Self-supervised Learning for Few-shot Medical Image Segmentation

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu and Daniel Rueckert, *Fellow, IEEE*

***Abstract*—Fully-supervised deep learning segmentation models are inflexible when encountering new unseen semantic classes and their fine-tuning often requires significant amounts of annotated data. Few-shot semantic segmentation (FSS) aims to solve this inflexibility by learning to segment an arbitrary unseen semantically meaningful class by referring to only a few labeled examples, without involving fine-tuning. State-of-the-art FSS methods are typically designed for segmenting natural images and rely on abundant annotated data of training classes to learn image representations that generalize well to unseen testing classes. However, such a training mechanism is impractical in annotation-scarce medical imaging scenarios. To address this challenge, in this work, we propose a novel self-supervised FSS framework for medical images, named SSL-ALPNet, in order to bypass the requirement for annotations during training. The proposed method exploits superpixel-based pseudo-labels to provide supervision signals. In addition, we propose a simple yet effective adaptive local prototype pooling module which is plugged into the prototype networks to further boost segmentation accuracy. We demonstrate the general applicability of the proposed approach using three different tasks: organ segmentation of abdominal CT and MRI images respectively, and cardiac segmentation of MRI images. The proposed method yields higher Dice scores than conventional FSS methods which require manual annotations for training in our experiments.**

***Index Terms*—Self-supervised learning; Few-shot segmentation; Representation learning**

I. Introduction

When trained on abundant well-annotated training data, a fully-supervised deep learning segmentation model usually achieves good performance. However, the performance of a fully-supervised model typically deteriorates severely when labeled training data is scarce [1], [2]. Unfortunately, in medical imaging, there is often a lack of large, well-annotated medical image dataset due to the prohibitive cost for manual labeling, making it often impractical to train a data-consuming fully-supervised deep model. Even more problematic is that fully-supervised models are inflexible when faced with arbitrary new classes of potential segmentation targets (anatomical structures or lesions). It is impractical to train a new fully-supervised model for every single new segmentation class, since training from scratch or fine-tuning are time consuming and they require expertise.

A potential solution to the challenges of annotation scarcity and inflexibility to new classes is few-shot learning [3]–[8]. During *testing*, a few-shot learning model extracts discriminative representations of a previously-unseen class from only a few labeled examples (called *support*), and is then able to predict this unseen class on unlabeled data (called *query*), usually without additional fine-tuning. On medical images, most of previous few-shot segmentation methods might be limited by a common drawback: these methods require to be trained on a huge amount of annotated training class examples for learning image representations that are generalizable to unseen classes [9]–[20]. This hunger for large amounts of annotated training data leads to a chicken-and-egg problem due to a scarcity in annotations in medical images.

To circumvent the need for large amount of annotations, we propose to train a few-shot segmentation (FSS) model directly on unlabeled images, via self-supervised learning [21]–[28]. By training on a pretext task like patch in-painting [27] or instance discrimination [29]–[31] on unlabeled images, a self-supervised model learns image representations that are generalizable or customizable to downstream tasks like classification or semantics segmentation. Unfortunately, most of self-supervised learning techniques are only designed for generic transfer learning problem [30]–[32], and thus under-explore the uniqueness of few-shot segmentation problem. Therefore, we propose to tailor self-supervised learning to the FSS problem. Specifically, we propose to exploit prior knowledge of medical images via self-supervision, utilizing specially-designed pseudolabels and training objectives.

Another performance bottleneck for many popular FSS architectures lies in the inability to model differences between local patches within a semantic class. This bottleneck is particularly exaggerated in medical image segmentation in the context of class imbalance. As shown in Fig. 1 (a), the *background* class (i.e. regions outside the purple *right kidney*) is composed of patches with different textures and intensities. Under such a foreground-background segmentation scenario, popular works [11]–[14], [33] might unfortunately lead to ambiguity at the border between foreground and background. This is because they represent each semantic class including the background as a location-agnostic 1-D vector, where the distinctions between different local patches are unreasonably smoothed out. To tackle this problem, we argue that local information needs to be explicitly preserved by an FSS model.

In this paper, we instantiate our solutions to self-supervised few-shot segmentation problem as the proposed SSL-ALPNet: a novel few-shot segmentation framework for medical images. It is a synergy between a *superpixel-based self-supervised*

This work is support by EPSRC Programme Grant EP/P001009/1 (SmartHeart) and Innovate UK Grant 104691 (London Medical Imaging & AI Centre for Value Based Healthcare). (Corresponding author: Cheng Ouyang)

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu and Daniel Rueckert are with the Department of Computing, Imperial College London, London SW7 2RH, U.K. Daniel Rueckert is also with the Institute for AI and Informatics in Medicine, Klinikum rechts der Isar, Technical University of Munich, Germany. C.B. and C.C. contributed equally. (E-mail: c.ouyang@imperial.ac.uk)

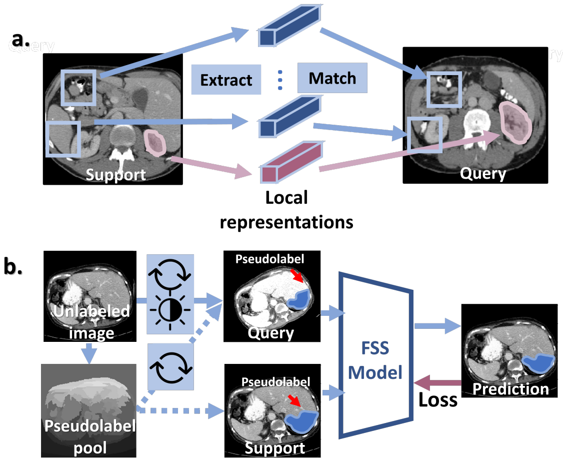

Fig. 1. (a). Intuition behind the proposed segmentation mechanism based on explicitly modeling local image information: First, local representations of each class including the background are extracted from the support images. Each representation only represents a local region, and matches a corresponding local patch in query. Since the background is not homogeneous, performing extract-and-match between local patches yields higher spatial accuracy, compared with methods that use one single 1-D representation to match the entire background. (b). The proposed superpixel-based self-supervision strategy: Superpixel-based pseudolabels are pre-computed from unlabeled images. In each training iteration, an image and one of its randomly sampled pseudolabel serve as the candidates for both support and query. To simulate inter-patient varieties in real-world, random intensity and geometric transformations are applied to the query in order to slightly alter its appearance and shape. The self-supervision task is retrieving a same pseudolabel on the query by referring to the support.

learning (SSL) strategy and an *adaptive local prototype pooling network* (ALPNet). The SSL strategy is designed to bypass the reliance on manually annotated training images, by exploiting unlabeled images and pseudolabels. As shown in Fig. 1 (b), to learn image representations tailored to few-shot segmentation, we use superpixels as pseudolabels. We argue that image representations learned on superpixels are well-generalizable to real semantically meaningful objects. This is because superpixels naturally share a piece-wise smoothness prior with real semantic objects [34], [35]. In addition to leveraging the smoothness prior, we exploit a boundary prior learned from superpixels. This is based on the observation that the boundaries of both superpixels and real semantic objects share similar properties. In terms of network architecture, the ALPNet is designed to improve segmentation accuracy of prototype networks. This is achieved by utilizing the proposed adaptive local prototype pooling module (ALP): a plug-in module added to a prototype network [13]. It explicitly models each local image patch as a distinct representation prototype.

Overall, we summarize our contributions as follows:

- We propose a novel superpixel-based self-supervised learning (SSL) strategy for few-shot medical image segmentation. In our experiments, it achieves the state-of-the-art performance for FSS on medical images without using manual annotations during training, when applied to our network architecture.
- We propose an adaptive local prototype pooling network (ALPNet): a simple network architecture which significantly outperforms the baseline prototype network in few-shot segmentation.
- We demonstrate the robustness and flexibility of our SSL-ALPNet framework on a wide range of medical image segmentation tasks. In particular, we report on a comprehensive evaluation on multiple segmentation classes, imaging modalities and different number of shots in testing. We also explored the challenging weakly-annotated testing scenario. We believe the established evaluation protocol facilitates future works on few-shot medical image segmentation and self-supervised representation learning.

This paper is a substantial extension to our conference paper [36], especially in the following aspects: First, we re-interpret the proposed SSL technique as an early investigation on tailoring representation learning to few-shot segmentation task. In particular, we systematically analyzed the intrinsic properties of medical images under FSS setting, and propose to fully exploit their patch-level image prior knowledge using self-supervised learning (see detailed analysis in Sec. III-C1). Second, we further propose and validate a boundary prior for SSL on medical imaging data, which brings consistent performance gains throughout different datasets compared with [36]. Third, we extend the previous one-shot testing scenario to multiple shots, demonstrating that the proposed framework can efficiently utilize multiple reference examples when available. Fourth, we explore the challenging weakly-annotated few-shot segmentation problem where annotations are extremely scarce and coarse in testing. Powered by strong prior knowledge learned through SSL, the proposed model which takes only bounding-box annotations as reference in testing, still achieve reasonable segmentation accuracy (see Sec. IV-E). Finally, in Sec. II-B we draw connections between the proposed SSL technique and recent contrastive representation learning [29]–[31], [37]–[41], particularly a closely related work [2]. We elaborate their similar practice of using image transformations but different intuitions and derivations.

II. Related Work

A. *Few-shot semantic segmentation*

Most of current few-shot segmentation techniques focus on network architecture design, and can be roughly categorized as methods based on *implicit feature interaction*, or methods based on *explicit representation comparison*. The former category starts from [15], where both the support and the query images are sent to a network, and the support features implicitly guide the network to make predictions on the query. Recent works [1], [16], [42] exploit more sophisticated network components to construct stronger inductive biases that are favorable for information propagation between support and query.

The earliest work based on representation comparison is by [10], where the label of the query is decided by making comparisons with the support. Recent works include [12], [17], [19], [43]. A major stream along this line of research is *prototypical networks* (PN) [3], [11], [13], [33], where representation prototypes of the support are calculated for measuring pixel-wise similarities with the query feature map.

As the similarities measured between prototypes and feature maps of the query in PN can be used to visualize the prediction process, which is highly desirable for medical image applications where network interpretability is beneficial, our framework follows the principle of PN. Specifically, we choose one of the state-of-the-art prototypical alignment network (PANet) [13] as our baseline. Compared with other state-of-the-art methods, PANet is conceptually simple and elegant, with only an off-the-shelf feature extractor network and a proposed alignment regularization term. Using such a generic model could highlight our self-supervised technique as a universal training strategy.

Almost all of the works above assume the availability of abundant annotated training data, and therefore focus solely on network architecture design. In contrast, our work aims to tackle the unsolved training data scarcity problem, and therefore focuses on designing a novel self-supervised learning technique tailored to image segmentation.

In medical imaging applications, FSS has been previously interpreted as using a few annotated samples for training or for fine-tuning [2], [44], [45], [45]–[47], [47]–[52]. These methods are therefore out-of-scope in our discussion as fine-tuning requires expertise in deep learning, thus being inflexible in clinical settings. SE-Net [1] utilizes squeeze-and-excite blocks [53] for implicit feature interaction, and can be applied to unseen classes without fine-tuning. Concurrent with or after our previous work [36], Feyjie *et al.* [54] employ image denosing as an auxiliary task for regularizing few-shot medical image segmentation. Additionally, Sun *et al.* [55] extend SE-Net by introducing attention mechanism and by regularizing intra-class and inter-class distances. Yu *et al.* [56] enforce a strong local constraint to a prototype network.

B. *Self-supervised learning in semantic segmentation*

In semantic segmentation, the majority of the self-supervised techniques aim to obtain a pre-trained model to facilitate further supervised training. These approaches learn image representations through handcrafted pre-text tasks such as image in-painting [27], patch reordering [22], rotation regression [57], motion prediction [58] and so on. Similar methods have also been applied to medical imaging: [2], [59]–[62]. However, almost all of them still require a second-stage fine-tuning after self-supervised pre-training. In contrast, our proposed framework can be directly applied to few-shot segmentation of real semantic labels without fine-tuning. In addition, in conventional self-supervised learning, there is no guarantee that image representations that are learned through a pretext task to be fully transferable to downstream tasks. This is due to the potential gap between two tasks (*e.g.* image-level rotation prediction versus pixel-level segmentation). In our method, features learned from SSL are well transferable to few-shot segmentation, since the two tasks are highly related: they share a unified problem formulation, and pseudo-labels in SSL share similar image properties with real semantic labels.

Although derived from different perspectives, our self-supervision task shares a common practice with contrastive representation learning [29]–[31], [38]–[41]: In both methods, image intensity and geometric transformations are employed to boost invariance of learned representations. In addition, our self-supervision task can be also interpreted as instance discrimination on a dense pixel level (see Sec. III-C5). However, most of contrastive representation learning works stem from the principle of mutual information maximization [63] or feature uniformity [64], and they are designed for generic transfer learning. They leave the unique natures of few-shot segmentation under-explored. Our method is instead tailored to few-shot segmentation task, and innovatively exploits image prior knowledge in representations. In parallel with our conference work [36], Chaitanya *et al.* [2] employ contrastive learning on rectangular patches and image instances for semi-supervised medical image segmentation. Unlike [2], which requires a supervised fine-tuning stage, our method is designed for the tuning-free few-shot segmentation scenario.

C. *Superpixels*

Superpixels are small, compact image patches composed of pixels sharing common intensities or textures [35], [65], [66]. Superpixels are usually generated by unsupervised algorithms like graph-cuts [35], under assumptions like piece-wise smoothness [34]. Concurrent with our previous work [36], [67] employs superpixels as clues for grouping features in semi-supervised few-shot segmentation. However, their approach still heavily relies on abundant annotated training data [67]. In contrast, the training phase of our method is purely free of manual annotations.

III. Method

A few-shot segmentation (FSS) model segments an unseen class by referring to only a few labeled references. Of note, in FSS, segmenting unseen classes do not require additional fine-tuning. In the following section, we first introduce the problem formulation of general few-shot segmentation in Sec. III-A, then we describe the proposed adaptive local prototype pooling network (ALPNet) in Sec. III-B. Finally, we present the novel superpixel-based self-supervised learning strategy (SSL) in Sec. III-C, with a detailed discussion of both intuition and instantiation of the two image priors utilized in SSL.

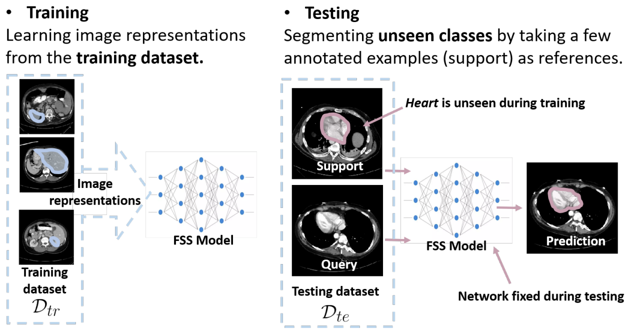
• Training
Learning image representations from the **training dataset.**
Image representations
FSS Model
Training dataset
$\mathcal{D}_{tr}$
• Testing
Segmenting **unseen classes** by taking a few annotated examples (support) as references.
Heart is unseen during training
Support
Query
Testing dataset
$\mathcal{D}_{te}$
FSS Model
Prediction
Network fixed during testing

Fig. 2. Problem formulation of few-shot segmentation (FSS).

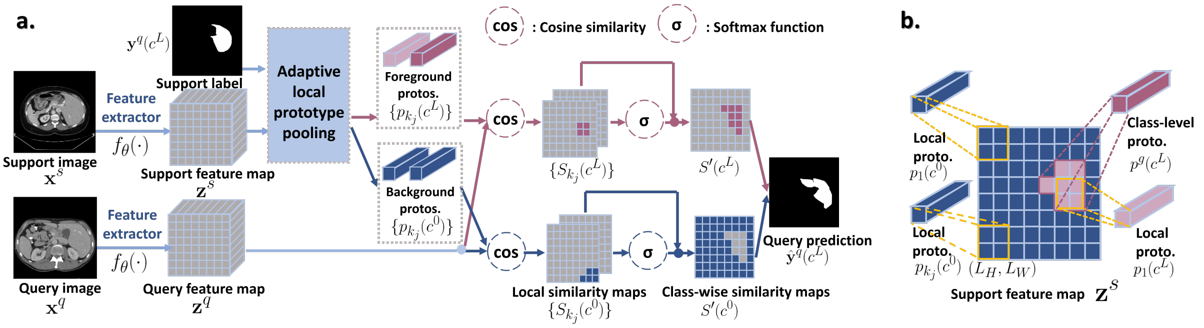

Fig. 3. (a). Workflow of the proposed adaptive local prototype pooling network: The feature extractor $f_\theta(\cdot)$ extracts feature maps of the support: $\mathbf{z}^s$ and that of the query: $\mathbf{z}^q$ from support and query images. Then, the support feature map and the label are sent to the proposed adaptive local prototype pooling module to compute an ensemble of local representation prototypes $p_k(c^j)$'s and a class-level prototype $p^g(c^j)$ for class c^j. In this example $j \in \{0, L\}$, where c^L is the foreground class *Liver*, and c^0 denotes the background class. These representation prototypes serve as references of each class, for measuring spatial similarities with the feature map of the query. These similarity measurements are fused into the final prediction. (b). Mechanism of the adaptive local prototype pooling module: Local prototypes (*e.g.* $p_1(c^L)$) are computed by taking averages of support features within pooling windows (orange boxes) along the spatial dimensions; class-level prototypes (*e.g.* $p^g(c^L)$) are computed under the entire support label (purple region).

A. Problem Formulation

To train an FSS network, a training set $\mathcal{D}_{tr}$ with images of training classes $\mathcal{C}_{tr}$ (e.g., $\mathcal{C}_{tr} = \{liver, spleen, kidney\}$) is required. The training classes are assumed to be known and their labeled samples are assumed to be available, from which the network learns image representations that are generalizable to unseen classes. Specifically, $\mathcal{D}_{tr} = \{(\mathbf{x}, \mathbf{y}(c^{\hat{j}}))\}$ is composed of images $\mathbf{x} \in \mathcal{X}$ and corresponding binary masks $\mathbf{y}(c^{\hat{j}}) \in \mathcal{Y}$ of classes $c^{\hat{j}} \in \mathcal{C}_{tr}$, where $\mathcal{X}$ is the image space, $\mathcal{Y}$ is the label space, and $\hat{j} = 1, 2, 3, ...$ is the class index. After training, the network is fixed and then evaluated on a testing set $\mathcal{D}_{te}$, which is defined in the same way as $\mathcal{D}_{tr}$ but contains images of unseen testing classes $\mathcal{C}_{te}$ (e.g., $\mathcal{C}_{te} = \{heart\}$) where $\mathcal{C}_{tr} \cap \mathcal{C}_{te} = \emptyset$. An illustration of the training and testing phases of (conventional) few-shot segmentation models is shown in Fig. 2. Of note, the background class c^0 does not belong to neither training nor testing classes. We use a different notation $j = 0, 1, 2, 3, ...$ to index all classes including the *background*, *i.e.* $\{c^j\} = \{c^0\} \cup \{c^{\hat{j}}\}$.

Most of the recent FSS models are both trained and tested in *episodes* [9]–[11], [15]. An episode $(\mathcal{Q}, \mathcal{S})$ is sampled from $\mathcal{D}_{tr}$ during training or sampled from $\mathcal{D}_{te}$ during testing. Each episode consists of a query set $\mathcal{Q} = \{\mathbf{x}^q\}$: unlabeled images to be segmented, and a support set $\mathcal{S} = \{(\mathbf{x}^s_l, \mathbf{y}^s_l(c^{\hat{j}}))\}$: images $\mathbf{x}^s_l$'s and masks $\mathbf{y}^s_l(c^{\hat{j}})$'s which are references for segmenting class $c^{\hat{j}}$. One episode comprises an N-shot segmentation sub-problem with N image-label pairs of class $c^{\hat{j}}$ provided in the support. The subscript $l = 1, 2, 3, ..., N$ denotes the l-th sample pair in the support. For the ease of illustration, without losing generality, in the following section, we assume only one foreground class is present at a time, i.e. $|\{c^{\hat{j}}\}| = 1$, and the few-shot segmentation reduces to foreground ($c^{\hat{j}}$) – background (c^0) segmentation.

B. Adaptive Local Prototype Pooling Network

1) Overview: As depicted in Fig. 3, the proposed ALPNet is composed of three major components: (1) a generic feed-forward convolutional network parameterized by θ: $f_\theta(\cdot) : \mathcal{X} \rightarrow \mathcal{E}$, which extracts representations from images ($\mathcal{E}$ denotes the feature space); (2) the adaptive local prototype pooling (ALP) module: $\mathcal{E} \times \mathcal{Y} \rightarrow \mathcal{E}$, which is used for computing representation prototypes from the support feature map; (3) a proposed local-to-global similarity-based prediction process: $\mathcal{E} \times \mathcal{E} \rightarrow \mathcal{Y}$ for making the final segmentation.

As shown in Fig. 3 (a), both support and query feature maps: $\mathbf{z}^s_l, \mathbf{z}^q \in \mathcal{E}$, are extracted by passing the support image $\mathbf{x}^s_l$ and the query image $\mathbf{x}^q$ to the network $f_\theta(\cdot)$. Then, the support feature map $\mathbf{z}^s_l$ and the corresponding binary semantic mask $\mathbf{y}^s_l(c^{\hat{j}})$ are used by the ALP module to compute ensembles of representation prototypes $\{\mathcal{P}(c^j)\}$ for each class c^j (including the *background* class), as shown in dotted boxes in Fig. 3 (a). Then, for each class, we measure how similar the query feature map is to each prototype. These measurements are in the form of similarity maps (as seen in the right half of Fig. 3 (a)). These similarity maps are stitched together by each class of the prototype to form the prediction.

2) Adaptive local prototype pooling module: The proposed adaptive local prototype pooling module takes the support feature map $\mathbf{z}^s_l$ and the binary mask $\mathbf{y}^s_l(c^{\hat{j}})$ of the foreground class $c^{\hat{j}}$ as input, and computes *local prototypes* and *class-level prototypes*, of class $c^{\hat{j}}$ or of the background c^0. Specifically, *local prototypes* are obtained by locally averaging the support feature map $\mathbf{z}^s_l \in \mathbb{R}^{D \times H \times W}$ (D to be the channel depth and H, W to be the spatial sizes) with local pooling windows of size (L_H, L_W). This process is illustrated in Fig. 3 (b), where the pooling windows are drawn as orange boxes. This pooling window size decides the spatial extent over which each representation prototype covers.

In practice, this local averaging operation is achieved by passing the support feature map $\mathbf{z}^s_l$ to an average pooling layer, yielding the average-pooled support feature $\text{avgpool}(\mathbf{z}^s_l) \in \mathbb{R}^{D \times \frac{H}{L_H} \times \frac{W}{L_W}}$. Each 1-D feature vector of the average-pooled support feature map at the location $(m, n) \in \mathbb{N}^2$ is now a local prototype. We note the local prototype arises from spatial position (m, n) of $\text{avgpool}(\mathbf{z}^s_l)$ as $p_{l,mn}(c) \in \mathbb{R}^{D \times 1 \times 1}$. The above process is written as follows:

$$p_{l,mn}(c) = \text{avgpool}(\mathbf{z}_l^s)(m,n) = \frac{1}{L_H L_W}\sum_h \sum_w \mathbf{z}_l^s(h,w), \tag{1}$$

$$\text{where } mL_H \leq h < (m+1)L_H,\ nL_W \leq w < (n+1)L_W$$
$$m = 0,1,2,3,...,{}^H/_{L_H}-1,\ n = 0,1,2,3,...,{}^H/_{L_W}-1.$$

Here c is the class of this prototype, yet to be undecided at this stage.

We then decide the semantic class c of each local prototype $p_{l,mn}(c)$. This is done by average-pooling the binary mask $\mathbf{y}_l^s$ to the same spatial size $(\frac{H}{L_H}, \frac{W}{L_W})$ as avgpool$(\mathbf{z}_l^s)$, and fetching the corresponding value at the same spatial location (m,n). Let $y_{l,mn}^a \in [0,1]$ to be the value of average-pooled $\mathbf{y}_l^s$ at location (m,n), the class c of the prototype $p_{l,mn}(c)$ is then given by:

$$c = \begin{cases} c^0 & y_{l,mn}^a < T \\ c^{\hat{j}} & y_{l,mn}^a \geq T \end{cases} \quad \text{where} \quad y_{l,mn}^a = \text{avgpool}(\mathbf{y}_l^s(c^{\hat{j}}))(m,n), \tag{2}$$

where T is a threshold for categorizing the prototype as either class $c^{\hat{j}}$ (the foreground class) or as the background class c^0. T is empirically set to 0.95. Through the process shown above, we have now obtained all the local prototypes.

To further obtain a holistic representation of the foreground class $c^{\hat{j}}$, as well as to account for objects smaller than the local pooling window (L_H, L_W) in $\mathbf{z}_l^s$, we also compute a *class-level prototype* $p_l^g(c^{\hat{j}})$ for the foreground class, using masked-average pooling [13]. This is done by spatially averaging the support feature map $\mathbf{z}_l^s$ underneath the entire binary mask $\mathbf{y}_l^s$ of the object, namely:

$$p_l^g(c^{\hat{j}}) = \frac{\sum\limits_h \sum\limits_w \mathbf{y}_l^s(c^{\hat{j}})(h,w)\mathbf{z}_l^s(h,w)}{\sum\limits_h \sum\limits_w \mathbf{y}_l^s(c^{\hat{j}})(h,w)}. \tag{3}$$

For convenience, we put local prototypes $p_{l,mn}$'s and class-level prototypes p_l^g's together, and bin them into different prototype ensembles $\{\mathcal{P}(c^j)\}$ according to the class c^j of each prototype. Each prototype ensemble $\mathcal{P}(c^j)$ contains all the prototypes of class c^j. We re-index prototypes in each ensemble $\mathcal{P}(c^j)$ using subscript $k_j = 1,2,3,...,K_j$, namely, $p_{k_j}(c^j)$ to be the k_j-th prototype of class c^j, $\mathcal{P}(c^j) = \{p_{k_j}(c^j)\}$, and $K_j = |\mathcal{P}(c^j)|$.

Under an $N > 1$-shot scenario (*i.e.* multiple support samples $(\mathbf{x}_l^s, \mathbf{y}_l^s(c^j))$'s are available during testing time), we repeat the same prototype computation process for each support image. We gather all obtained prototypes together and bin them to the prototype ensembles by their classes, regardless of which support sample that a prototype is from, and then continue with the segmentation in one pass using the same process as described below. We have also experimented with an alternative mechanism: N independent 1-shot predictions are first made based on N support samples. Then, these N predictions are blended together to form the final prediction. However, we did not observe any consistent benefit compared to our default settings.

3) Local-to-global similarity-based prediction process: Once prototypes are computed, we use them as references of each class c^j, to measure how similar each feature vector $\mathbf{z}^q(h,w) \in \mathcal{R}^{D\times 1\times 1}$ at spatial location (h,w) of the query feature map $\mathbf{z}^q$, is to each class c^j. We then predict the class of the query at location (h,w) to be the class of the most similar prototype. Intuitively, as most of prototypes are computed over a small pooling window (L_H, L_W) instead of the entire object, each prototype is only expected to match the most similar local part in the query (*e.g.* to match a *heart*, a prototype whose pooling window falls over the *left ventricle* region, only tries to match a *left-ventricle*-like region in the query, rather than to match the entire *heart*). As shown in Fig. 3 (a), these similarity measurements are termed as *local similarity maps*, noted as $\{S_{k_j}(c^j)\}$. Each $S_{k_j}(c^j)$ corresponds to a prototype $p_{k_j}(c^j)$ and reflects how similar each vector of the query feature map $\mathbf{z}^q$ is to a particular prototype $p_{k_j}(c^j)$.

Then, as shown in the part of Fig. 3 after the *cos* icon, all the local similarities $\{S_{k_j}(c^j)\}$ corresponding to a same class c^j are stitched together into a global pixel-wise similarity map called *class-wise similarity*, noted as $S'(c^j)$ (*e.g.* local similarities of all *ventricles* and *atria* are stitched together into an overall similarity to the *heart* class). The class-wise similarity $S'(c^j)$ reflects how similar each feature vector of the query feature map $\mathbf{z}^q$ to the class c^j. The final prediction is obtained by normalizing all class-wise similarities into probabilities.

Specifically, for each prototype $p_{k_j}(c^j) \in \mathcal{P}$, we compute a local similarity map $S_{k_j}(c^j)$ between the prototype and the query feature map $\mathbf{z}^q$. The local similarity score $S_{k_j}(c^j)(h,w)$ at location (h,w) of the local similarity map $S_{k_j}(c^j)$ is given by

$$S_{k_j}(c^j)(h,w) = \alpha \text{sim}(p_{k_j}(c^j), \mathbf{z}^q(h,w)), \tag{4}$$

where sim$(\cdot,\cdot)$ is a similarity measurement, which we configure as consine similarity [13]. α is a temperature factor $(\alpha > 0)$ [13]. During training α controls the strength of penalty on any pairs of $(p_{k_j}(c^j), \mathbf{z}^q(h,w))$ that yields a high local similarity score $S_{k_j}(c^j)(h,w)$ but with $p_{k_j}(c^j)$ and $\mathbf{z}^q(h,w)$ coming from different classes. We set $\alpha = 20$ to be consistent with [13].

Then, as shown in the part of Fig. 3 after the *cos* icon, for each class c^j, we stitch all K_j local similarity maps $S_{k_j}(c^j)$'s together to form the class-wise similarity $S'(c^j)$. Specifically, we take each entry $S'(c^j)(h,w)$ at location (h,w) to be the (soft) maximum of all local similarities $\{S_{k_j}(c^j)(h,w)\}$ at the same location (h,w), namely:

$$S'(c^j)(h,w) = \sum_{k_j} S_{k_j}(c^j)(h,w)\ [\underset{k'_j}{\text{softmax}}(\{S_{k'_j}(c^j)(h,w)\})](k_j). \tag{5}$$

Here $[\underset{k'_j}{\text{softmax}}(\{S_{k'_j}(c^j)(h,w)\})](k_j)$ refers to the operation that is composed of the following three steps: (1) stacking all (in total, K_j) $S_{k_j}(c^j)(h,w)$'s along the channel dimension, yielding a tensor with shape $K_j \times 1 \times 1$; (2) then computing the softmax function along all K_j channels; (3) fetching the k_j-th channel of the output obtained by the softmax function. By repeating this computation process for all classes, we obtain class-wise similarity maps $S'(c_j)$'s between the query feature map $\mathbf{z}^q$ and each class $c_j \in \{c^0, c^{\hat{j}}\}$.

In the end, to obtain the final prediction $\hat{\mathbf{y}}^q$ in the form of probability, we stack all class-wise similarities $S'(c^j)$'s along the channel dimension, and apply another softmax function along the channel dimension, namely:

$$\hat{\mathbf{y}}^q(h,w) = \underset{c^j}{\mathrm{softmax}}(\{S'(c^j)(h,w)\}). \tag{6}$$

We then interpolate $\hat{\mathbf{y}}^q$ back to the original size of input images.

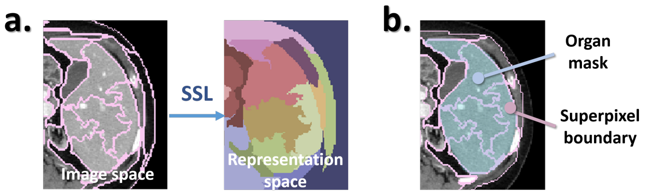

Fig. 4. (a). Conceptual illustration of the piece-wise smoothness prior: In image space, neighboring pixels within a superpixel (delineated by pink boundaries) often share similar textures. This similarity property in texture can also be observed in neighboring pixels within semantically meaningful labels. We therefore argue this smoothness prior is generalizable and can be leveraged for segmentation. Therefore, we encourage the network to learn to preserve this piece-wise smoothness property of images in the corresponding representation space, via the SSL. In this figure, each color in representation space denotes one distinct cluster of similar feature vectors. (b). Boundaries of real semantic labels (organs) and pseudolabels are usually associated with drastic changes in image gradient, and both types of boundaries sometimes overlap. This suggests that boundary prior learned on superpixels are generalizable to real objects. Of note, in our experiments on abdominal images, objects of testing classes are strictly excluded in training images even though unlabeled.

C. Superpixel-based self-supervised learning

In this part, we first explain the intuition behind the proposed superpixel-based self-supervised learning approach. We then introduce the process for computing pseudolabels. Finally, we discuss the detailed mathematical formulation of the loss functions and the overall training process.

1) Intuition: To tailor self-supervised learning to few-shot segmentation, the proposed superpixel-based self-supervised learning (SSL) technique learns image representations that are robust against inter-patient variability. It also fully exploits patch-level image priors shared between pseudolabels and semantic labels. This allows the network to learn image representations that enable high spatial accuracy in few-shot segmentation. In principle, for a similarity-based segmentation process as described in Sec. III-B3, it is crucial for the features to be invariant of variabilities among instances of a same semantic class (e.g. differences in intensity and shape of a same organ between two patients). This invariance ensures that the retrieval from the support to the query of a same class to be robust. Meanwhile, to ensure representations learned on SSL to be well-generalizable to semantically meaningful labels, it is desirable for the pseudolabels employed in pretext task to share the same prior knowledge and properties with real labels that the model may encounter in the downstream segmentation task. Despite the recent hype of self-supervised learning [30], [32], [39], [41], exploiting image priors for downstream dense prediction tasks is hardly discussed.

The proposed SSL encourages feature invariance and it exploits two shared prior knowledge: a piece-wise smoothness property [34] of pixels in both pseudolabels and semantic labels, which we refer to as the *smoothness prior*; and a property shared between boundaries of pseudolabels and real semantic labels that both types of boundaries are usually associated with large values of the image gradient. We refer to this property as the *boundary prior*. An illustration of two priors are shown in Fig. 4.

As shown in Fig. 5, the SSL task is formulated as segmenting superpixel-based pseudolabels, by referring to their randomly transformed copies. In this process, invariances to these applied transformations are naturally enforced by the gradients that are back-propagated from the similarity-based segmentation mechanism. In addition, as shown in Fig. 4 (a), to embed smoothness prior of semantic labels into learned representations, we deliberately employ superpixels rather than simple rectangular patches [2] as pseudolabels. This is because superpixels themselves are usually generated following the piece-wise smoothness model [34], [35]. Moreover, to boost segmentation accuracy, we exploit the boundary prior shared between pseudolabels and real semantically meaningful regions. As shown in Fig. 4 (b), both boundaries of pseudolabels and those of real semantically meaningful labels often occur at regions where intensity drastically changes. Therefore, if the learned representations were able to precisely grasp superpixel boundaries, this capability of capturing boundaries would be likely to generalize to real semantic labels. This boundary prior therefore benefits segmentation accuracy. In practice, the boundary prior is injected to the learned representations with a simple but effective boundary loss [68] between the ground truth boundary map and the soft edge map of prediction.

As shown in Fig. 5, the proposed SSL stategy comprises an *offline unsupervised pseudolabel generation* phase and an *online episode-based training* phase.

2) Unsupervised pseudolabel generation: As shown in the left part of Fig. 5, a set $\mathbf{Y}_i^p$ of pseudolabel candidates are essentially superpixels of an unlabeled image $\mathbf{x}_i$, where p stands for *pseudolabel*. These superpixels are computed efficiently with an offline unsupervised graph-based method [35] before the online training phase.

3) Online episode composition: The network is trained in *episodes*: few-shot segmentation sub-problems comprised by randomly chosen support and query sets. Specifically, we formulate the self-supervision task as a foreground-background segmentation problem: the network is required to retrieve a randomly sampled superpixel (*i.e.* the *foreground*) in a transformed image.

As shown in Fig. 5, for each episode i, the support $\mathcal{S}_i$ is formed by a randomly-sampled image $\mathbf{x}_i$ in together with one of its randomly chosen superpixel $\mathbf{y}_i^r(c^p) \in \mathbf{Y}_i^p$, namely, $\mathcal{S}_i = \{(\mathbf{x}_i, \mathbf{y}_i^r(c^p))\}$. $\mathbf{y}_i^r(c^p)$ is in the form of binary mask with index $r = 1, 2, 3, ..., |\mathbf{Y}_i^p|$. The upperscript p of c^p stands for the *pseudolabel* class (*i.e.* the *foreground* class), which is assigned to the chosen superpixel. The rest of the image

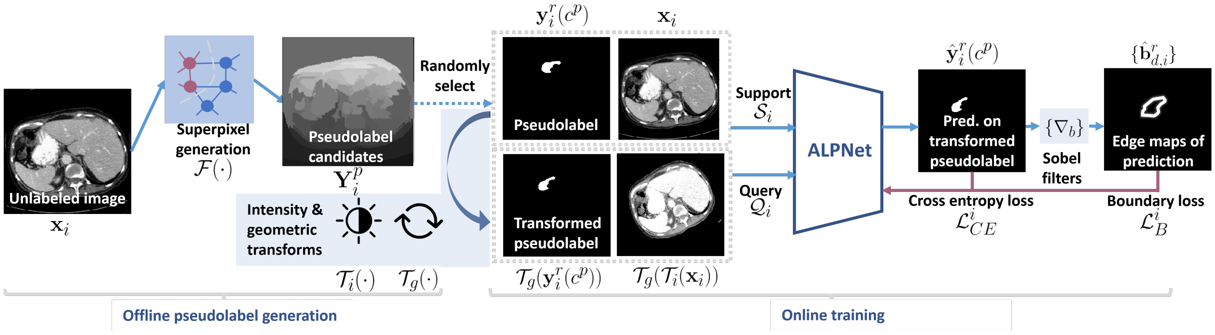

Fig. 5. Workflow of the proposed superpixel-based self-supervised learning technique.

is therefore given the *background* class c^0. The mask of the *background* class is naturally computed as $1 - \mathbf{y}_i^r(c^p)$. Of note, regardless of which specific superpixel is chosen as the pseudolabel, we always give the pseudolabel a same class label c^p across iterations. This is because the prototype-based 1-way few-shot segmentation problem can be interpreted as retrieving the same object across the labeled and the unlabeled images, where fixed class-id's are unnecessary.

We obtain the corresponding query set by applying a random geometric transformation $\mathcal{T}_g(\cdot)$ to $\mathcal{S}_i$, and a random intensity transformation $\mathcal{T}_i(\cdot)$ to $\mathbf{x}_i$, namely $\mathcal{Q}_i = \{(\mathcal{T}_g(\mathcal{T}_i(\mathbf{x}_i)))\}, \mathcal{T}_g(\mathbf{y}_i^r(c^p)))$. Specifically, $\mathcal{T}_g(\cdot)$ includes affine transformation and elastic transformation and $\mathcal{T}_i(\cdot)$ is gamma transformation. By this mean, one training episode $(\mathcal{S}_i, \mathcal{Q}_i)$ is formed. In practice, we employ the simplest 1-shot scenario for each episode for SSL during training, as we haven't observed significant benefit of using higher shots in SSL even when tested with $N > 1$ shots.

In each iteration, the prediction $\hat{\mathbf{y}}_i^r(c^p)$ of query pseudolabel $\mathcal{T}_g(\mathbf{y}_i^r(c^p))$ is obtained by feeding $(\mathcal{S}_i, \mathcal{Q}_i)$ to the ALPNet, and taking steps described in Section III.

4) Boundary prior and boundary loss: To embed the boundary prior into the learned representations, we employ an l-1 boundary loss between the predicted pseudolabel and its ground truth [68]. Specifically, as shown in the right part of Fig. 5, we first obtain the soft edge maps $\{\hat{\mathbf{b}}_{d,i}^r\}$ of the prediction $\hat{\mathbf{y}}_i^r(c^p)$ and the ground truth boundary maps $\{\mathbf{b}_{d,i}^r\}$ of the ground truth $\mathcal{T}_g(\mathbf{y}_i^r(c^p))$. This is obtained by applying 3×3 Sobel filters $\{\nabla_d\}$ along the height dimension and the width dimension separately. Here the lowerscript $d \in \mathcal{D}_{irs} = \{height, width\}$ denotes the direction of Sobel filter (along which image gradient is computed). We have:

$$\begin{aligned} \hat{\mathbf{b}}_{d,i}^r &= \nabla_d * \hat{\mathbf{y}}_i^r(c^p) \text{ and} \\ \mathbf{b}_{d,i}^r &= \nabla_d * \mathcal{T}_g(\mathbf{y}_i^r(c^p)), \text{ where } d \in \mathcal{D}_{irs}, \end{aligned} \tag{7}$$

where $*$ denotes convolution. we then compute l-1 distances between two sets of maps, and obtain the boundary loss $\mathcal{L}_B^i(\theta; \mathcal{S}_i, \mathcal{Q}_i)$ for episode i, namely:

$$\mathcal{L}_B^i(\theta; \mathcal{S}_i, \mathcal{Q}_i) = \frac{1}{HW} \sum_{d \in \mathcal{D}_{irs}} \sum_h^H \sum_w^W |\, \hat{\mathbf{b}}_{d,i}^r(h,w) - \mathbf{b}_{d,i}^r(h,w) \,| \,. \tag{8}$$

An alternative boundary loss can be found in [69].

5) End-to-end training: The SSL framework is trained in a straightforward end-to-end manner, where each iteration takes an episode, and the gradients back-propagate to the network weights θ in a similar way as in fully-supervised scenarios. On top of the boundary loss discussed above, we also employ the commonly-used cross-entropy loss between $\hat{\mathbf{y}}_i^r(c^p)$ and $\mathcal{T}_g(\mathbf{y}_i^r(c^p))$. For each episode i we have

$$\mathcal{L}_{CE}^i(\theta; \mathcal{S}_i, \mathcal{Q}_i) = -\frac{1}{HW} \sum_h^H \sum_w^W \sum_{j \in \{0,p\}} \mathcal{T}_g(\mathbf{y}_i^r(c^j))(h,w) \log(\hat{\mathbf{y}}_i^r(c^j)(h,w)). \tag{9}$$

Of note, from the perspective of contrastive representation learning, this cross-entropy loss is in analogy with InfoNCE loss [32], [41], applied to superpixel-based pseudolabels.

Following the practice of [13], we employ *prototypical alignment regularization* [13]. This improves model performance by aligning features of a same class between the support images and the predicted query image. This is achieved by swapping the roles of the support and the query: It first takes the query image $\mathcal{T}_g(\mathcal{T}_i(\mathbf{x}_i))$ and its predicted segmentation $\hat{\mathbf{y}}_i^r(c^p))$ as a new support $\mathcal{S}_i'$, and it takes the original support image as a new query $\mathcal{Q}_i' = \{\mathbf{x}_i\}$. Then, it enforces the new support $\mathcal{S}_i' = \{\mathcal{T}_g(\mathcal{T}_i(\mathbf{x}_i)), \hat{\mathbf{y}}_i^r(c^p)\}$ being able to be used as reference to segmentation $\mathbf{x}_i$. This regularization is written as follows:

$$\mathcal{L}_{REG}^i(\theta; \mathcal{S}_i', \mathcal{Q}_i') = -\frac{1}{HW} \sum_h^H \sum_w^W \sum_{j \in \{0,p\}} \mathbf{y}_i^r(c^j)(h,w) \log(\tilde{\mathbf{y}}_i^r(c^j)(h,w)). \tag{10}$$

Here $\tilde{\mathbf{y}}_i^r(c^p)$ is the prediction of $\mathbf{y}_i^r(c^p)$ taking $\mathbf{x}_i$ as query (*i.e.* $\hat{\mathbf{y}}_i^r(c^p)$ denotes the prediction in the first pass, while $\tilde{\mathbf{y}}_i^r(c^p)$ denotes the prediction after the support and the predicted query are swapped).

Overall, in each iteration (episode) i, the training objective is defined as follows:

$$\mathcal{L}^i(\theta; \mathcal{S}_i, \mathcal{Q}_i) = \mathcal{L}_{CE}^i + \lambda_B \mathcal{L}_B^i + \lambda_{REG} \mathcal{L}_{REG}^i, \tag{11}$$

where weights λ_B and λ_{REG} are both empirically set to 1.0. Once the iterative training process is finished, the network can be directly used for few-shot segmentation on real semantic objects, without any fine-tuning process.

IV. Experiments

A. Datasets

We comprehensively evaluated our method on three different combinations of imaging modalities and anatomical structures: abdominal CT [70], abdominal T2-SPIR MRI [71] and cardiac bSSFP MRI [72]. In each dataset, images are equally separated into 5 folds for cross-validation. In each fold, the validation set is further split into *disjoint* sets of support and query images. An illustration of dataset split scheme is shown in Fig. 6. Images in all these datasets are 3-D and contain many slices/regions which do not contain any labeled classes. These slices/regions allow us to compute sufficient pseudolabels. Each 3-D image is reformated as 2-D slices and resampled to 256×256. We replicate the 1-channel 2-D image for three times and stack them along the channel dimension to fit into the network. Following common practices, intensity normalization is applied to 3-D images.

To directly compare the performance of a same set of semantic classes between datasts and modalities, for abdominal CT and abdominal MRI, we constructed a shared label set of *left kidney* (LK), *right kidney* (RK), *liver* and *spleen*. For cardiac MRI, *left ventricle blood pool* (LV-BP), *left ventricle myocardium* (LV-MYO) and *right ventricle* (RV) are included. This label set covers a wide range of objects with various shape, size and textures under different imaging modalities.

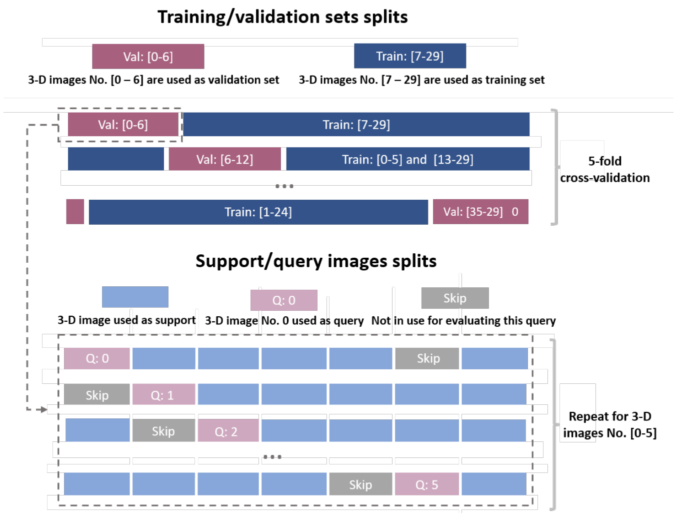
Training/validation sets splits
Val: [0-6]
Train: [7-29]
3-D images No. [0 – 6] are used as validation set
3-D images No. [7 – 29] are used as training set
Val: [0-6]
Train: [7-29]
Val: [6-12]
Train: [0-5] and [13-29]
Train: [1-24]
Val: [35-29] 0
5-fold cross-validation
Support/query images splits
Q: 0
Skip
3-D image used as support
3-D image No. 0 used as query
Not in use for evaluating this query
Q: 0
Skip
Skip
Q: 1
Skip
Q: 2
Skip
Q: 5
Repeat for 3-D images No. [0-5]

Fig. 6. Upper part: training-validation sets splittings for Abdominal CT dataset which contains 30 3D images, numbered as No. 0 - No.29. Lower part: support-query splittings for a validation set, in a 5-shot segmentation scenario, taking the validation of the first validation set as an example.

B. Evaluation protocol

1) Evaluating 3-D segmentation on 2-D models: We employed Dice score [73] as the evaluation metric. It measures the overlaps between the prediction and the ground-truth. As we segment 3-D images using a 2-D model, Dice scores are computed per-patient after re-stacking 2-D predictions back into 3-D volumes.

To assign support slices to query slices, we employed the protocol in [1]. In both query set and support set, for each 3-D image, we first equally divide its region-of-interest into C equal-sized chunks. Then, for each chunk in query, its corresponding support samples are the set of central slices of corresponding chunks from all support scans. We set C=3 throughout our experiments. To rigorously observe the generalization ability of an FSS model to unseen semantic classes in abdominal images, we first group labels into the upper-abdomen and the low-abdomen groups. In each experiment, classes of one group are taken as the training classes while those of the other group comprise the testing classes. We then strictly exclude any slices containing testing classes from the training set. For cardiac images, each time we take one label for testing and the rest for training. Excluding slices containing testing classes is unfeasible due to the view-point constraint.

2) Implementation: We implement the proposed framework using PyTorch, based on vanilla PANet implementation[1] [13]. An off-the-shelf ResNet-101 [74] network pretrained on part of MS-COCO is employed as the feature extractor $f_\theta(\cdot)$ [2]. It has an input dimension of $3 \times 256 \times 256$ and yields a feature map of $256 \times 32 \times 32$. We train the proposed SSL-ALPNet using 1-shot configuration for 100k iterations with an SGD optimizer. Hyper-parameters of the proposed method and baseline methods are decided by manually searching with different combinations.

C. Quantitative and qualitative results

Table I - II show the comparisons of the proposed SSL-ALPNet with the baseline vanilla PANet [13] and the SE-Net[3] [1], a method particularly designed for medical images, as mentioned in Sec. II-A. The proposed SSL-ALPNet consistently outperforms baseline methods by at least 10 points in terms of Dice score in both 1-shot and 5-shot scenarios. Fig. 7 qualitatively demonstrates the desirable performance of the proposed method under different imaging modalities and anatomical structures. In addition, SSL-ALPNet yields consistent performance gains when expanding from 1-shot to 5-shot in testing, demonstrating its data efficiency in terms of fully exploiting additional support examples if available. We have also included the upper bounds by fully-supervised methods for abdominal image segmentation in the last two rows of Table I.

D. Ablation study

1) ALPNet architecture: To justify the benefit of exploiting local information, as argued in Sec. I, in Table I and Table II, we compared ALPNet with the PANet architecture when trained using either manual annotations (*Vanilla PANet* versus *ALPNet*) or self-supervised pseudolabels (*SSL-PANet* versus *SSL-ALPNet w/o BP*, where *BP* is short for *boundary prior*).

[1] https://github.com/kaixin96/PANet

[2] This initialization alone does not contribute much to segmenting medical images [36].

[3] https://github.com/abhi4ssj/few-shot-segmentation

TABLE I
QUANTITATIVE EVALUATION OF THE PROPOSED METHOD AND BASELINE METHODS ON ABDOMINAL IMAGES, MEASURED IN DICE SCORE. "BP" IS SHORT FOR "BOUNDARY PRIOR"

Method	Manual Anno.?	Abdominal-CT					Abdominal-MRI				
		Lower		Upper		Mean	Lower		Upper		Mean
		LK	RK	Spleen	Liver		LK	RK	Spleen	Liver	
Inference with N=1 shot.											
SE-Net [1]	✓	32.70	23.60	32.53	38.20	31.76	63.85	64.56	11.78	55.08	48.82
Vanilla PANet [13]	✓	33.29	16.27	36.47	50.10	34.04	53.95	35.05	58.50	54.14	49.41
ALPNet	✓	28.43	26.68	40.21	55.78	37.78	58.59	57.28	56.77	60.14	58.19
SSL-PANet	×	37.58	34.69	43.73	61.71	44.42	47.71	47.95	58.73	64.99	54.85
SSL-ALPNet w/o BP	×	63.34	54.82	60.25	73.65	63.02	73.63	78.39	67.02	73.05	73.02
SSL-ALPNet w/ BP	×	**66.04**	**62.14**	**68.39**	**73.90**	**67.62**	**78.77**	**83.44**	**70.02**	**75.01**	**76.81**
Inference with N=5 shots.											
Vanilla PANet [13]	✓	31.13	19.28	36.88	57.62	36.23	53.52	34.37	56.13	56.84	50.21
ALPNet	✓	39.90	32.75	45.61	60.90	44.79	67.38	71.84	61.36	64.13	66.18
SSL-PANet	×	38.83	36.42	42.40	69.16	46.70	44.32	42.18	54.49	66.59	51.90
SSL-ALPNet w/o BP	×	71.78	69.31	70.94	**82.05**	73.52	79.90	85.96	72.30	80.40	79.64
SSL-ALPNet w/ BP	×	**74.34**	**71.61**	**75.74**	81.96	**75.91**	**82.28**	**86.23**	**72.42**	**80.70**	**80.16**
Upper bounds by fully-supervised segmentation											
Zhou et al. [75]	Ful. Sup.	95.3	92.0	96.8	97.4	95.4			-		
Isenseen et al. [76]	Ful. Sup.			-					-		94.6

TABLE II
QUANTITATIVE EVALUATION OF THE PROPOSED METHOD AND BASELINE METHODS ON CARDIAC IMAGES.

Method	Manual Anno.?	LV-BP	LV-MYO	RV	Mean
Inference with N=1 shot.					
SE-Net [1]	✓	58.04	25.18	12.86	32.03
Vanilla PANet [13]	✓	53.64	35.72	39.52	42.96
ALPNet	✓	73.08	49.53	58.50	60.34
SSL-PANet	×	70.42	46.79	69.52	62.25
SSL-ALPNet w/o BP	×	**83.99**	66.74	79.96	76.90
SSL-ALPNet w/ BP	×	83.98	**67.68**	**82.15**	**77.94**
Inference with N=5 shots.					
Vanilla PANet [13]	✓	60.69	39.44	41.66	47.26
ALPNet	✓	82.65	52.61	69.13	68.14
SSL-PANet	×	70.62	46.03	67.16	61.27
SSL-ALPNet w/o BP	×	86.88	70.56	85.12	80.85
SSL-ALPNet w/ BP	×	**86.89**	**72.14**	**85.95**	**81.66**
Upper bound by fully-supervised segmentation					
FCN-ResNet-101 [74], [77]	Ful. Sup.	95.81	87.00	93.57	92.13

In both scenarios, consistent performance gains of ALPNet are shown. Of note, this performance boost is achieved with the almost-negligible additional computational cost of introducing the adaptive local prototype pooling module.

2) SSL with image priors: As discussed in Sec. II-C, the proposed SSL leverages two image priors: the smoothness prior enforced by using superpixels as pseudolabels, and the boundary prior (BP) enforced by the boundary loss. By comparing the same ALPNet's trained on manual annotations versus those trained using SSL w/o BP, we can observe the superiority of self-supervised learning in FSS. By a close-up comparison between SSL-ALPNet's with boundary prior and those without, we can observe the performance gains in Table I - II[4]. All these results highlights the benefit of exploiting proper patch-level image priors in self-supervised learning for segmentation, which is usually under-explored in previous FSS or representation learning works.

3) Granularity of prototypes: We also examined the effect of granularity of prototypes: the extend under which each local prototype is computed over the feature map of the support. As shown in Fig. 3-B and Equ. 1, this extend is controlled by (L_H, L_W) (orange boxes in Fig. 3-B). They are empirically set to be $(4, 4)$ during training and $(2, 2)$ during testing, both over support feature maps that have a spatial size of 32×32. This configuration is based on the intuition that during both training and testing, (L_H, L_W) should be reasonably smaller than the potential segmentation target to capture the fine-grained details of the image. However, to enlarge the receptive field of a prototype, during training, (L_H, L_W) should be reasonably large.

TABLE III
EFFECT OF VARYING PROTOTYPE WINDOW SIZE (L_H, L_W) DURING TRAINING ON SEGMENTATION RESULTS ON ABDOMINAL CT

(L_H, L_W)	LK	RK	Spleen	Liver	Mean
(2,2)	**66.42**	58.78	67.21	**73.94**	66.58
(4,4) (reported)	66.04	**62.14**	**68.39**	73.90	**67.62**
(8,8)	61.68	53.89	61.94	69.25	61.69

We therefore experimented with different (L_H, L_W)'s during training, and tested the obtained models on 1-shot abdominal CT segmentation. These results are reported in Table III. These results agree with our intuitions on selecting the size of (L_H, L_W). It is interesting to note that choosing an oversmall (L_H, L_W) during training does not severely affect the performance. This might be because the backbone network has already provided a reasonably large receptive field.

E. Testing with weak annotations

We have also examined the behavior of the model trained on SSL under an extremely low-resource scenario, which we term as *weakly-annotated testing*. Under this scenario, only

[4] For rigorousness we performed single-sided Wilxocon signed-rank tests for scenarios where the performance gain of introducing BP is smaller than +1.5 out of 100 in terms of mean Dice scores, across Table I-II & IV. p-values for all tests are below 0.01 except for the 5-shot abdominal MRI segmentation.

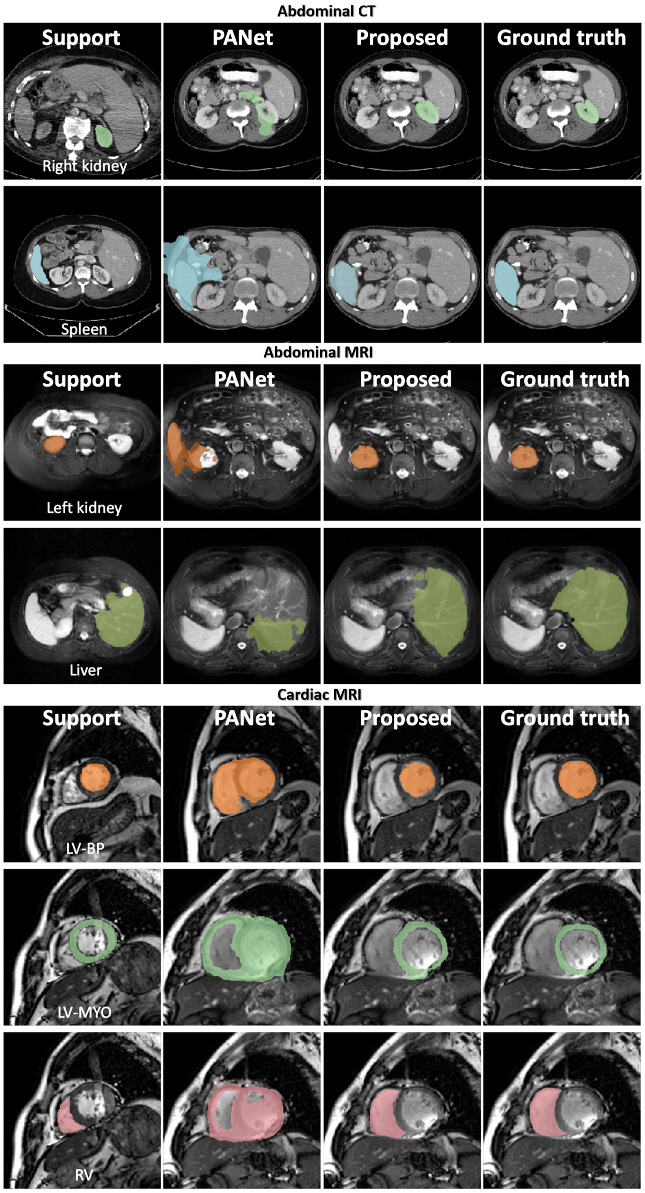

Fig. 7. Qualitative results of the proposed method, in comparison with baseline method PANet [13].

TABLE IV
QUANTITATIVE RESULTS OF WEAK (1-SHOT WITH BOUNDING BOX ONLY) ANNOTATION DURING TESTING

Method	Abdominal CT	Abdominal MRI	Cardiac MRI
ALPNet	38.21	52.43	41.46
SSL-ALPNet w/o BP	48.73	64.80	46.71
SSL-ALPNet w/ BP	**57.56**	**68.25**	**47.17**

bounding-box-level annotations are available for each class under a 1-shot setting. As shown in Table IV and Fig. 8, by leveraging two image priors in learned representations, the model is still able to reasonably predict the rough shapes and boundaries of unseen anatomical structures.

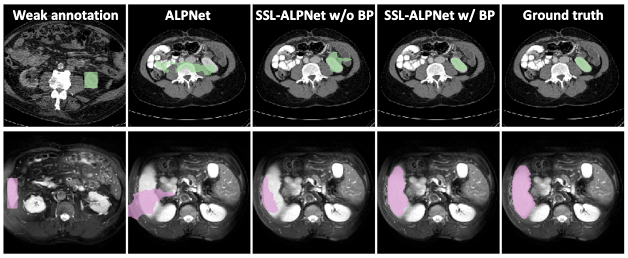

Fig. 8. Examples of predictions under weak (1-shot with bounding box) annotation during testing

V. Conclusion and Discussion

In this work, we have proposed a self-supervised framework, named SSL-ALPNet, for few-shot medical image segmentation. Without using any manual labels during training, the proposed method successfully outperforms previous methods that rely on large amounts of annotated data of training classes.

From a broader perspective, self-supervised representation learning is a family of promising techniques for learning from rich but under-exploited unlabeled medical images. In comparison with 2-D RGB images, medical images features their own unique properties. For example, medical images are usually in 3-D and are highly structured. These unique properties promise future works on self-supervised learning for medical imaging applications. In our method, the piece-wise smoothness prior and the boundary prior of medical images are proven to be beneficial for few-shot segmentation.

Although the proposed self-supervised learning technique has demonstrated superiority over conventional methods, some extensions remain to be made. For example, the current method is designed for 1-way segmentation: only one label class to be segmented at a time. To expand to multi-way segmentation where more than one classes are to be segmented, adjustments to the SSL technique need to be made. For example, sampling multiple superpixels at one time and labeling them with different pseudolabels would be a straightforward solution. Also, lesions are often more difficult for a prototype-based network to segment compared with organs, since many types of lesions do not have regular textures or shapes. Therefore, both the SSL strategy and the prototype-based network need to be upgraded to account for these lesions. Interestingly, a recent work [78] which employs a similar self-supervision technique to ours, has demonstrated promising results on one-shot lesion retrieval. To further automate few-shot segmentation during testing time, it is also desirable to simplify the chunking mechanism for assigning support 2-D slices to query images.

For few-shot medical image segmentation in general, several questions remain to be investigated. For example, as 3-D networks are generally regarded as superior to 2-D counterparts in terms of segmentation accuracy [79], a fully 3-D few-shot medical image segmentation technique with high computational efficiency remains to be proposed. Meanwhile, as medical images are usually accompanied by non-image information like radiology reports, leveraging multi-modality information in self-supervised learning will be promising.

Also, designing universal self-supervised learning techniques which target at a wider range of downstream tasks is of great practical value.

In summary, we have successfully designed a self-supervised few-shot medical image segmentation framework. Circumventing the need for manually annotated training data, this work potentially expands future applications of few-shot segmentation in medical images. We hope that the proposed superpixel-based self-supervision technique would inspire future investigations on few-shot segmentation and unsupervised representation learning.

References

[1] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "'squeeze & excite' guided few-shot segmentation of volumetric images," *Medical image analysis*, vol. 59, p. 101587, 2020.

[2] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *arXiv preprint arXiv:2006.10511*, 2020.

[3] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[4] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[5] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[6] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[7] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[8] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, "One shot learning of simple visual concepts," in *Proceedings of the annual meeting of the cognitive science society*, vol. 33, 2011.

[9] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018.

[10] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[11] N. Dong and E. Xing, "Few-shot semantic segmentation with prototype learning." in *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, vol. 3, no. 4, 2018, p. 79.

[12] M. Siam, B. N. Oreshkin, and M. Jagersand, "Amp: Adaptive masked proxies for few-shot segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5249–5258.

[13] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9197–9206.

[14] X. Zhang, Y. Wei, Y. Yang, and T. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *arXiv preprint arXiv:1810.09091*, 2018.

[15] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Few-shot segmentation propagation with guided networks," *arXiv preprint arXiv:1806.07373*, 2018.

[16] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9587–9595.

[17] M. Siam, N. Doraiswamy, B. N. Oreshkin, H. Yao, and M. Jagersand, "Weakly supervised few-shot object segmentation using co-attention with visual and semantic inputs," *arXiv preprint arXiv:2001.09540*, 2020.

[18] P. Tian, Z. Wu, L. Qi, L. Wang, Y. Shi, and Y. Gao, "Differentiable meta-learning model for few-shot semantic segmentation," *arXiv preprint arXiv:1911.10371*, 2019.

[19] T. Hu, P. Mettes, J.-H. Huang, and C. G. Snoek, "Silco: Show a few images, localize the common object," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5067–5076.

[20] S. M. Hendryx, A. B. Leach, P. D. Hein, and C. T. Morrison, "Meta-learning initializations for image segmentation," *arXiv preprint arXiv:1912.06290*, 2019.

[21] D. Lieb, A. Lookingbill, and S. Thrun, "Adaptive road following using self-supervised learning and reverse optical flow." in *Robotics: science and systems*, 2005, pp. 273–280.

[22] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430.

[23] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Advances in neural information processing systems*, 2014, pp. 766–774.

[24] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision*. Springer, 2016, pp. 577–593.

[25] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84.

[26] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8059–8068.

[27] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[28] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666.

[29] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9865–9874.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[31] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[33] J. Liu and Y. Qin, "Prototype refinement network for few-shot segmentation," *arXiv preprint arXiv:2002.03579*, 2020.

[34] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on pure and applied mathematics*, vol. 42, no. 5, pp. 577–685, 1989.

[35] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

[36] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervision with superpixels: Training few-shot medical image segmentation without annotation," in *European Conference on Computer Vision*. Springer, 2020, pp. 762–780.

[37] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[38] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 535–15 545.

[39] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.

[40] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[41] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[42] X. Cao and X. Zhen, "Few-shot semantic segmentation with democratic attention networks," 2020.

[43] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot

learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.

[44] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transforms for one-shot medical image segmentation," in *CVPR*, 2019.

[45] A. K. Mondal, J. Dolz, and C. Desrosiers, "Few-shot 3d multi-modal medical image segmentation using generative adversarial learning," *arXiv preprint arXiv:1810.12241*, 2018.

[46] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert, "Data efficient unsupervised domain adaptation for cross-modality image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 669–677.

[47] H. Yu, S. Sun, H. Yu, X. Chen, H. Shi, T. S. Huang, and T. Chen, "Foal: Fast online adaptive learning for cardiac motion estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4313–4323.

[48] C. Chen, C. Qin, H. Qiu, C. Ouyang, S. Wang, L. Chen, G. Tarroni, W. Bai, and D. Rueckert, "Realistic adversarial data augmentation for mr image segmentation," *arXiv preprint arXiv:2006.13322*, 2020.

[49] H. Cui, D. Wei, K. Ma, S. Gu, and Y. Zheng, "A unified framework for generalized low-shot medical image segmentation with scarce data." *IEEE Transactions on Medical Imaging*, 2020.

[50] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *Advances in Neural Information Processing Systems*, 2019, pp. 6447–6458.

[51] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised and task-driven data augmentation," in *International conference on information processing in medical imaging*. Springer, 2019, pp. 29–41.

[52] A. Makarevich, A. Farshad, V. Belagiannis, and N. Navab, "Metamedseg: Volumetric meta-learning for few-shot organ segmentation," *arXiv preprint arXiv:2109.09734*, 2021.

[53] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[54] A. R. Feyjie, R. Azad, M. Pedersoli, C. Kauffman, I. B. Ayed, and J. Dolz, "Semi-supervised few-shot learning for medical image segmentation," *arXiv preprint arXiv:2003.08462*, 2020.

[55] L. Sun, C. Li, X. Ding, Y. Huang, Z. Chen, G. Wang, Y. Yu, and J. Paisley, "Few-shot medical image segmentation using a global correlation network with discriminative embedding," *Computers in biology and medicine*, vol. 140, p. 105067, 2022.

[56] Q. Yu, K. Dang, N. Tajbakhsh, D. Terzopoulos, and X. Ding, "A location-sensitive local prototype network for few-shot medical image segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 262–266.

[57] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2051–2060.

[58] X. Zhan, X. Pan, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised learning via conditional motion propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1881–1889.

[59] A. Jamaludin, T. Kadir, and A. Zisserman, "Self-supervised learning for spinal mris," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 294–302.

[60] Z. Zhou, V. Sodha, M. M. R. Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, "Models genesis: Generic autodidactic models for 3d medical image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 384–393.

[61] W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews, and D. Rueckert, "Self-supervised learning for cardiac mr image segmentation by anatomical position prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 541–549.

[62] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101539, 2019.

[63] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[64] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," *arXiv preprint arXiv:2005.10242*, 2020.

[65] X. Ren and J. Malik, "Learning a classification model for segmentation," in *null*. IEEE, 2003, p. 10.

[66] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[67] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 142–158.

[68] C. Chen, C. Ouyang, G. Tarroni, J. Schlemper, H. Qiu, W. Bai, and D. Rueckert, "Unsupervised multi-modal style transfer for cardiac mr segmentation," *arXiv preprint arXiv:1908.07344*, 2019.

[69] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International conference on medical imaging with deep learning*. PMLR, 2019, pp. 285–296.

[70] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," 2015.

[71] A. E. Kavur, N. S. Gezer, M. Barış, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar *et al.*, "Chaos challenge–combined (ct-mr) healthy abdominal organ segmentation," *arXiv preprint arXiv:2001.06535*, 2020.

[72] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 12, pp. 2933–2946, 2018.

[73] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[75] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, and A. L. Yuille, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 672–10 681.

[76] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert *et al.*, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[77] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[78] K. Yan, J. Cai, D. Jin, S. Miao, A. P. Harrison, D. Guo, Y. Tang, J. Xiao, J. Lu, and L. Lu, "Self-supervised learning of pixel-wise anatomical embeddings in radiological images," *arXiv preprint arXiv:2012.02383*, 2020.

[79] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.