Self-supervision with Superpixels: Training Few-Shot Medical Image Segmentation Without Annotation

Cheng Ouyang(✉), Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert

BioMedIA Group, Department of Computing, Imperial College London, London, UK
c.ouyang@imperial.ac.uk

Abstract. Few-shot semantic segmentation (FSS) has great potential for medical imaging applications. Most of the existing FSS techniques require abundant annotated semantic classes for training. However, these methods may not be applicable for medical images due to the lack of annotations. To address this problem we make several contributions: (1) A novel self-supervised FSS framework for medical images in order to eliminate the requirement for annotations during training. Additionally, superpixel-based pseudo-labels are generated to provide supervision; (2) An adaptive local prototype pooling module plugged into prototypical networks, to solve the common challenging foreground-background imbalance problem in medical image segmentation; (3) We demonstrate the general applicability of the proposed approach for medical images using three different tasks: abdominal organ segmentation for CT and MRI, as well as cardiac segmentation for MRI. Our results show that, for medical image segmentation, the proposed method outperforms conventional FSS methods which require manual annotations for training.

1 Introduction

Automated medical image segmentation is a key step for a vast number of clinical procedures and medical imaging studies, including disease diagnosis and follow-up [1–3], treatment planning [4,5] and population studies [6,7]. Fully supervised deep learning based segmentation models can achieve good results when trained on abundant labeled data. However, the training of these networks in medical imaging is often impractical due to the following two reasons: there is often a lack of sufficiently large amount of expert-annotated data for training due the considerable clinical expertise, cost and time associated with annotation; This problem is further exacerbated by differences in image acquisition procedures

C. Biffi, C. Chen and T. Kart—Equal contribution.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58526-6_45) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12374, pp. 762–780, 2020.
https://doi.org/10.1007/978-3-030-58526-6_45

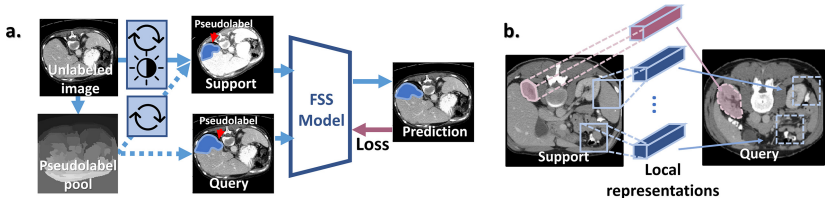

Fig. 1. (a). Proposed superpixel-based self-supervised learning. For each unlabeled image, pseudolabels are generated on superpixels. In each iteration during training, a randomly selected pseudolabel and the original image serve as the candidate for both support and query. Then, random transforms (marked in blue boxes) are applied between the support and the query. The self-supervision task is designed as segmenting the pseudolabel on the query with reference to the support, despite the transforms applied in between. (b). The proposed ALPNet solves the class-imbalance-induced ambiguity problem by adaptively extracting multiple local representations of the large background class (in blue). Each of them only represents a local region of background. (Color figure online)

across medical devices and hospitals, often resulting in datasets containing few manually labeled images; Moreover, the number of possible segmentation targets (different anatomical structures, different types of lesions, etc.) are countless. It is impractical to cover every single unseen class by training a new, specific model.

As a potential solution to these two challenges, few-shot learning has been proposed [8–13]. During *inference*, a few-shot learning model distills a discriminative representation of an unseen class from only a few labeled examples (usually denoted as *support*) to make predictions for unlabeled examples (usually denoted as *query*) without the need for re-training the model. If applying few-shot learning to medical images, segmenting a rare or novel lesion can be potentially efficiently achieved using only a few labeled examples.

However, *training* an existing few-shot semantic segmentation (FSS) model for medical imaging has not had much success in the past, as most of FSS methods rely on a large training dataset with many annotated training classes to avoid overfitting [14–19,21–25]. In order to bypass this unmet need of annotation, we propose to train an FSS model on unlabeled images instead via self-supervised learning, an unsupervised technique that learns generalizable image representations by solving a carefully designed task [26–33]. Another challenge for a lot of state-of-the-art FSS network architectures is the loss of local information within a spatially variant class in their learned representations. This problem is in particular magnified in medical images since extreme foreground-background imbalance commonly exists in medical images. As shown in Fig. 1(b), the background class is large and spatially inhomogeneous whereas the foreground class (in purple) is small and homogeneous. Under this scenario, an ambiguity in prediction on foreground-background boundary might happen if the distinct appearance information of different local regions (or saying, parts) in the background is unreasonably averaged out. Unfortunately, this loss of intra-class local

information exists in a lot of recent works, where each class is spatially averaged into a 1-D representation prototype [16,18,19,34] or weight vectors of a linear classifier [17]. In adjust to this problem, we instead encourage the network to preserve intra-class local information, by extracting an ensemble of local representations for each class.

In order to break the deadlock of training data scarcity and to boost segmentation accuracy, we propose SSL-ALPNet, a self-supervised few-shot semantic segmentation framework for medical imaging. The proposed framework exploits *superpixel-based self-supervised learning* (SSL), using superpixels for eliminating the need for manual annotations, and an *adaptive local prototype pooling* enpowered prototypical network (ALPNet), improving segmentation accuracy by preserving local information in learned representations. As shown in Fig. 1(a), to ensure image representations learned through self-supervision are well-generalizable to real semantic classes, we generate pseudo-semantic labels using superpixels, which are compact building blocks for semantic objects [35–37]. In addition, to improve the discriminative ability of learned image representations, we formulate the self-supervision task as one-superpixel-against-the-rest segmentation. Moreover, to enforce invariance in representations between support and query, which is crucial for few-shot segmentation in real-world, we synthesis variants in shape and intensity by applying random geometric and intensity transforms between support and query. In our experiments, we observed that by purely training with SSL, our network outperforms those trained with manual annotated classes by considerable margins. Besides, as shown in Fig. 1, to boosts segmentation accuracy, we designed adaptive local prototype module (ALP) for preserving local information of each class in their prototypical representations. This is achieved by extracting an ensemble of local representation prototypes, each focuses on a different region. Of note, the number of prototypes are allocated adaptively by the network based on the spatial size of each class. By this mean, ALP alleviates ambiguity in segmentation caused by insufficient local information.

Overall, the proposed SSL-ALPNet framework has the following major advantages: Firstly, compared with current state-of-the-art few-shot segmentation methods which in general rely on a large number of annotated classes for training, the proposed method eliminates the need for annotated training data instead. By completely detaching representation extraction from manual labeling, the proposed method potentially expands the application of FSS in annotation-scarce medical images. In addition, unlike most of self-supervised learning methods for segmentation where fine-tuning on labeled data is still required before testing [27–30,32,38], the proposed method requires no fine-tuning after SSL. Moreover, compared to some of novel modules [39–41] used in FSS where slight performance gain are at the cost of heavy computations, the proposed ALP is simple and efficient in contrast to its significant performance boosting. No trainable parameters is contained in ALP.

Our contributions are summarized as follows:

- We propose SSL-ALPNet, the first work that explores self-supervised learning for few-shot medical image segmentation, to the best of our knowledge. It outperforms peer FSS methods, which usually require training with manual annotations, by merely training on unlabeled images.
- We propose adaptive local prototype pooling, a local representation computation module that significantly boosts performance of the state-of-the-art prototypical networks on medical images.
- We for the first time evaluated FSS on different imaging modalities, segmentation classes and with the presence of patient pathologies. The established evaluation strategy not only highlights wide applicability of our work, but also facilitates future works that seek to evaluate FSS in a more realistic scenario.

2 Related Work

2.1 Few-Shot Semantic Segmentation

Recent work by [31] firstly introduces self-supervised learning into few-shot image classification. However, few-shot segmentation is often more challenging: dense prediction needs to be performed at a pixel level. To fully exploit information in limited support data, most of popular FSS methods directly inject support to the network as guiding signals [20,21,42,43], or construct discriminative representations from support as reference to segment query [15–19]. The pioneering work [15] learns to generate classifier weights from support; [17] extends weights generation to multi-scale. [14] instead directly use support to condition segmentation on query by fusing their feature maps. Exploiting network components such as attention modules [39,44] and graph networks [40,41], recent works boost segmentation accuracy [21] and enable FSS with coarse-level supervisions [22,24,42]. Exploiting learning-based optimization, [23,25] combine meta-learning with FSS. However, almost all of these methods assume abundant annotated (including weakly annotated) training data to be available, making them difficult to translate to segmentation scenarios in medical imaging.

One main stream of FSS called *prototypical networks* focuses on exploiting *representation prototypes* of semantic classes extracted from the support. These prototypes are utilized to make similarity-based prediction [8,18,34] on query, or to tune representations of query [16]. Recently, prototypical alignment network (PANet) [18] has achieved state-of-the-art performance on natural images. This is achieved simply with a generic convolutional network and an alignment regularization. However, these works aim to improve performance on training-classes-abundant natural images. Their methodologies focus on network design. Our work, by contrast, focuses on utilizing unlabeled medical image for training by exploiting innovative training strategies and pesudolabels. Nevertheless, since PANet is one of state-of-the-art and is conceptually simple, we take this method

as our baseline to highlight our self-supervised learning as a generic training strategy.

In medical imaging, most of recent works on few-shot segmentation only focus on training with less data [45–49]. These methods usually still require retraining before applying to unseen classes, and therefore they are out-of-scope in our discussion. Without retraining on unseen classes, the SE-Net [43] introduces squeeze and excite blocks [50] to [14]. To the best of our knowledge, it is the first FSS model specially designed for medical images, with which we compared our method in experiments.

2.2 Self-supervised Learning in Semantic Segmentation

A series of self-supervision tasks have been proposed for semantic segmentation. Most of these works focus on intuitive handcrafted supervision tasks including spatial transform prediction [51], image impainting [32], patch reordering [27], image colorization [33], difference detection [52], motion interpolation [53] and so on. Similar methods have been applied to medical images [38,54–56]. However, most of these works still require a second-stage fine-tuning after initializing with weights learned from self-supervision. In addition, features learned from handcrafted tasks may not be sufficiently generalizable to semantic segmentation, as two tasks might not be strongly related [57]. In contrast, in our work, segmenting superpixel-based pseudolabels is directly related to segmenting real objects. This is because superpixels are compact building blocks for semantic masks for real objects. Recent works [48,58,59] on medical imaging rely on second-order optimization [60]. These works differ from our work in key method and task.

Our proposed SSL technique shares a similar spirit as [61] (or arguably, as some recent works on contrastive learning [62–65]) in methodology. Both methods encourage invariance in image representation by intentionally creating variants. While [61] focuses on visual information clustering, we focus on the practical but challenging few-shot medical image segmentation problem.

2.3 Superpixel Segmentation

Superpixels are small, compact image segments which are usually piece-wise smooth [35,66]. Superpixels are generated by clustering local pixels using statistical models with respect to low-level image features. These models include Gaussian mixture [37] and graph cut [67]. In this work, we employed off-the-shelf, efficient and unsupervised graph-cut-based algorithm by [68]. Compared with the popular SLIC method [37], superpixels generated by [68] are more diverse in shape. Training with these superpixels intuitively improves generalizability of the network to unseen classes in various shapes.

3 Method

We first introduce problem formulation for few-shot semantic segentation (FSS). Then, the ALPNet architecture is introduced with a focus on adaptive local

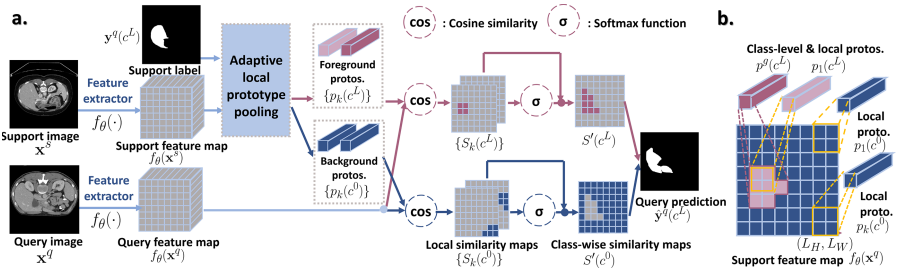

Fig. 2. (a). Workflow of the proposed network: The feature extractor $f_\theta(\cdot)$ takes the support image and query image as input to generate feature maps $f_\theta(\mathbf{x}^s)$ for support and $f_\theta(\mathbf{x}^q)$ for query. The proposed adaptive local prototype pooling module then takes support feature map and support label as input to obtain an ensemble of representation prototypes $p_k(c^j)$'s. These prototypes are used as references for comparing with query feature map $f_\theta(\mathbf{x}^q)$. Similarity maps generated by these comparisons are fused together to form the final segmentation. This figure illustrates a 1-way segmentation setting, where c^L is the foreground class, c^0 is the background. (b). Illustration of the adaptive local prototype pooling module: Local prototypes are calculated by spatially averaging support feature maps within pooling windows (orange boxes); class-level prototypes are averaged under the entire support label (purple region). (Color figure online)

prototype pooling and the corresponding inference process. We highlight our superpixel-based self-supervised learning (SSL) with details in pseudolabel generation process and episode formation in Sect. 3.3. Finally, we introduce the overall end-to-end training objective under the proposed SSL technique. Of note, after the proposed self-supervised learning, ALPNet can be directly applied to unseen classes with its weights fixed, and with reference to a few human-labeled support slices. There is no fine-tuning required in this testing phase.

3.1 Problem Formulation

The aim of few-shot segmentation is to obtain a model that can segment an unseen semantic class, by just learning from a few labeled images of this unseen class during inference without retraining the model. In few-shot segmentation, a training set $\mathcal{D}_{tr}$ containing images with training semantic classes $\mathcal{C}_{tr}$ (e.g., $\mathcal{C}_{tr} = \{liver, spleen, spine\}$), and a testing set $\mathcal{D}_{te}$ of images containing testing unseen classes $\mathcal{C}_{te}$ (e.g., $\mathcal{C}_{te} = \{heart, kidney\}$), are given, where $\mathcal{C}_{tr} \cap \mathcal{C}_{te} = \emptyset$. The task is to train a segmentation model on $\mathcal{D}_{tr}$ (e.g. labeled images of livers, spleens and spines) that can segment semantic classes $\mathcal{C}_{te}$ in images in $\mathcal{D}_{te}$, given a few annotated examples of $\mathcal{C}_{te}$ (e.g. to segment *kidney* with reference to a few labeled images of kidney), without re-training. $\mathcal{D}_{tr} = \{(\mathbf{x}, \mathbf{y}(c^{\hat{j}}))\}$ is composed of images $\mathbf{x} \in \mathcal{X}$ and corresponding binary masks $\mathbf{y}(c^{\hat{j}})$'s $\in \mathcal{Y}$ of classes $c^{\hat{j}} \in \mathcal{C}_{tr}$, where $\hat{j} = 1, 2, 3, ..., N$ is the class index. $\mathcal{D}_{te}$ is defined in the same way but for testing images and masks with $\mathcal{C}_{te}$. In each inference pass, a *support* set $\mathcal{S}$ and a *query* set $\mathcal{Q}$ are given. The support $\mathcal{S} = \{(\mathbf{x}_l^s, \mathbf{y}_l^s(c^{\hat{j}}))\}$ contains images $\mathbf{x}_l^s$

and masks $\mathbf{y}_l^s(c^{\hat{j}})$, and it serves as examples for segmenting $c^{\hat{j}}$'s; the query set $\mathcal{Q} = \{\mathbf{x}^q\}$ contains images $\mathbf{x}^q$'s to be segmented. Here, the superscripts denote an image or mask is from support (s) or query (q). And $l = 1, 2, 3, ..., K$ is the index for each image-mask pair of class $c^{\hat{j}}$. One support-query pair $(\mathcal{S}, \mathcal{Q})$ comprises an *episode*. Every episode defines an N-way K-shot segmentation sub-problem if there are N classes (also called N tasks) to be segmented and K labeled images in $\mathcal{S}$ for each class. Note that the background class is denoted as c^0 and it does not count toward $\mathcal{C}_{tr}$ or $\mathcal{C}_{te}$.

3.2 Network Architecture

Overview: Our network is composed of: (a) a generic *feature extractor* network $f_\theta(\cdot) : \mathcal{X} \rightarrow \mathcal{E}$ parameterized by θ, where $\mathcal{E}$ is the representation space (i.e. feature space) on which segmentation operates; (b) the proposed *adaptive local prototype pooling module* (ALP) $g(\cdot, \cdot) : \mathcal{E} \times \mathcal{Y} \rightarrow \mathcal{E}$ for extracting representation *prototypes* from support features and labels; (c) and a *similarity based classifier* $sim(\cdot, \cdot) : \mathcal{E} \times \mathcal{E} \rightarrow \mathcal{Y}$ for segmention by comparing prototypes and query features.

As shown in Fig. 2, in inference, the feature extractor network $f_\theta(\cdot)$ provides ALP with feature maps by mapping both $\mathbf{x}_l^s$'s and $\mathbf{x}^q$'s to feature space $\mathcal{E}$, producing feature maps $\{(f_\theta(\mathbf{x}^q), f_\theta(\mathbf{x}_l^s))\} \in \mathcal{E}$. ALP takes each $(f_\theta(\mathbf{x}_l^s), \mathbf{y}_l^s(c^{\hat{j}}))$ pair as input to compute both *local prototypes* and *class-level prototypes* of semantic class $c^{\hat{j}}$ and background c^0. These prototypes will later be used as references of each class for segmenting query images. Prototypes of all c^j's forms a prototype ensemble $\mathcal{P} = \{p_k(c^j)\}, j = 0, 1, 2, ..., N$ where k is prototype index and $k \geq 1$ for each c^j. This prototype ensemble is used by the classifier $sim(\cdot, \cdot)$ to predict the segmentation for the query image, saying $\hat{\mathbf{y}}^q = sim(\mathcal{P}, f_\theta(\mathbf{x}^q))$. This is achieved by first measuring similarities between each $p_k(c^j)$'s and query feature map $f_\theta(\mathbf{x}^q)$, and then fusing these similarities together.

Adaptive Local Prototype Pooling: In contrast to previous works [17,18,34], where intra-class local information is unreasonably spatially averaged out underneath the semantic mask, we propose to preserve local information in prototypes by introducing adaptive local prototype pooling module (ALP). In ALP, each *local prototype* is only computed within a *local pooling window* overlaid on the support and only represents one part of object-of-interest.

Specifically, we perform average pooling with a pooling window size (L_H, L_W) on each $f_\theta(\mathbf{x}_l^s) \in \mathbb{R}^{D \times H \times W}$ where (H, W) is the spatial size and D is the channel depth. Of note, (L_H, L_W) determines the spatial extent under which each local prototype is calculated in the representation space $\mathcal{E}$. The obtained local prototype $p_{l,mn}(c)$ with undecided class c at spatial location (m, n) of the average-pooled feature map is given by

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

$$\text{where } mL_H \leq h < (m+1)L_H, \; nL_W \leq w < (n+1)L_W.$$

To decide the class c of each $p_{l,mn}(c)$, we average-pool the binary mask $\mathbf{y}_l^s(c^{\hat{j}})$ of the foreground class $c^{\hat{j}}$ to the same size $(\frac{H}{L_H}, \frac{W}{L_W})$. Let $y^a_{l,mn}$ be the value of $\mathbf{y}_l^s(c^{\hat{j}})$ after average pooling at location (m, n), c is assigned as:

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

T is the lower-bound threshold for foreground which is empirically set to 0.95.

To ensure at least one prototype is generated for objects smaller than the pooling window (L_H, L_W), we also compute a *class-level prototype* $p_l^g(c^{\hat{j}})$ using masked average pooling [18,19]:

$$p_l^g(c^{\hat{j}}) = \frac{\sum\limits_{h,w} \mathbf{y}_l^s(c^{\hat{j}})(h, w) f_\theta(\mathbf{x}_l^s)(h, w)}{\sum\limits_{h,w} \mathbf{y}_l^s(c^{\hat{j}})(h, w)}. \tag{3}$$

In the end, $p_{l,mn}(c^j)$'s and $p_l^g(c^{\hat{j}})$'s are re-indexed with subscript k's for convenience, and hence comprise the representation prototype ensemble $\mathcal{P} = \{p_k(c^j)\}$. This ensemble therefore preserves more intra-class local distinctions by explicitly representing different local regions into separate prototypes.

Similarity-Based Segmentation: The similarity-based classifier $sim(\cdot, \cdot)$ is designed to make dense prediction on query by exploiting local image information in $\mathcal{P}$. This is achieved by firstly matching each prototype to a corresponding local region in query, and then fusing the local similarities together.

As a loose interpretation, to segment a large *liver* in query, in the first stage, a local prototype $p_k(c^L)$ with class c^L = *liver*, whose pooling window falls over the *right lobe* of the *liver* particularly finds a similar region which looks like a *right lobe* in query (instead of matching the entire *liver*). Then, to get an entire liver, results from *right lobe*, and *left lobe* are fused together to form a *liver*.

Specifically, $sim(\cdot, \cdot)$ first takes query feature map $f_\theta(\mathbf{x}^q)$ and prototype ensemble $\mathcal{P} = \{p_k(c^j)\}$ as input to compute *local similarity maps* $S_k(c^j)$'s between $f_\theta(\mathbf{x}^q)$ and all $p_k(c^j)$'s respectively. Each entry $S_k(c^j)(h, w)$ at spatial location (h, w) corresponding to $f_\theta(\mathbf{x}^q)$ is given by

$$S_k(c^j)(h, w) = \alpha p_k(c^j) \odot f_\theta(\mathbf{x}^q)(h, w), \tag{4}$$

where $\odot$ denotes cosine similarly, which is bounded, same as in [18]: $a \odot b = \frac{\langle a,b \rangle}{\|a\|_2 \|b\|_2}$, $a, b \in \mathbb{R}^{D \times 1 \times 1}$, α is a multiplier, which helps gradients to back-propagate in training [69]. In our experiments, α is set to 20, same as in [18].

Then, to obtain similarity maps (unnormalized) with respect to each class c^j as a whole, local similarity maps $S_k(c^j)$'s are fused for each class separately into *class-wise similarities* $S'(c^j)$, this is done through a softmax function:

$$S'(c^j)(h, w) = \sum_k S_k(c^j)(h, w) \operatorname*{softmax}_k [S_k(c^j)(h, w)]. \tag{5}$$

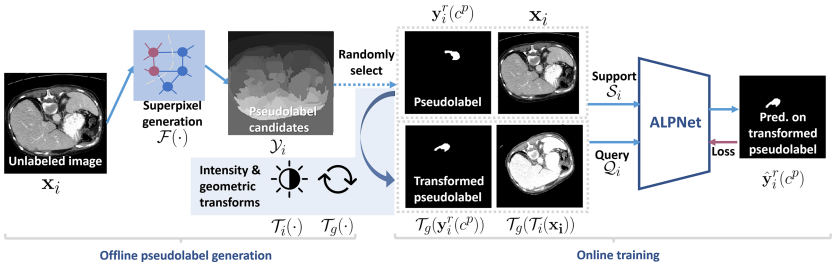

Fig. 3. Workflow of the proposed superpixel-based self-supervised learning technique.

$\underset{k}{\text{softmax}}[S_k(c^j)(h,w)]$ refers to the operation of first stacking all $S_k(c^j)(h,w)$'s along channel dimension and then computing softmax function along channels.

To obtain the final dense prediction, in the end, class-wise similarities are normalized into probabilities:

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

3.3 Superpixel-Based Self-supervised Learning

To obtain accurate and robust results, two properties are highly desirable for similarity-based classifiers. For each class, the representations should be *clustered* in order to be discriminative under a similarity metric; meanwhile, these representations should be *invariant* across images (in our case any combinations of support and query) to ensure robustness in prediction [61].

These two properties are encouraged by the proposed superpixel-based self-supervised learning (SSL). As annotations for real semantic classes are unavailable, SSL exploits pseudolabels to enforce *clustering* at a superpixel-level. This is naturally achieved by back-propagating segmentation loss via cosine-similarity-based classifier. Here, the superpixel-level clustering property can be transferred to real semantic classes, since one semantic mask is usually composed of several superpixels [35,36]. Additionally, to encourage representations to be invariant against shape and intensity differences between images, we perform geometric and intensity transforms between support and query. This is because shape and intensity are the largest sources of variations in medical images [70].

The proposed SSL framework consists of two phases: offline pseudolabel generation and online training. The entire workflow can be seen in Fig. 3.

Unspervised Pseudolabel Generation: To obtain candidates for pseudolabels, a collection of superpixels $\mathcal{Y}_i = \mathcal{F}(\mathbf{x}_i)$ are generated for every image $\mathbf{x}_i$. This is efficiently done with the unsupervised algorithm [68] denoted by $\mathcal{F}(\cdot)$.

Online Episode Composition: For each episode i, an image $\mathbf{x}_i$ and a randomly chosen superpixel $\mathbf{y}_i^r(c^p) \in \mathcal{Y}_i$ form the support $\mathcal{S}_i = \{(\mathbf{x}_i, \mathbf{y}_i^r(c^p))\}$. Here $\mathbf{y}_i^r(c^p)$ is a binary mask with index $r = 1, 2, 3, ..., |\mathcal{Y}_i|$ and c^p denotes the *pseudolabel*

class (corresponding background mask $\mathbf{y}_i^r(c^0)$ is given by $1-\mathbf{y}_i^r(c^p)$). Meanwhile, the query set $\mathcal{Q}_i = \{(\mathcal{T}_g(\mathcal{T}_i(\mathbf{x}_i)))\}, \mathcal{T}_g(\mathbf{y}_i^r(c^p)))$ is constructed by applying random geometric and intensity transforms: $\mathcal{T}_g(\cdot)$ and $\mathcal{T}_i(\cdot)$ to the support. By this mean, each $(\mathcal{S}_i, \mathcal{Q}_i)$ forms a 1-way 1-shot segmentation problem. In practice, $\mathcal{T}_g(\cdot)$ includes affine and elastic transforms, $\mathcal{T}_i(\cdot)$ is gamma transform.

End-to-End Training: The network is trained end-to-end, where each iteration i takes an episode $(\mathcal{S}_i, \mathcal{Q}_i)$ as input. Cross entropy loss is employed where the segmentation loss $\mathcal{L}_{seg}^i$ for each iteration is written as:

$$\mathcal{L}_{seg}^i(\theta; \mathcal{S}_i, \mathcal{Q}_i) = -\frac{1}{HW}\sum_h^H\sum_w^W\sum_{j\in\{0,p\}} \mathcal{T}_g(\mathbf{y}_i^r(c^j))(h,w)\log(\hat{\mathbf{y}}_i^r(c^j)(h,w)), \quad (7)$$

where $\hat{\mathbf{y}}_i^r(c^p)$ is the prediction of query pseudolabel $\mathcal{T}_g(\mathbf{y}_i^r(c^p))$ and is obtained as described in Sect. 3.2. In practice, weightings of 0.05 and 1.0 are given to c^0 and c^p separately for mitigating class imbalance. We also inherited the *prototypical alignment regularization* in [18]: taking prediction as support, i.e. $\mathcal{S}' = (\mathcal{T}_g(\mathcal{T}_i(\mathbf{x}_i)), \hat{\mathbf{y}}_i^r(c^p))$, it should correctly segment the original support image $\mathbf{x}_i$. This is presented as

$$\mathcal{L}_{reg}^i(\theta; \mathcal{S}_i', \mathcal{S}_i) = -\frac{1}{HW}\sum_h^H\sum_w^W\sum_{j\in\{0,p\}} \mathbf{y}_i^r(c^j)(h,w)\log(\bar{\mathbf{y}}_i^r(c^j)(h,w)), \quad (8)$$

where $\bar{\mathbf{y}}_i^r(c^p)$ is the prediction of $\mathbf{y}_i^r(c^p)$ taking $\mathbf{x}_i$ as query.

Overall, the loss function for each epioside is:

$$\mathcal{L}^i(\theta; \mathcal{S}_i, \mathcal{Q}_i) = \mathcal{L}_{seg}^i + \lambda\mathcal{L}_{reg}^i, \quad (9)$$

where λ controls strength of regularization as in [18].

After self-supervised learning, the network can be directly used for inference on unseen classes.

4 Experiments

Datasets: To demonstrate the general applicability of our proposed method under different imaging modalities, segmentation classes and health conditions of the subject, we performed evaluations under three scenarios: abdominal organs segmentation for CT and MRI (Abd-CT and Abd-MRI) and cardiac segmentation for MRI (Card-MRI). All three datasets contain rich information outside their regions-of-interests, which benefits SSL by providing sources of superpixels. Specifically,

- **Abd-CT** is from MICCAI 2015 Multi-Atlas Abdomen Labeling challenge [71]. It contains 30 3D abdominal CT scans. Of note, this is a clinical dataset containing patients with various pathologies and variations in intensity distributions between scans.

- **Abd-MRI** is from ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge (Task 5) [72]. It contains 20 3D T2-SPIR MRI scans.
- **Card-MRI** is from MICCAI 2019 Multi-sequence Cardiac MRI Segmentation Challenge (bSSFP fold) [73], with 35 clinical 3D cardiac MRI scans.

To unify experiment settings, all images are re-formated as 2D axial (Abd-CT and Abd-MRI) or 2D short-axis (Card-MRI) slices, and resized to 256×256 pixels. Prepossessings are applied following common practices. Each 2D slice is repeated for three times in channel dimension to fit into the network.

To comparatively evaluate the results on classes with various shapes, locations and textures between partically-pathologic, imhomogeneous Abd-CT and all-healthy, homogeneous Abd-MRI, we construct a shared label set containing left kidney, right kidney, spleen and liver; For Card-MRI, the label set contains left-ventricle blood pool (LV-BP), left-ventricle myocardium (LV-MYO) and right-ventricle (RV). In all experiments, we perform five-fold cross-validation.

Evaluation: To measure the overlapping between prediction and ground truth, we employ Dice score (0–100, 0: mismatch; 100: perfectly match), which is commonly used in medical image segmentation researches. To evaluate 2D segmentation on 3D volumetric images, we follow the evaluation protocol established by [43]. In a 3D image, for each class $c^{\hat{j}}$, images between the top slice and the bottom slice containing $c^{\hat{j}}$ are divided into C equally-spaced chunks. The middle slice in each chunk from the support scan is used as reference for segmenting all the slices in corresponding chunk in query. In our experiments C is set to be 3. Of note, the support and query scans are from different patients.

To evaluate generalization ability to unseen testing classes, beyond the standard few-shot segmentation experiment setting for medical images established by [43] (**setting 1**), where testing class might appear as background in training data, we introduce a **setting 2**. In setting 2, we force testing classes (even unlabeled) to be completely unseen by removing any image that contains a testing class, from the training dataset.

Labels are therefore partitioned differently according to the settings and types of supervision. In setting 1, when training with SSL, no label partitioning is required for training. When training with annotated images, each time we take one class for testing and the rest for training. To observe if the learned representations encode spatial concepts like left and right, we deliberately group ⟨left/ right kidney⟩ to appear together in training or testing. In setting 2, as ⟨spleen, liver⟩, or ⟨left/ right kidney⟩ usually appear together in a 2D slice respectively, we group them into *upper* abdomen and *lower* abdomen groups separately. In each experiment all slices containing the testing group will be removed from training data. For Card-MRI, only setting 1 is examined as all the labels usually appear together in 2D slices, making label exclusion impossible.

To simulate the scarcity of labeled data in clinical practice, all our experiments in this section are performed under 1-way 1-shot setting.

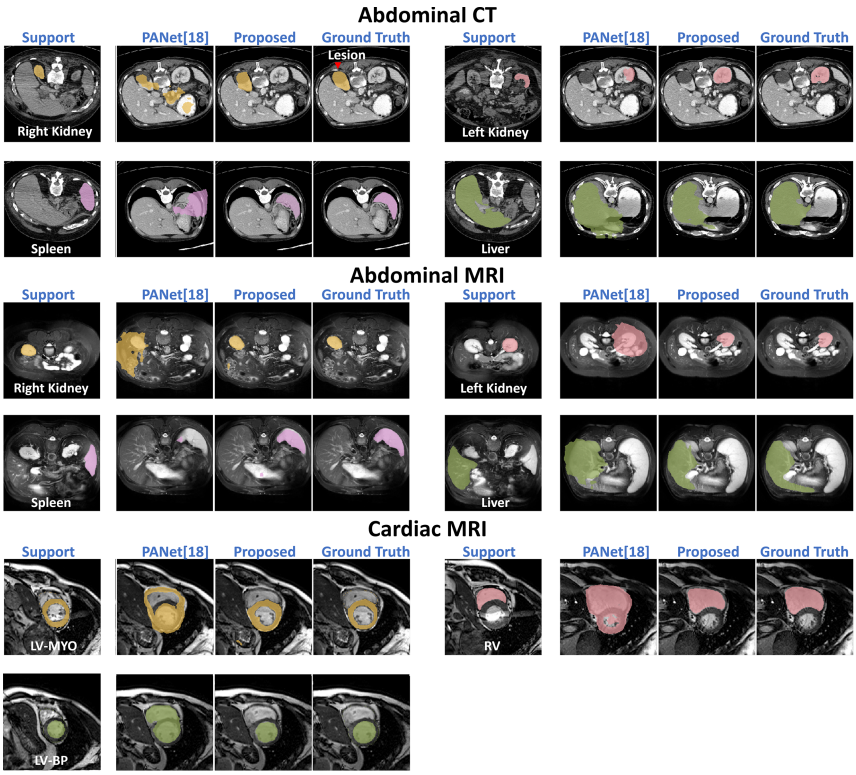

Fig. 4. Qualitative results of our method on all three combinations of imaging modalities and segmentation tasks. The proposed method achieves desirable segmentation results which are close to ground truth. To highlight the strong generalization ability, examples of results from the proposed method on Abd-CT and Abd-MRI are from setting 2, where images containing testing classes are strictly excluded in training set even though they are unlabeled. See supplemental materials for more examples.

Implementation Details: The network is implemented with PyTorch based on official PANet implementation[1] [18]. To obtain high spatial resolutions in feature maps, $f_\theta(\cdot)$ is configured as an off-the-shelf fully-convolutional ResNet101, which is pre-trained on part of MS-COCO for higher segmentation performance [18,74] (same for vanilla PANet in our experiments). It takes a $3 \times 256 \times 256$ image as input and produces a $256 \times 32 \times 32$ feature map. Local pooling window (L_H, L_W) for prototypes is set to 4×4 for training and 2×2 for inference on feature maps. The loss in Eq. 9 is minimized for 100k iterations using stochastic gradient descent with a batch size of 1. The learning rate is 0.001 with a stepping decay rate of 0.98 per 1000 iterations. The self-supervised training takes ∼3h on a single Nvidia RTX 2080Ti GPU, consuming 2.8 GBs of memory.

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

Table 1. Experiment results (in Dice score) on abdominal images under setting 2.

Method	Manual Anno.?	Abdominal-CT					Abdominal-MRI				
		Lower		Upper		Mean	Lower		Upper		Mean
		LK	RK	Spleen	Liver		LK	RK	Spleen	Liver	
SE-Net [43]	✓	32.83	14.34	0.23	0.27	11.91	62.11	61.32	51.80	27.43	50.66
Vanilla PANet [18]	✓	32.34	17.37	29.59	38.42	29.43	53.45	38.64	50.90	42.26	46.33
ALPNet-init	-	13.90	11.61	16.39	41.71	20.90	19.28	14.93	23.76	37.73	23.93
ALPNet	✓	34.96	30.40	27.73	47.37	35.11	53.21	58.99	52.18	37.32	50.43
SSL-PANet	×	37.58	34.69	43.73	61.71	44.42	47.71	47.95	58.73	64.99	54.85
SSL-ALPNet	×	**63.34**	**54.82**	**60.25**	**73.65**	**63.02**	**73.63**	**78.39**	**67.02**	**73.05**	**73.02**

Table 2. Experiment results (in Dice score) on abdominal images under setting 1.

Method	Manual Anno.?	Abdominal-CT					Abdominal-MRI				
		Kidneys		Spleen	Liver	Mean	Kidneys		Spleen	Liver	Mean
		LK	RK				LK	RK			
SE-Net [43]	✓	24.42	12.51	43.66	35.42	29.00	45.78	47.96	47.30	29.02	42.51
Vanilla PANet [18]	✓	20.67	21.19	36.04	49.55	31.86	30.99	32.19	40.58	50.40	38.53
ALPNet	✓	29.12	31.32	41.00	65.07	41.63	44.73	48.42	49.61	62.35	51.28
SSL-PANet	×	56.52	50.42	55.72	60.86	57.88	58.83	60.81	61.32	71.73	63.17
SSL-ALPNet	×	**72.36**	**71.81**	**70.96**	**78.29**	**73.35**	**81.92**	**85.18**	**72.18**	**76.10**	**78.84**
Zhou et al. [75]	Ful. Sup	95.3	92.0	96.8	97.4	95.4	-				
Isenseen et al. [76]	Ful. Sup	-					-				94.6

4.1 Quantitative and Qualitative Results

Comparison with State-of-the-Art Methods: Tables 1, 2 and 3 show the comparisons of our method with vanilla PANet, one of state-of-the-art methods on natural images and SE-Net[2] [43], the lastest FSS method for medical images. Without using any manual annotation, our proposed SSL-ALPNet consistently outperforms them by an average Dice score of >25. As shown in Fig. 4, the proposed framework yields satisfying results on organs with various shapes, sizes and intensities. Of note, for all evaluated methods, results on Abd-MRI are in general higher than those on Abd-CT. This not surprising as Abd-MRI is more homogeneous, and most of organs in Abd-MRI have distinct contrast to surrounding tissues, which helps to reduce ambiguity at boundaries.

Importantly, Table 1 demonstrates the strong generalization ability of our method to unseen classes. This implies that the proposed superpixel-based self-supervised learning has successfully trained the network to learn more diverse and generalizable image representations from unlabeled images.

The upperbounds obtained by fully-supervised learning on all labeled images are shown in Table 2 for reference.

Performance Boosts by ALP and SSL: The separate performance gains obtained by introducing adaptive local prototype pooling or self-supervised learning can be observed in rows *ALPNet* and *SSL-PANet* in Tables 1, 2 and

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

Table 3. Experiment results (in Dice score) on cardiac images under setting 1.

Method	Manual anno.?	LV-BP	LV-MYO	RV	Mean
SE-Net [43]	✓	58.04	25.18	12.86	32.03
Vanilla PANet [18]	✓	53.64	35.72	39.52	42.96
ALPNet	✓	73.08	49.53	58.50	60.34
SSL-PANet	×	70.43	46.79	69.52	62.25
SSL-ALPNet	×	**83.99**	**66.74**	**79.96**	**76.90**

Table 4. Ablation study on types transformations.

Int	Geo	LK	RK	Spleen	Liver	Mean
×	×	45.49	48.40	53.05	73.60	55.13
✓	×	55.56	49.12	59.20	73.39	59.31
×	✓	59.32	51.45	57.74	**78.93**	61.86
✓	✓	**63.34**	**54.82**	**60.25**	73.65	**63.02**

Table 5. Ablation study on minimum pseudolabel sizes.

Min. size (px)	LK	RK	Spleen	Liver	Mean
100	52.92	47.45	53.16	68.40	55.49
400	**63.34**	**54.82**	**60.25**	73.65	**63.02**
1600	51.74	44.83	56.99	**74.73**	57.08
Avg. Size in 2D (px)	798	799	1602	5061	

3. These results suggests both SSL and ALP contribute greatly. The performance gains of SSL highlight the benefit of a well-designed training strategy that encourages learning generalizable features, which is usually overlooked in recent few-shot segmentation methods. More importantly, the synergy between them (*SSL-ALPNet*) leads to significant performance gains by learning richer image representations and by constructing more effective inductive bias. To be assured that MS-COCO initialization alone cannot do FSS, we also include the results when the ALPNet is directly tested after initialization, shown in *ALPNet-init* in Table 1.

Robustness Under Patient Pathology: As shown in Fig. 4, despite the large dark lesion on right-kidney in Abd-CT, the proposed method stably produces satisfying results.

4.2 Ablation Studies

Ablation studies are performed on Abd-CT under setting 2. This scenario is challenging but close to clinical scenario in practice.

Importance of Transforms Between the Support and Query: To demonstrate the importance of geometric and intensity transformations in our method, we performed ablation studies as shown in Table 4. Unsurprisingly, the highest and lowest overall results are obtained by applying both or no transforms, proving the effectiveness of introducing random transforms. Interestingly, applying intensity transform even hurts performance on liver. This implies that the configuration of intensity transforms in our experiments may deviate from the actual intensity distribution of livers in the dataset.

Effect of Pseudolabel Sizes: To investigate the effect of pseudolabel sizes on performance, we experimented with pseudolabel sets with different minimum superpixel sizes. Table 5 shows that the granularity of superpixels should be reasonably smaller than sizes of actual semantic labels. This implies that too-coarse or too-fine-grained pseudolabels might divert the granularity of clusters in the learned representation space from that of real semantic classes.

5 Conclusion

In this work, we propose a self-supervised few-shot segmentation framework for medical imaging. The proposed method successfully outperforms state-of-the-art methods without requiring any manual labeling for training. In addition, it demonstrates strong generalization to unseen semantic classes in our experiments. Moreover, the proposed superpixel-based self-supervision technique provides an effective way for image representation learning, opening up new possibilities for future works in semi-supervised and unsupervised image segmentation.

Acknowledgements. This work is supported by the EPSRC Programme Grant EP/P001009/1. This work is also supported by the UK Research and Innovation London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare. The authors would like to thank Konstantinos Kamnitsas and Zeju Li for insightful comments.

References

1. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation. Annu. Rev. Biomed. Eng. **2**(1), 315–337 (2000)
2. Sharma, N., Aggarwal, L.M.: Automated medical image segmentation techniques. J. Med. Phys. Assoc. Med. Phys. India **35**(1), 3 (2010)
3. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Alzheimer's Disease Neuroimaging, et al.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage **55**(3), 856–867 (2011)
4. Ei Naqa, I., et al.: Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning a. Med. Phys. **34**(12), 4738–4749 (2007)
5. Zaidi, H., El Naqa, I.: Pet-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur. J. Nucl. Med. Mol. Imaging **37**(11), 2165–2187 (2010)
6. De Leeuw, F., et al.: Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the rotterdam scan study. J. Neurol. Neurosurg. Psychiatry **70**(1), 9–14 (2001)
7. Petersen, S.E., et al.: Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK biobank-rationale, challenges and approaches. J. Cardiovasc. Magn. Reson. **15**(1), 46 (2013)
8. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, pp. 4077–4087 (2017)
9. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1199–1208 (2018)

10. Garcia, V., Bruna, J.: Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043 (2017)
11. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in Neural Information Processing Systems, pp. 3630–3638 (2016)
12. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 594–611 (2006)
13. Lake, B., Salakhutdinov, R., Gross, J., Tenenbaum, J.: One shot learning of simple visual concepts. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 33 (2011)
14. Rakelly, K., Shelhamer, E., Darrell, T., Efros, A., Levine, S.: Conditional networks for few-shot semantic segmentation. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May, 2018, Workshop Track Proceedings (2018)
15. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. arXiv preprint arXiv:1709.03410 (2017)
16. Dong, N., Xing, E.: Few-shot semantic segmentation with prototype learning. In: British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, 3–6 September 2018, vol. 3, p. 79(2018)
17. Siam, M., Oreshkin, B.N., Jagersand, M.: AMP: adaptive masked proxies for few-shot segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5249–5258 (2019)
18. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9197–9206 (2019)
19. Zhang, X., Wei, Y., Yang, Y., Huang, T.: Sg-one: Similarity guidance network for one-shot semantic segmentation. arXiv preprint arXiv:1810.09091 (2018)
20. Rakelly, K., Shelhamer, E., Darrell, T., Efros, A.A., Levine, S.: Few-shot segmentation propagation with guided networks. arXiv preprint arXiv:1806.07373 (2018)
21. Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q., Yao, R.: Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9587–9595 (2019)
22. Siam, M., Doraiswamy, N., Oreshkin, B.N., Yao, H., Jagersand, M.: Weakly supervised few-shot object segmentation using co-attention with visual and semantic inputs. arXiv preprint arXiv:2001.09540 (2020)
23. Tian, P., Wu, Z., Qi, L., Wang, L., Shi, Y., Gao, Y.: Differentiable meta-learning model for few-shot semantic segmentation. arXiv preprint arXiv:1911.10371 (2019)
24. Hu, T., Mettes, P., Huang, J.H., Snoek, C.G.: Silco: show a few images, localize the common object. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5067–5076 (2019)
25. Hendryx, S.M., Leach, A.B., Hein, P.D., Morrison, C.T.: Meta-learning initializations for image segmentation. arXiv preprint arXiv:1912.06290 (2019)
26. Lieb, D., Lookingbill, A., Thrun, S.: Adaptive road following using self-supervised learning and reverse optical flow. In: Robotics: Science and Systems, pp. 273–280 (2005)
27. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1422–1430 (2015)

28. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 766–774 (2014)
29. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 577–593. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_35
30. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
31. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8059–8068 (2019)
32. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
33. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
34. Liu, J., Qin, Y.: Prototype refinement network for few-shot segmentation. arXiv preprint arXiv:2002.03579 (2020)
35. Ren, X., Malik, J.: Learning a classification model for segmentation. In: null, p. 10. IEEE (2003)
36. Stutz, D., Hermans, A., Leibe, B.: Superpixels: an evaluation of the state-of-the-art. Comput. Vis. Image Underst. **166**, 1–27 (2018)
37. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels. Technical report, EPFL (2010)
38. Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J.: Models genesis: generic autodidactic models for 3D medical image analysis. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 384–393. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_42
39. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
40. Schlichtkrull, M., et al.: Modeling relational data with graph convolutional networks. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 593–607. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_38
41. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
42. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: 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, pp. 5217–5226 (2019)
43. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: 'squeeze & excite' guided few-shot segmentation of volumetric images. Med. Image Anal. **59**, 101587 (2020)
44. Hu, T., Yang, P., Zhang, C., Yu, G., Mu, Y., Snoek, C.G.: Attention-based multi-context guiding for few-shot semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8441–8448 (2019)

45. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transforms for one-shot medical image segmentation. In: CVPR (2019)
46. Mondal, A.K., Dolz, J., Desrosiers, C.: Few-shot 3D multi-modal medical image segmentation using generative adversarial learning. arXiv preprint arXiv:1810.12241 (2018)
47. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D.: Data efficient unsupervised domain adaptation for cross-modality image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 669–677. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_74
48. Yu, H., et al.: Foal: fast online adaptive learning for cardiac motion estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4313–4323(2020)
49. Chen, C., et al.: Realistic adversarial data augmentation for MR image segmentation. arXiv preprint arXiv:2006.13322 (2020)
50. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
51. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2051–2060 (2017)
52. Shimoda, W., Yanai, K.: Self-supervised difference detection for weakly-supervised semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5208–5217 (2019)
53. Zhan, X., Pan, X., Liu, Z., Lin, D., Loy, C.C.: Self-supervised learning via conditional motion propagation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1881–1889 (2019)
54. Jamaludin, A., Kadir, T., Zisserman, A.: Self-supervised learning for Spinal MRIs. DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 294–302. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_34
55. Bai, W., et al.: Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 541–549. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_60
56. Chen, L., et al.: Self-supervised learning for medical image analysis using image context restoration. Med. Image Anal. **58**, 101539 (2019)
57. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: disentangling task transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3712–3722 (2018)
58. Dou, Q., de Castro, D.C., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. In: Advances in Neural Information Processing Systems, pp. 6447–6458 (2019)
59. Wu, Y., Rosca, M., Lillicrap, T.: Deep compressed sensing. arXiv preprint arXiv:1905.06723 (2019)
60. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400 (2017)
61. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9865–9874 (2019)

62. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
63. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: Advances in Neural Information Processing Systems, pp. 15535–15545 (2019)
64. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
65. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
66. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Commun. Pure Appl. Math. **42**(5), 577–685 (1989)
67. Liu, M.Y., Tuzel, O., Ramalingam, S., Chellappa, R.: Entropy rate superpixel segmentation. In: CVPR 2011, pp. 2097–2104. IEEE (2011)
68. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vision **59**(2), 167–181 (2004)
69. Oreshkin, B., López, P.R., Lacoste, A.: Tadam: task dependent adaptive metric for improved few-shot learning. In: Advances in Neural Information Processing Systems, pp. 721–731 (2018)
70. Heimann, T., Meinzer, H.P.: Statistical shape models for 3D medical image segmentation: a review. Med. Image Anal. **13**(4), 543–563 (2009)
71. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-Atlas labeling beyond the cranial vault-workshop and challenge (2015)
72. Kavur, A.E., et al.: Chaos challenge-combined (CT-MR) healthy abdominal organ segmentation. arXiv preprint arXiv:2001.06535 (2020)
73. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. **41**(12), 2933–2946 (2018)
74. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
75. Zhou, Y., et al.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10672–10681 (2019)
76. Isensee, F., et al.: nnU-Net: self-adapting framework for U-Net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)