

# Exemplar Learning for Medical Image Segmentation

Qing En<sup>1</sup>

QingEn@cunet.carleton.ca

Yuhong Guo<sup>1,2</sup>

YuhongGuo@carleton.ca

<sup>1</sup> Carleton University

Ottawa, Canada

<sup>2</sup> Canada CIFAR AI Chair

Amii, Canada

## Abstract

Medical image annotation typically requires expert knowledge and hence incurs time-consuming and expensive data annotation costs. To alleviate this burden, we propose a novel learning scenario, Exemplar Learning (EL), to explore automated learning processes for medical image segmentation with a single annotated image example. This innovative learning task is particularly suitable for medical image segmentation, where all categories of organs can be presented in one single image and annotated all at once. To address this challenging EL task, we propose an Exemplar Learning-based Synthesis Net (ELNet) framework for medical image segmentation that enables innovative exemplar-based data synthesis, pixel-prototype based contrastive embedding learning, and pseudo-label based exploitation of the unlabeled data. Specifically, ELNet introduces two new modules for image segmentation: an exemplar-guided synthesis module, which enriches and diversifies the training set by synthesizing annotated samples from the given exemplar, and a pixel-prototype based contrastive embedding module, which enhances the discriminative capacity of the base segmentation model via contrastive representation learning. Moreover, we deploy a two-stage process for segmentation model training, which exploits the unlabeled data with predicted pseudo segmentation labels. To evaluate this new learning framework, we conduct extensive experiments on several organ segmentation datasets and present an in-depth analysis. The empirical results show that the proposed exemplar learning framework produces effective segmentation results.

## 1 Introduction

Medical image analysis is becoming increasingly important for clinical diagnosis and surgical planning due to the rapid advancement of medical imaging technologies [1, 2]. Notably, medical image segmentation is one of the critical steps in quantitative medical image analysis, aiming to automatically identify the target region from medical images pixel-by-pixel [3, 4, 5]. Fully supervised deep neural networks have been demonstrated to yield desirable segmentation results by using large amounts of labeled training data [6, 7]. However, obtaining abundant annotated medical images at the pixel-level entails substantial labour and financial expenses because annotating medical images requires the knowledge of clinical experts that is not always available. To reduce the annotation cost, several techniques have been

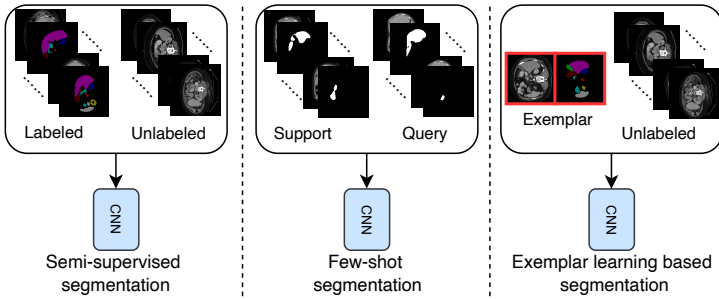


Figure 1: Exemplar learning vs. semi-supervised segmentation and few-shot segmentation. In semi-supervised segmentation, multiple labeled samples and many unlabeled samples are present. In few-shot segmentation, many support and query samples are available from the base categories. By contrast, only one labeled sample and many unlabeled samples are available in exemplar learning.

developed to perform medical image segmentation with less annotated data [4, 8], including semi-supervised segmentation [13, 18, 30] and few-shot segmentation [20, 26].

Although some notable improvements have been achieved, current solutions are still unable to eliminate the labelling conundrum. Most few-shot segmentation methods rely on extensive auxiliary training datasets with exhaustive annotated data to transfer knowledge from the support set to the query set [20, 28]. Semi-supervised segmentation methods usually focus on exploiting the consistency property of unlabeled data, but they still require a nontrivial portion of the densely annotated data [13, 22, 30]. We observe that the images for medical segmentation tasks often contain variations of the same set of organ categories, while a proper example image can cover all the parts for the whole organ category set. Motivated by this observation, we propose a novel learning scenario called Exemplar Learning (EL) to set up the working environment for a new set of medical image segmentation techniques that require only one single expert annotated image. The differences between this new EL setting and the previous semi-supervised segmentation and the few-shot segmentation settings are illustrated in Figure 1.

The fundamental challenges for exemplar learning lie in the following two aspects: (1) Data diversity is severely deficient, and the number of foreground-background pixels is imbalanced in medical image datasets. Since only one annotated image is available, the model can easily be overfitted to the labeled sample. (2) The contrast level between the organ and the background is low, and the differences among multiple organs’ appearances are barely discernible, easily resulting in the lack of discriminative capacity for the segmentation models. This phenomenon makes it difficult to distinguish the boundaries between organs, leading to over-segmentation issues. The fact that humans can learn by analogy [10, 21, 56] motivates us to address the abovementioned challenges by enriching the sample diversity and the discriminability of models.

In this paper, we propose a novel framework, *ELNet*, for learning to segment medical images effectively with only one annotated image. *ELNet* enriches the diversity of the labeled data by synthesizing training data and enhances the discriminability of the base segmentation model by performing pixel-prototype based contrastive embedding learning. Specifically, given the exemplar image with all organ categories labeled, we first devise an exemplar-guided synthesis module (ESM) to enlarge our training set by taking crops of foreground organs and pasting them through various transformations onto different background

images. This can increase the invariance of organ representations to different backgrounds, while enriching the diversity of the sample. Next, we design a pixel-prototype based contrastive embedding module (PCEM) to decompose the organs into distinct and consistent parts by capturing homogeneous components of the same type through contrastive embedding learning. This module enables pixels belonging to the same organ to be similar, in contrast to the case where pixels belong to different organs, and hence is expected to improve the discriminability of the segmentation model. Moreover, we deploy a two-stage process for segmentation model training, which exploits the unlabeled data with predicted pseudo segmentation labels to further improve the segmentation model. The main contributions of our paper can be summarized as follows:

- We propose a novel learning scenario, Exemplar Learning, which investigates medical image segmentation with a single annotated image.
- We propose a novel ELSNet framework to segment medical images in the EL scenario by creating exemplar-based synthetic data, learning pixel-prototype based contrastive embeddings, and exploiting unlabeled data with pseudo-labels.
- Experimental results on two medical image segmentation datasets show that the proposed ELSNet can effectively perform the medical semantic segmentation task.

## 2 Related Work

**Semi-Supervised Medical Image Segmentation.** Semi-supervised semantic segmentation has received increasing attention to train models by reducing the mask labeling cost [13, 15, 24, 25, 37]. The technique has also been applied to the field of medical image segmentation [11, 18, 22, 30]. Reiß et al. [18] proposed a multi-label deep supervision model to supervise low-resolution features and applied it to multiple medical supervision signals. Wu et al. [30] presented a semi-supervised polyp segmentation model by collaborative and adversarial learning. Moreover, Seibold et al. [22] used labeled images as references to generate more accurate pseudo-labels. Luo et al. [11] proposed a dual-task-consistency semi-supervised framework for medical image segmentation. These semi-supervised methods require multiple annotated images for model training. By contrast, we propose to train segmentation models with only one annotated image, which is more challenging.

**Few-Shot Medical Image Segmentation.** Few-shot segmentation has been exploited in the medical image domain [9, 14, 16, 32, 35]. Most of these methods require many base categories to be annotated during the training phase and require fine-tuning for unseen classes. SE-Net [20] and RP-Net [28] focus on designing models to segment unseen classes without retraining. Sli2Vol [31] propagated the 2D image segmentation with an affinity matrix directly to reconstruct the rest of the image in 3D volumes in a self-supervised manner. Ouyang *et al.* [17] generated superpixel-based pseudo-labels and used the adaptive local prototype information for training the self-supervised FSS framework. These methods require the support set in the test phase, and the predicted masks only contain foreground and background categories. By contrast, the proposed exemplar learning does not require a support set, and the predicted masks have semantic information.

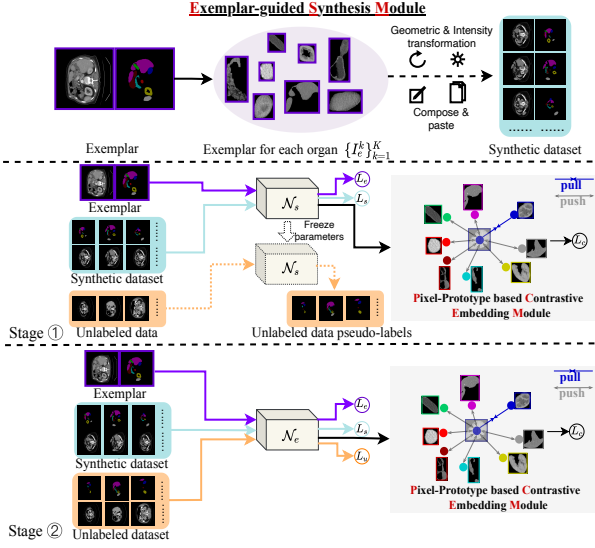


Figure 2: An overview of the proposed ELSNet framework. First, ESM enriches the training data diversity. In stage 1, the segmentation network  $\mathcal{N}_s$  is trained on the exemplar and synthetic data with PCEM and then used to generate the pseudo-labels of unlabeled data. In stage 2, the exemplar, synthetic and unlabeled data are all used to train the segmentation network  $\mathcal{N}_e$  with PCEM.

### 3 Proposed Method

In the setting of the exemplar learning, one labeled training image (*i.e.* Exemplar) and  $T$  unlabeled training images are given, and denoted as  $\mathcal{D}_E = (I_e, Y_e)$  and  $\mathcal{D}_U = \{(I_u^t)\}_{t=1}^T$ . An input image is defined as  $I \in \mathbb{R}^{1 \times H \times W}$ , and the label is defined as  $Y_e \in \{0, 1\}^{K \times H \times W}$ , where  $K, H, W$  are the number of categories in the dataset, height and width of the input image, respectively. In this setting, the single exemplar image contains one segmentation instance for each category. We propose an Exemplar Learning-based Synthesis Net (ELSNNet) framework to train a good segmentation model from the given input images.

The overall architecture of the ELSNet is illustrated in Figure 2. We first create the synthetic dataset from the given exemplar by the exemplar-guided synthesis module, which enriches and diversifies the training set. Moreover, we deploy a two-stage process for training the segmentation model by using the pixel-prototype based contrastive embedding module, which enhances the discriminative capacity of the base segmentation model and exploits the unlabeled data with predicted pseudo segmentation labels. In the first stage, a synthetic segmentation network  $\mathcal{N}_s$  is trained on the exemplar and the synthetic dataset, which is then used to generate pseudo-labels for unlabeled data. In the second stage, an exemplar learning segmentation network  $\mathcal{N}_e$  is trained on the exemplar, the synthetic dataset and the unlabeled dataset (with pseudo labels). The two segmentation networks share the same structure, consisting of an encoder  $f_{encoder} : \mathbb{R}^{1 \times H \times W} \rightarrow \mathbb{R}^{c \times h \times w}$  and a decoder  $f_{decoder} : \mathbb{R}^{c \times h \times w} \rightarrow \mathbb{R}^{K \times H \times W}$ , where  $c, h$ , and  $w$  are the channel, height, and width of the embedding matrix, which is represented as  $X = f_{encoder}(I)$ . The decoder's output is used to produce the segmentation masks. Below we elaborate the two modules of the ELSNet framework and the training process.

### 3.1 Exemplar-Guided Synthesis Module

This module aims to synthesize the segmentation instance of each label category from the single exemplar into various backgrounds, thus creating a synthetic training dataset. The diagram of this module is shown in Figure 2. Given an exemplar, we first obtain the segmentation instance of each organ according to their annotations, then transform them to imitate the various appearances of the organs in medical images, and finally paste the transformed organs onto background images. We define a series of geometric transform operations and intensity transform operations as  $\mathcal{T}_g$  and  $\mathcal{T}_i$ , respectively. Then in principle, a synthetic sample  $I_s$  and its label  $Y_s$  can be generated by performing different transformations on  $I_e$  and  $Y_e$  as follows:

$$(I_s, Y_s) = \Omega(\mathcal{T}_g(\mathcal{T}_i(I_e)), \mathcal{T}_g(Y_e), \mathcal{T}_g(\mathcal{T}_i(I_b))), \quad (1)$$

where  $\Omega$  represents the proposed exemplar-guided synthesis operation that copies, transforms and pastes the exemplar onto the selected background image  $I_b$ . We choose black images and images that do not contain any organs as background images. Meanwhile, the corresponding label  $Y_s$  is generated based on the transformed  $Y_e$ . By this means, the synthetic training dataset can be created by using Eq. (1) with sufficient variations from one exemplar.

To better accommodate various transformations, we propose to implement the  $\Omega$  operation in a category-wise manner. First, we segregate the exemplar  $I_e$  into different categories of organs as follows:

$$I_e^k = I_e \otimes Y_e^k, \quad (2)$$

where  $Y_e^k$  and  $I_e^k$  indicate the mask and the exemplar instance for the  $k$ th organ category, respectively. Then, for the segmentation instance of each category from the single exemplar, geometric and intensity transformations are applied to imitate scale change, rotation, blur and intensity variations across the dataset. Finally, we compose the transformed exemplar organs and paste them onto the background images, creating a synthetic sample. Following this procedure, a synthetic dataset  $\mathcal{D}_s = \{(I_s^b, Y_s^b)\}_{b=1}^B$  can be created, where  $I_s^b \in \mathbb{R}^{1 \times H \times W}$ ,  $Y_s^b \in \{0, 1\}^{K \times H \times W}$  and  $B$  is the number of synthetic samples. With this synthetic dataset, a segmentation model can be trained without extra annotation effort. Moreover, as the basic features of all organs are present in the exemplar image, the synthetic dataset can overcome the limitations of previous works that rely on unrealistic virtual data [8, 9].

### 3.2 Pixel-Prototype Based Contrastive Embedding Module

In this module, we calculate prototypes of different categories of organs and deploy a contrastive learning paradigm over the prototypes to improve the discriminability of the model. Specifically, the organ prototypes are calculated based on the predicted masks from the decoder output. As previously stated,  $X$  represents the embedding features of the input image, and we use  $f_{decoder}$  to generate the predicted mask  $\hat{Y}$  as follows:

$$\hat{Y} = \operatorname{argmax}(\operatorname{softmax}(f_{decoder}(X))), \quad (3)$$

where  $\operatorname{softmax}$  denotes the class-wise softmax function and  $\hat{Y} \in \mathbb{R}^{K \times H \times W}$  is the predicted label indicator matrix. We resize the predicted mask  $\hat{Y} \in \mathbb{R}^{K \times H \times W}$  to the same size as the embedding features  $X$  via bilinear interpolation, which is denoted by  $\hat{Y}_x \in \mathbb{R}^{K \times h \times w}$ . With  $\hat{Y}_x$ , we leverage global average pooling [33, 34] over the foreground to integrate the pixel

features belonging to the same category into a feature vector, which is seen as the prototype of the corresponding organ. The prototype for the  $k$ th category,  $v_k$ , is computed as follows:

$$v_k = \frac{\sum_{(i,j)} X^{(i,j)} \mathbb{1}[\hat{Y}_x^{(k,i,j)} \neq 0]}{\sum_{(i,j)} \mathbb{1}[\hat{Y}_x^{(k,i,j)} \neq 0]}, \quad (4)$$

where  $(k, i, j)$  indicates the spatial location index of the  $k$ th category and  $\mathbb{1}$  represents the indicator function. The prototypes computed through the masked average pooling can extract global object representations for the target organs.

To maximize the representation similarity of the same organ among different images while simultaneously minimizing the similarity of different organs, we propose to deploy a contrastive learning loss to learn discriminative embeddings. We perform the calculation in a batch of  $N$  samples (e.g., a mini-batch). Specifically, for each category  $k$  and a prototype  $v_k^n$  from the  $n$ th image, we randomly select a prototype  $v_k^m \in \{v_k^i, i \neq n | v_k^n\}$  from the other  $N - 1$  images in the current batch as a positive sample, and use the prototypes from all other categories of the  $N - 1$  images as negative samples. Hence the prototype-based contrastive loss is defined as follows:

$$L_c = - \sum_{n=1}^N \sum_{k=1}^K \log \frac{\exp(v_k^n \cdot v_k^m / \tau)}{\exp(v_k^n \cdot v_k^m / \tau) + \sum_{j \neq k} \sum_{i \neq n} \exp(v_k^n \cdot v_j^i / \tau)}, \quad (5)$$

where  $\tau$  is the temperature hyper-parameter;  $n$  is the index of the images.

This proposed module is designed with two main differences from previous self-supervised approaches. First, our contrastive loss over the prototypes of various organs relies on the prediction mask rather than the entire image, which is directly related to the objective of the segmentation task. Second, instead of augmenting the inputs to produce multiple copies, we perform contrastive learning across multiple images, aiming to capture the semantic representation of the same organ that is invariant across different images. Overall, by enforcing the embeddings of the same organ category to be similar in different images than that of different organ categories, PCEM can enhance the discriminability of the embedding learning.

### 3.3 Two-Stage Training

The ELSNet is trained in two stages. In the first stage, the exemplar  $\mathcal{D}_E$  and the synthetic dataset  $\mathcal{D}_S$  are used as training data to train the segmentation network  $\mathcal{N}_s$  by minimizing the following joint loss function:

$$L_{s1} = L_e + \lambda_s L_s + \lambda_c L_c \quad (6)$$

where  $\lambda_s$  and  $\lambda_c$  are trade-off hyperparameters,  $L_c$  denotes the prototype-based contrastive loss defined in Eq.(5),  $L_e$  and  $L_s$  denote the segmentation losses computed from the exemplar and the synthetic dataset respectively, such that  $L_e = \mathcal{L}_{seg}(\hat{Y}_e, Y_e)$  and  $L_s = \mathcal{L}_{seg}(\hat{Y}_s, Y_s)$ . The segmentation loss  $\mathcal{L}_{seg}$  is defined as follows:

$$\mathcal{L}_{seg}(\hat{Y}, Y) = 0.5 * l_{ce}(\hat{Y}, Y) + 0.5 * l_{dice}(\hat{Y}, Y), \quad (7)$$

where  $l_{ce}$  is the cross-entropy loss function and  $l_{dice}$  is the Dice loss function;  $\hat{Y}$  denotes the predicted output of the segmentation network during the training stage.

After training the segmentation network  $\mathcal{N}_s$ , we use it to segment each image  $I_u$  in the unlabeled set  $\mathcal{D}_U$  and obtain its predicted segmentation mask  $Y_u$  as the pseudo-labels, thereby

constructing a pseudo-labeled set  $\mathcal{D}_U = \{(I_u^t, Y_u^t)\}_{t=1}^T$ . Then in the second stage of training, we train our final segmentation network  $\mathcal{N}_e$  by using all three sets of images, the exemplar  $\mathcal{D}_E$ , the synthetic dataset  $\mathcal{D}_S$  and the pseudo-labeled  $\mathcal{D}_U$  by minimizing the following joint loss function:

$$L_{s2} = L_e + \lambda_s L_s + \lambda_c L_c + \lambda_u L_u \quad (8)$$

where  $\lambda_u$  is a trade-off hyperparameter, and the segmentation loss on the unlabeled data,  $L_u = \mathcal{L}_{seg}(\hat{Y}_u, Y_u)$ , is incorporated into training with their pseudo-labels. After the two-stage training,  $\mathcal{N}_e$  can be used for inference on test images.

## 4 Experiments

### 4.1 Experimental Setting

**Implementation details.** We adopt a U-Shape transformer-based structure [29] as the basic structure of  $\mathcal{N}_s$  and  $\mathcal{N}_e$  in the proposed ELSNet. The weights of the proposed ELSNet are randomly initialized. The input image size is set to  $224 \times 224$  with random rotation and flipping. Adam optimizes the proposed ELSNet with a weight decay of 0.0001 and a learning rate of  $1e-4$ . The batch size is set to 12 and the value  $\tau$  is set to 0.07. We divide all medical 3D volumes into individual images during the testing stage for inference [8, 29]. We randomly select a sample containing all categories in the training set as the exemplar. In order to reduce the influence of fluctuations in the results, average results over five runs are reported for each experiment.

**Datasets and Evaluation Metrics.** We evaluate the proposed framework on the Synapse dataset<sup>1</sup> and the ACDC dataset<sup>2</sup>. Synapse is a multi-label organ dataset containing 30 abdominal clinical CT cases with 2211 images, and we use 18 cases for training and 12 cases for testing [8, 29]. ACDC is a cardiac MRI dataset that contains 100 cases from MRI scanners with 1300 images. We used 70 cases for training, 20 for evaluation, and 10 for testing. Following [8], we evaluated the proposed framework based on two metrics, namely, the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95).

### 4.2 Experimental Results

#### 4.2.1 Comparison Results

We compared the ELSNet with three state-of-the-art image segmentation methods under the same experimental setting on the ACDC and Synapse datasets: UNet [19], MT-UNet [29] and MLDS [18]. We re-implemented these methods and trained them under the same setting as the proposed ELSNet. The comparison results on the two datasets are reported in Table 1 and Table 2, respectively. On the ACDC dataset, the proposed ELSNet achieves considerable improvements, outperforming the second-best method, the semi-supervised MLDS, by 0.221 and 23.39 in terms of the class average DSC and HD95 results, respectively. The Synapse dataset contains more organ categories with different sizes. The foregrounds and background are more difficult to distinguish in the images of this dataset. Nevertheless, on the Synapse dataset, ELSNet again outperforms all three comparison methods and produces the best class

<sup>1</sup><https://www.synapse.org/#!/Synapse:syn3193805/wiki/217789>

<sup>2</sup><https://www.creatis.insa-lyon.fr/Challenge/acdc/>



Table 1: Comparison results on the ACDC dataset. The class average results and the results for individual classes in terms of DSC and HD95 are reported.

Method	DSC.Avg↑	RV	Myo	LV	HD95.Avg↓	RV	Myo	LV
UNet[ <a href="#">18</a> ]	0.142	0.140	0.112	0.174	43.30	63.76	35.60	30.80
MT-UNet[ <a href="#">19</a> ]	0.142	0.119	0.126	0.182	74.20	83.91	61.48	77.22
MLDS[ <a href="#">18</a> ]	0.189	0.144	0.165	0.258	50.03	72.13	30.20	47.77
ELNet	<b>0.410</b>	<b>0.293</b>	<b>0.374</b>	<b>0.563</b>	<b>26.64</b>	<b>47.63</b>	<b>16.58</b>	<b>15.73</b>

Table 2: Comparison results on the Synapse dataset. The class average DSC and HD95 results and the DSC results for all individual classes are reported.

Method	HD95↓	DSC↑	Aor	Gal	Kid(L)	Kid(R)	Liv	Pan	Spl	Sto
UNet[ <a href="#">18</a> ]	132.42	0.160	0.026	0.167	0.177	0.154	0.649	0.015	0.059	0.033
MTUNet[ <a href="#">19</a> ]	154.60	0.112	0.066	0.108	0.155	0.053	0.352	0.008	0.046	0.102
MLDS[ <a href="#">18</a> ]	159.26	0.221	0.057	0.147	0.306	0.183	0.638	0.038	<b>0.306</b>	0.090
ELNet	<b>109.70</b>	<b>0.315</b>	<b>0.319</b>	<b>0.372</b>	<b>0.381</b>	<b>0.219</b>	<b>0.784</b>	<b>0.067</b>	0.276	<b>0.104</b>

average DSC result of 0.315 and the best class average HD95 result of 109.70. Moreover, ELSNet achieves the best results in almost all categories except *Spl*. The improvements are particularly large in the Aor and Gal categories. Overall, these results validate the efficacy of the proposed ELSNet for EL.

#### 4.2.2 Qualitative Evaluation

To further validate the segmentation performance of the proposed ELSNet, several visualized segmentation examples for Baseline, MLDS, and ELSNet are presented in Figure 3. Baseline refers to the base segmentation network that is directly trained from the single annotated image. Typically, the tiny sizes of some organs can make segmentation very challenging, not to mention that there is only one annotation example. However, we can see from Figure 3 that compared with the existing state-of-the-art method, ELSNet can segment more accurately, even under conditions of deformation, edge ambiguity, shape complexity and background clutter shown in the examples.

#### 4.2.3 Ablation Studies

**Impact of the proposed modules.** We tested the empirical contributions of the proposed modules on the two datasets and the results are reported in Table 3. Following the general experimental setup in [[18](#), [19](#)], BS denotes the baseline that uses only the exemplar as supervision, which achieves 0.112 and 0.142 in terms of DSC measure on the two datasets, Synapse and ACDC, respectively. By adding the proposed exemplar-guided synthesis module, “+ESM” substantially improves the performance to 0.234 and 0.273 in terms of DSC on the two datasets. Such performance gains highlight the impact of the synthetic dataset produced by ESM, which enriches the diversity of the labeled samples and increases the generalization capability of the model. By further including the proposed pixel-prototype based contrastive embedding module in stage 1, the results on the two datasets reach DSC values of 0.264 and 0.355 respectively. This demonstrates the impact of the PCEM module on enhancing the discriminative ability of the segmentation model. Finally, the full model,



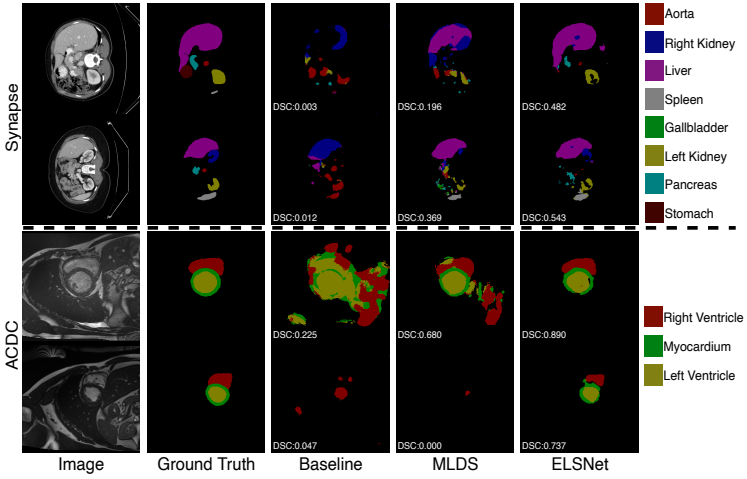


Figure 3: Visualized results obtained with the Baseline, the MLDS and the proposed ELSNet on Synapse and ACDC datasets. The DSC value of each sample is also presented.

Table 3: Ablation study of the proposed modules on the Synapse and ACDC datasets. BS denotes the baseline. +ESM denotes the variant that further includes the proposed ESM. +ESM+PCEM\_S1 denotes the variant produced when both ESM and PCEM modules are used in stage 1. +ESM+PCEM\_S1&2 denotes the full approach.

Method	Synapse				ACDC			
	DSC↑	$\Delta$ DSC	HD95↓	$\Delta$ HD95	DSC↑	$\Delta$ DSC	HD95↓	$\Delta$ HD95
BS	0.112	-	154.60	-	0.142	-	74.20	-
+ESM	0.234	<b>+0.122</b>	120.59	<b>-34.01</b>	0.273	<b>+0.131</b>	38.35	<b>-35.85</b>
+ESM+PCEM_S1	0.264	<b>+0.152</b>	<b>101.00</b>	<b>-53.60</b>	0.355	<b>+0.213</b>	40.80	<b>-33.40</b>
+ESM+PCEM_S1&2	<b>0.315</b>	<b>+0.203</b>	109.70	<b>-44.90</b>	<b>0.410</b>	<b>+0.268</b>	<b>26.64</b>	<b>-47.56</b>

" +ESM+PCEM\_S1&2", produces the best results on both datasets by using all of the proposed modules in stage 1 and 2. These results validate the impact of the proposed modules on the overall performance of the proposed ELSNet framework.

**Impact of different loss functions.** To demonstrate the effectiveness of the loss functions involved in the training stages of ELSNet, we summarize the ablation results over multiple loss terms on the left side of Table 4. The results in the second row show that creating the synthetic dataset with ESM ( $L_s$ ) can significantly improve the DSC results to 0.273. Using two modules, ESM ( $L_s$ ) and PCEM ( $L_c^{e+s}$ ), simultaneously to train the model in stage 1 can lead to further substantial improvements, as shown in the third row. Also, note that ignoring the synthetic dataset but still using the pseudo-labels of unlabeled data in stage 2 can obtain a result value of 0.359, which shows the unlabeled data is useful. When the contrastive loss  $L_c$  is only used in either stage 1 or stage 2, the DSC results (in the sixth row and fifth row) degrade from the full model (the last row) that uses  $L_c$  in both stages. The full model with all the loss terms yields the best results.

Table 4: Ablation results on the ACDC dataset. **Left side** presents the ablation results with different loss functions.  $L_c^{e+s}$  indicates that  $L_c$  is used in Eq.(6) for stage 1.  $L_c^{e+s+u}$  indicates that  $L_c$  is used in Eq.(8) for stage 2. **Right side** presents the ablation results with different transformation strategies. Int.E and Int.B indicate intensity transformations are applied on Exemplar and Background images. Similarly, Geo.E/Geo.B indicates geometry transformations.

$L_e$	$L_s$	$L_c^{e+s}$	$L_c^{e+s+u}$	$L_u$	RV	Myo	LV	DSC.Avg↑	Int.E	Int.B	Geo.E	Geo.B	RV	Myo	LV	DSC.Avg↑
✓	-	-	-	-	0.119	0.126	0.182	0.142	-	-	-	-	0.136	0.159	0.232	0.176
✓	✓	-	-	-	0.192	0.265	0.362	0.273	✓	-	✓	-	0.255	0.317	0.362	0.311
✓	✓	✓	-	-	0.249	0.315	0.503	0.355	✓	✓	-	-	0.191	0.308	0.448	0.315
✓	-	✓	✓	✓	0.263	0.323	0.488	0.359	-	-	✓	✓	0.210	0.356	0.451	0.339
✓	✓	-	✓	✓	0.218	0.355	0.462	0.345	✓	-	✓	✓	0.242	0.333	0.464	0.346
✓	✓	✓	-	✓	0.290	0.354	0.529	0.390	✓	✓	✓	-	0.272	0.334	0.502	0.370
✓	✓	✓	✓	✓	<b>0.293</b>	<b>0.374</b>	<b>0.563</b>	<b>0.410</b>	✓	✓	✓	✓	<b>0.293</b>	<b>0.374</b>	<b>0.563</b>	<b>0.410</b>

**Impact of the transformation strategies.** We tested different transformation strategies for synthesizing images in ESM by applying Geometric (Geo: scaling, rotation) and Intensity (Int: blur, intensity variations) transformations on the exemplar and background images. The results of different variants are summarized and reported on the right side of Table 4. When none of the transformations is utilized, *i.e.*, we paste the exemplar into different background images to build the synthetic dataset, minor improvement is achieved over the baseline that only uses the exemplar, as shown in the first row. When both geometric and intensity transformations are applied to the exemplar, the average DSC performance is substantially improved to 0.311 (second row). By applying the two types of transformations on both the exemplar and the background images, the performance of the full model improves by another 10% or so to 0.410 (the bottom row). We also observe from the third and fourth rows that dropping either type of transformations (Geo or Int) can degrade the performance. These results suggest that both types of transformation are effective for generating useful synthetic data and the proposed ESM is reasonable. Moreover, dropping the intensity transformation on the background (fifth row) leads to larger performance degradation than dropping the geometric transformation on the background (sixth row), which indicates that the intensity diversity of the background is more important than the geometric diversity.

## 5 Conclusions

This paper introduced a new experimental scenario, Exemplar Learning, and proposed a novel framework, ELSNet, to learn segmentation models from only one annotated image. ELSNet uses an exemplar-guided synthesis module (ESM) to enrich and diversify the training data by synthesizing annotated samples from the given exemplar, and uses a pixel-prototype based contrastive embedding module (PCEM) to increase the discriminative ability of the segmentation model by contrastive embedding learning. A two-stage training process is deployed to exploit the unlabeled data via pseudo-labels. Experimental results demonstrate that the proposed framework is effective and outperforms existing segmentation methods under scenarios with very limited supervision information.

## References

- [1] Moshe Bar. The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7):280–289, 2007.
- [2] Chen Chen, Chen Qin, Huaqi Qiu, Cheng Ouyang, Shuo Wang, Liang Chen, Giacomo Tarroni, Wenjia Bai, and Daniel Rueckert. Realistic adversarial data augmentation for mr image segmentation. In *MICCAI*, 2020.
- [3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [4] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. In *NeurIPS*, 2021.
- [5] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2015.
- [6] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017.
- [7] James S Duncan and Nicholas Ayache. Medical image analysis: Progress over two decades and the challenges ahead. *IEEE transactions on pattern analysis and machine intelligence*, 22(1):85–106, 2000.
- [8] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.
- [9] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille. Domain adaptive relational reasoning for 3d multi-organ segmentation. In *MICCAI*, 2020.
- [10] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.
- [11] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *AAAI*, 2021.
- [12] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.
- [13] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1369–1379, 2019.
- [14] Arnab Kumar Mondal, Jose Dolz, and Christian Desrosiers. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241*, 2018.

- [15] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020.
- [16] Cheng Ouyang, Konstantinos Kamnitsas, Carlo Biffi, Jinming Duan, and Daniel Rueckert. Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *MICCAI*, 2019.
- [17] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *ECCV*, 2020.
- [18] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *CVPR*, 2021.
- [19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.
- [20] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. ‘squeeze & excite’guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020.
- [21] Chaitanya Ryali, David J Schwab, and Ari S Morcos. Learning background invariance improves generalization and robustness in self-supervised learning on imagenet and beyond. In *NeurIPS Workshop on ImageNet: Past, Present, and Future*, 2021.
- [22] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. In *AAAI*, 2022.
- [23] Neeraj Sharma, Lalit M Aggarwal, et al. Automated medical image segmentation techniques. *Journal of medical physics*, 35(1):3, 2010.
- [24] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.
- [25] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, 2017.
- [26] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.
- [27] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.
- [28] Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. In *ICCV*, 2021.

- [29] Hongyi Wang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xian-Hua Han, Yen-Wei Chen, and Ruofeng Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP*, 2022.
- [30] Huisi Wu, Guilian Chen, Zhenkun Wen, and Jing Qin. Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation. In *ICCV*, 2021.
- [31] Pak-Hei Yeung, Ana IL Namburete, and Weidi Xie. Sli2vol: Annotate a 3d volume from a single slice with self-supervised learning. In *MICCAI*, 2021.
- [32] Hanchao Yu, Shanhui Sun, Haichao Yu, Xiao Chen, Honghui Shi, Thomas S Huang, and Terrence Chen. Foal: Fast online adaptive learning for cardiac motion estimation. In *CVPR*, 2020.
- [33] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *CVPR*, 2019.
- [34] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 2020.
- [35] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *CVPR*, 2019.
- [36] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. In *AAAI*, 2021.
- [37] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *ICLR*, 2020.