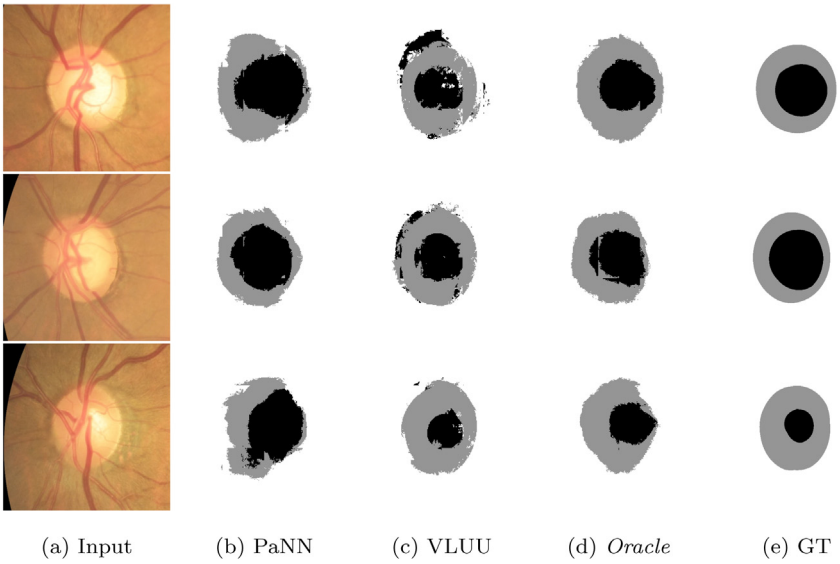


Fig. 9. Qualitative comparison on partially supervised optic disc-and-cup segmentation with $n = 3$. GT denotes the ground truth. The segmentation network is ResNet-FCN. n denotes the number of images with optic disc annotated. A FCN trained with VLUU and partial labels can generate prediction masks which are qualitatively comparable with the masks predicted by a FCN trained with complete labels.

perfectly. Again, for a fair comparison, we use a ResNet-FCN as the network backbone and use the same set of hyperparameters in Section 5.1.3. The performance metric is the mIOU between the unprocessed[3] prediction masks and ground truth masks on optic disc and optic cup.

In contrast to CXRs, the fundus images are color images with RGB channels. To generate a vicinal image, we concatenate two sampled images from the two partially labeled datasets along the RGB channels, i.e. the vicinal images now have 6 ($3K$ where $K = 2$) channels. Eqs. (2) and (3) still hold. In the training of VLUU, we rearrange the training data as two partially labeled datasets. The small fully labeled dataset is split into two sub-datasets containing the same images, where one sub-dataset only contains labels for the optic disc and is treated as the new partially labeled dataset for the optic disc. The other sub-dataset with only labels for the optic cup is added into the partially labeled dataset for the optic cup.

5.2.3. Results

Compared with the experiments in Section 5.1, we use a more extreme setting to test the limit of partially supervised methods. We use only 10 images from P (i.e. 10 images with optic cup annotated) and n images from F (i.e. n images with both optic disc and optic cup annotated). There is a severe class imbalance here, as the ratio of the number of labels for cup to the number of labels for disc is $\frac{10+n}{n}$. The results measured in mIOU between the prediction masks and ground truth masks on optic disc and optic cup are presented in Table 6. With much smaller data size than before, EL fails. Besides, as EL is not designed for fully labeled datasets, the images with complete labels (from F) actually have a negative influence on the training. Meanwhile, PaNN cannot easily learn the image prior based on only a few fully labeled images. VLUU outperforms EL and PaNN by a large margin. Essentially, EL and PaNN do not solve the data scarcity problem, while VLUU can generate new vicinal examples. Moreover, a segmentation network trained with VLUU can even achieve comparable performance with the same network trained with complete labels (i.e. more supervision). Considering the existence of class imbalance and dataset shift, we conclude that VLUU is more robust on small-scale data. The visual comparison between PaNN, VLUU and *Oracle* is shown in Fig. 9. It can be seen that PaNN generates

Table 6
Quantitative comparison (mIOU) of PSL methods on partially supervised optic disc-and-cup segmentation with class imbalance. The segmentation network is ResNet-FCN. n denotes the number of images with optic disc annotated.

Method	Type	$n = 1$	$n = 2$	$n = 3$
EL [8]	PSL	0.1395	0.1596	0.1991
PaNN [6]	PSL/SSL	0.5976	0.5999	0.6299
VLUU	PSL	**0.6452**	**0.7605**	**0.7945**
Oracle	SL	0.6677	0.7045	0.7713

[3] In practice, the prediction masks could be further improved by image processing techniques.

unrealistic shapes for the optic disc and optic cup if not enough fully labeled data is available learn a reasonable image prior. Note, although VLUU can achieve comparable performance with *Oracle* in numerical results, there are artifacts caused by the uncertainty of the vicinal labels, e.g. as shown in Fig. 9, VLUU may generate optic cup predictions outside the optic disc.

6. Conclusion

In this paper, we discuss the robustness issue of partially supervised methods under the challenge of data scarcity. We present VLUU, an easy-to-implement framework, for medical image segmentation tasks with only small partially labeled data. Compared with previous methods, VLUU efficiently utilizes the human structure similarity. The experimental results show that VLUU is more robust than state-of-the-art partially supervised methods under various data scarcity situations. Our research suggests a new research direction in label-efficient DL with partial supervision by tackling the problem from the perspective of VRM.

CRediT authorship contribution statement

Nanqing Dong: Conceptualization, Methodology, Software, Investigation, Writing - original draft, Writing – review & editing, Visualization. **Michael Kampffmeyer:** Writing - original draft, Writing – review & editing, Project administration, Funding acquisition. **Xiaodan Liang:** Validation, Supervision. **Min Xu:** Writing - original draft, Supervision. **Irina Voiculescu:** Writing - original draft, Writing - review & editing, Validation, Supervision. **Eric Xing:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would also like to thank Huawei, Amazon, and Google for providing cloud computing service for this study. This work was partially funded by the Research Council of Norway grants no. 315029, 309439, and 303514.

References

[1] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: CVPR, 2015, pp. 3431–3440.
[2] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, 2015, pp. 234–241.
[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[4] G. González, G.R. Washko, R.S.J. Estépar, Multi-structure segmentation from partially labeled datasets. application to body composition measurements on CT scans, in: Image Analysis for Moving Organ, Breast, and Thoracic Images, Springer, 2018, pp. 215–224.
[5] O. Petit, N. Thome, A. Charnoz, A. Hostettler, L. Soler, Handling missing annotations for semantic segmentation with deep convnets, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 20–28.
[6] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, A.L. Yuille, Prior-aware neural network for partially-supervised multi-organ segmentation, in: ICCV, 2019, pp. 10672–10681.
[7] X. Fang, P. Yan, Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, IEEE Trans. Med. Imaging (2020).
[8] G. Shi, L. Xiao, Y. Chen, S.K. Zhou, Marginal loss and exclusion loss for partially supervised multi-organ segmentation, Med. Image Anal. (2021) 101979.
[9] R. Caruana, Multitask learning, Mach. Learn. 28 (1) (1997) 41–75.
[10] K. Dmitriev, A.E. Kaufman, Learning multi-class segmentations from single-class datasets, in: CVPR, 2019, pp. 9501–9511.
[11] O. Chapelle, J. Weston, L. Bottou, V. Vapnik, Vicinal risk minimization, in: NIPS, 2001, pp. 416–422.
[12] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: ICLR, 2018.
[13] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: ICCV, 2019, pp. 6023–6032.
[14] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, K. Doi, Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules, Am. J. Roentgenol. 174 (1) (2000) 71–74.
[15] X. Zhuang, K.S. Rhode, R.S. Razavi, D.J. Hawkes, S. Ourselin, A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI, IEEE Trans. Med. Imaging 29 (9) (2010) 1612–1625.
[16] X. Zhuang, W. Bai, J. Song, S. Zhan, X. Qian, W. Shi, Y. Lian, D. Rueckert, Multiatlas whole heart segmentation of CT data using conditional entropy for atlas ranking and selection, Med. Phys. 42 (7) (2015) 3822–3833.
[17] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence, Dataset Shift in Machine Learning, The MIT Press, 2009.
[18] A. Iscen, G. Tolias, Y. Avrithis, O. Chum, Label propagation for deep semi-supervised learning, in: CVPR, 2019, pp. 5070–5079.
[19] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: ICLR, 2017.
[20] B. Jiang, Z. Zhang, D. Lin, J. Tang, B. Luo, Semi-supervised learning with graph learning-convolutional networks, in: CVPR, 2019, pp. 11313–11320.
[21] Y. Ouali, C. Hudelot, M. Tami, Semi-supervised semantic segmentation with cross-consistency training, in: CVPR, 2020, pp. 12674–12684.
[22] F. Wang, C. Zhang, Label propagation through linear neighborhoods, IEEE Trans. Knowl. Data Eng. 20 (1) (2007) 55–67.
[23] X. Zhu, Z. Ghahramani, Learning from Labeled and Unlabeled Data with Label Propagation, Tech. Rep. CMU-CALD-02-107, Carnegie Mellon University, 2002.
[24] B. Triggs, J.J. Verbeek, Scene segmentation with CRFS learned from partially labeled images, in: NIPS, 2008, pp. 1553–1560.
[25] N. Natarajan, I.S. Dhillon, P.K. Ravikumar, A. Tewari, Learning with noisy labels, in: NIPS, 2013, pp. 1196–1204.
[26] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The scapes dataset for semantic urban scene understanding, in: CVPR, 2016, pp. 3213–3223.
[29] S. Vandenhende, S. Georgoulis, B. De Brabandere, L. Van Gool, Branched multi-task networks: Deciding what layers to share, in: BMVC, 2020.
[30] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, J. Yang, Pattern-affinitive propagation across depth, surface normal and semantic segmentation, in: CVPR, 2019, pp. 4106–4115.
[31] V. Vapnik, Statistical Learning Theory, Wiley, 1998.
[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014, pp. 2672–2680.
[33] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, E.P. Xing, SCAN: Structure correcting adversarial network for organ segmentation in chest X-rays, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 263–273.
[34] P. Moeskops, M. Veta, M.W. Lafarge, K.A. Eppenhof, J.P. Pluim, Adversarial training and dilated convolutions for brain MRI segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 56–64.
[35] K. Chen, D. Zhu, J. Lu, Y. Luo, An adversarial and densely dilated network for connectomes segmentation, Symmetry 10 (10) (2018) 467.
[36] Z. Han, B. Wei, A. Mercado, S. Leung, S. Li, Spine-GAN: Semantic segmentation of multiple spinal structures, Med. Image Anal. 50 (2018) 23–35.
[37] P. Luc, C. Couprie, S. Chintala, J. Verbeek, Semantic segmentation using adversarial networks, in: NIPS Workshop on Adversarial Training, 2016.
[38] N. Dong, M. Xu, X. Liang, Y. Jiang, W. Dai, E. Xing, Neural architecture search for adversarial medical image segmentation, in: MICCAI, 2019, pp. 828–836.
[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: NIPS, 2016, pp. 2234–2242.

[40] N. Dong, M. Kampffmeyer, X. Liang, Z. Wang, W. Dai, E. Xing, Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio, in: MICCAI, 2018, pp. 544–552.
[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: CVPR, 2016, pp. 2818–2826.
[42] R. Müller, S. Kornblith, G.E. Hinton, When does label smoothing help? in: NIPS, 2019, pp. 4696–4705.
[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[44] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, L. Van Gool, Multi-task learning for dense prediction tasks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[45] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR, 2015.
[46] S. Motiian, Q. Jones, S. Iranmanesh, G. Doretto, Few-shot adversarial domain adaptation, in: NIPS, 2017, pp. 6670–6680.
[47] Z. Wang, N. Dong, S.D. Rosario, M. Xu, P. Xie, E.P. Xing, Ellipse detection of optic disc-and-cup boundary in fundus images, in: ISBI, IEEE, 2019, pp. 601–604.
[48] S. Syc, C. Warner, S. Saidha, S. Farrell, A. Conger, E. Bisker, J. Wilson, T. Frohman, E. Frohman, L. Balcer, et al., Cup to disc ratio by optical coherence tomography is abnormal in multiple sclerosis, J. Neurol. Sci. 302 (1–2) (2011) 19–24.