Self-Supervised Tuning for Few-Shot Segmentation

Kai Zhu, Wei Zhai and Yang Cao*

University of Science and Technology of China
{zkzy, wzhai056}@mail.ustc.edu.cn, forrest@ustc.edu.cn

Abstract

Few-shot segmentation aims at assigning a category label to each image pixel with few annotated samples. It is a challenging task since the dense prediction can only be achieved under the guidance of latent features defined by sparse annotations. Existing meta-learning method tends to fail in generating category-specifically discriminative descriptor when the visual features extracted from support images are marginalized in embedding space. To address this issue, this paper presents an adaptive tuning framework, in which the distribution of latent features across different episodes is dynamically adjusted based on a self-segmentation scheme, augmenting category-specific descriptors for label prediction. Specifically, a novel self-supervised inner-loop is firstly devised as the base learner to extract the underlying semantic features from the support image. Then, gradient maps are calculated by back-propagating self-supervised loss through the obtained features, and leveraged as guidance for augmenting the corresponding elements in embedding space. Finally, with the ability to continuously learn from different episodes, an optimization-based meta-learner is adopted as outer loop of our proposed framework to gradually refine the segmentation results. Extensive experiments on benchmark PASCAL-5i and COCO-20i datasets demonstrate the superiority of our proposed method over state-of-the-art.

1 Introduction

Recently, semantic segmentation models [Long et al., 2015] have made great progress under full supervision, and some of them have even surpassed the level of human recognition. However, when the learned model is applied to a new segmentation task, it takes great cost to collect a large amount of full annotated data in pixel level. Furthermore, samples are not available in large quantities in some areas such as health care, security and so on. To address this problem, various few-shot segmentation methods are proposed.

*Corresponding author

One solution for solving few-shot segmentation [Shaban et al., 2017] is meta-learning [Munkhdalai and Yu, 2017], whose general idea is to utilize a large number of episodes similar to target task to learn a meta learner that generate an initial segmentation model, and a base learner that quickly tunes the model with few samples. In most methods, a powerful feature extractor with good migration ability is provided for the meta learner to map the query and support images into a shared embedding space. And the base learner generates a category-specific descriptor with the support set. The similarity between the query’s feature maps and the descriptor is measured under a parametric or nonparametric metric, and leveraged as guidance for dense prediction of the query branch.

However, when the low-level visual features of foreground objects extracted from the support images are too marginalized in the embedding space [Zhang et al., 2014], the generated descriptor by base learner is not category-specifically discriminative. In this case, the regions to be segmented in the query image may be ignored or even confused with other categories in the background. For example, as shown in Fig. 1, the bottle-specific descriptor generated from the support...
image is unable to identify the bottle itself in the same image, and therefore inapplicable for bottles in other complicated scenarios to be segmented.

To address the issue, this paper presents an adaptive tuning framework for few-shot segmentation, in which the marginalized distributions of latent features are dynamically adjusted by a self-supervised scheme. The core of the proposed framework is the base learner driven by a self-segmentation task, which augments category-specific constraint on each new episode and facilitate successive labeling. Specifically, the base learner is designed as an inner-loop task, where the support images are segmented under the supervision of the presented masks of input support images. By back-propagating self-supervised loss through the support feature map, the corresponding gradients are calculated and used as guidance for augmenting each element in the support embedding space.

Moreover, the self-segmentation task can be considered as a special case of few-shot segmentation with the query and support image to be the same. Therefore, we also utilize the resulting loss of the task (we call it auxiliary loss in this paper) to promote the training process. Since auxiliary loss is equivalent to performing data enhancement to annotated samples, it can involve more information for training with the same number of iterations. The evaluation is performed on the prevailing public benchmark datasets PASCAL-5i and COCO-20!, and the results demonstrate above-par segmentation performance and generalization ability of our proposed method.

Our main contributions are summarized as follows:

1. An adaptive tuning framework is proposed for few-shot segmentation, in which the marginalized distributions of latent category features are dynamically adjusted by a self-supervised scheme.
2. A novel base learner driven by a self-segmentation task is proposed, which augments category-specific feature description on each new episode, resulting in better performance on label prediction.
3. Experimental results on two public benchmark datasets PASCAL-5i and COCO-20! demonstrate the superiority of our proposed method over SOTA.

2 Related Work

**Few-shot learning.** Few-shot learning has recently received a lot of attention and substantial progress has been made based on meta-learning. Generally, these methods can be divided into three categories. Metric-based methods focus on the similarity metric function over the embeddings [Snell et al., 2017]. Model-based methods mainly utilize the internal architecture of the network (such as memory module [Santoro et al., 2016], etc.) to realize the rapid parameter adaptation in new categories. The optimization-based method aims at learning an update scheme for base learner [Munkhdalai and Yu, 2017] in each episode. In the latest study, [Lee et al., 2019] and [Bertinetto et al., 2018] introduce machine learning methods such as SVM and ridge regression into the inner loop of the base learner, and [Rusu et al., 2018] directly replaces the inner loop with an encoded-decode network. These methods have achieved state-of-the-art performance in few-shot classification task. Our model also takes inspiration of them.

**Semantic segmentation.** Semantic segmentation is an important task in computer vision and FCNs [Long et al., 2015] [Zhang et al., 2019b] have greatly promoted the development of the field. After that, DeepLabV3 [Chen et al., 2017] and PSPNet [Zhao et al., 2017] propose different global contextual modules, which pay more attention to the scale change and global information [Chen et al., 2019] in the segmentation process. [Zhu et al., 2019] considers the full-image dependencies from all pixels based on Non-local Networks, which shows superior performance in terms of reasoning [Qiao et al., 2019].

**Few-shot semantic segmentation.** While the work on few-shot learning is quite extensive, the research on few-shot segmentation [Hu et al., 2019] has been presented only recently. [Shaban et al., 2017] first proposes the definition and task of one-shot segmentation. Following this, [Rakelly et al., 2018] solves the problem with sparse pixel-wise annotations, and then extends their method to interactive image segmentation and video object segmentation. [Dong and Xing, 2018] generalizes the few-shot semantic segmentation problem from 1-way (class) to N-way (classes). [Zhang et al., 2019a] introduces an attention mechanism to effectively fuse information from multiple support examples and proposes an iterative optimization module to refine the predicted results. [Nguyen and Todorovic, 2019] and [Wang et al., 2019] leverage the annotations of the support images as supervision in different ways. Different from [Tian et al., 2019] which employs the base learner directly with linear classifier, our method devises a novel self-supervised base learner which is more intuitive and effective for few-shot segmentation. Compared to [Nguyen and Todorovic, 2019], the support images are used for supervision in both training and test stages of our framework.

3 Problem Description

Here we define an input triple $T_{ri} = (Q_i, S_i, T_Q^i)$, a label $T_Q$ and a relation function $F: A_i = F(Q_i, S_i, T_Q^i; \theta)$, where $Q_i$ and $S_i$ are the query and support images containing objects of the same class $i$, correspondingly. $T_Q^i$ are the pixel-wise labels corresponding to the $i^{th}$ class objects in $S_i$ and $Q_i$. $A_i$ is the actual segmentation result, and $\theta$ is all parameters to be optimized in function $F$. Our task is to randomly sample triples from the dataset, train and optimize $\theta$, thus minimizing the loss function $L$:

$$\theta^* = \arg\min_{\theta} L(A_i, T_Q^i).$$ (1)

We expect that the relationship function $F$ can segment object regions of the same class in another target image each time it sees few support images belonging to a new class. This is the embodiment of the meaning of few-shot segmentation. It should be mentioned that the classes sampled by the test set are not present in the training set, that is, $U_{\text{train}} \bigcap U_{\text{test}} = \emptyset$. The relation function $F$ in this problem is implemented by the model detailed in Sec. 4.3.
4 Method

4.1 Model Overview

Different from the existing methods, this paper proposes a novel self-supervised tuning framework for few-shot segmentation, which is mainly composed of an adaptive tuning mechanism, a self-supervised base learner and a meta learner. These three components will be illustrated in the next three subsections. In this subsection, we mainly present the description of the whole framework mathematically, and the symbolic representation is consistent with that in Sec. 3.

As shown in Fig. 2, a Siamese network \( f_e \) [Koch et al., 2015] is first proposed to extract the features of input query and support images [Zhang et al., 2012]. The mechanism of parameters sharing not only promotes optimization but also reduces the amount of calculation. In this step, the latent visual feature representations \( R_q \) and \( R_s \) are obtained as follows:

\[
R_q = f_e(Q; \theta_e), \tag{2} \\
R_s = f_e(S_i; \theta_e). \tag{3}
\]

Here \( \theta_e \) is the learnable parameter of the sharing encoder. To dynamically adjust the latent features, we firstly devise a novel self-supervised inner-loop as the base learner \( (f_b) \) to exploit the underlying semantic information \( (\theta_s) \).

\[
\theta_s = f_b(R_s, T_m^S; \theta_b). \tag{4}
\]

Then the distribution of low-level visual features is tuned \( (f_t) \) according to the high-level category-specific cues obtained above:

\[
R'_s = f_t(R_s; \theta_s). \tag{5}
\]

Inspired by Relation Network [Sung et al., 2018], a deep non-linear metric \( f_m \) is introduced into our meta learner. It measures the similarity between the feature map of query image and the tuned feature, and accordingly determines regions of interest in query images. Finally, a segmentation decoder \( f_d \) is utilized to refine the response area to the original image size:

\[
M = f_m(R_q, R'_s; \theta_m) \tag{6} \\
S = f_d(M; \theta_d) \tag{7}
\]

Similarly, \( \theta_m \) and \( \theta_d \) stand for the parameters of metric and decoding part. With the ability to continuously learn from different episodes, our meta learner gradually optimizes the whole process above (outer loop) and improves the performance of base learner, metric and decoder.

4.2 Self-Supervised Tuning Scheme

The self-supervised tuning scheme is the core of our proposed method, which is implemented by a base learner driven by self-segmentation and an adaptive tuning module. Different from the existing base learner used in [Lee et al., 2019] [Bertinetto et al., 2018], our proposed base learner is generated under the supervision of the presented masks of input support images. This method originates from an intuitive idea, that is, the premise of identifying the regions with the same category as target objects is to identify the object itself first. Therefore, we duplicate the features of the support image as the two inputs of Eq. 6 in Sec. 4.1 and calculate the standard cross-entropy loss \( (L_{cross}) \) with the corresponding support mask:

\[
M_{sup} = f_m(R_s, R'_s; \theta_m) \tag{8} \\
S_{sup} = f_d(M_{sup}; \theta_d) \tag{9} \\
L_{sup} = L_{cross}(S_{sup}; T_m^S) \tag{10}
\]

To exploit the above information to better implement the segmentation task of the query branch, the marginalized distribution is adjusted in embedding space based on the category-specific semantic constraint. Specifically, a gradient map that is calculated by back-propagating self-supervised
Table 1: Ablation study on PASCAL-5$^i$ dataset under the metric of mean-IoU. Bold fonts represent the best results.

(a) Results for 1-shot segmentation.

<table>
<thead>
<tr>
<th>+SSM</th>
<th>+SS Loss</th>
<th>i=0</th>
<th>i=1</th>
<th>i=2</th>
<th>i=3</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✔</td>
<td>49.6</td>
<td>62.6</td>
<td>48.7</td>
<td>48.0</td>
<td>52.2</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>53.2</td>
<td>63.6</td>
<td>48.7</td>
<td>47.9</td>
<td>53.4</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>51.1</td>
<td>64.9</td>
<td>51.9</td>
<td>50.2</td>
<td>54.5</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>54.4</td>
<td>66.4</td>
<td>57.1</td>
<td>52.5</td>
<td>57.6</td>
</tr>
</tbody>
</table>

(b) Results for 5-shot segmentation.

<table>
<thead>
<tr>
<th>+SSM</th>
<th>+SS Loss</th>
<th>i=0</th>
<th>i=1</th>
<th>i=2</th>
<th>i=3</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>✔</td>
<td>53.2</td>
<td>66.5</td>
<td>55.5</td>
<td>51.4</td>
<td>56.7</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>56.6</td>
<td>67.2</td>
<td>60.4</td>
<td>54.0</td>
<td>59.6</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>56.8</td>
<td>68.7</td>
<td>61.4</td>
<td>55.0</td>
<td>60.5</td>
</tr>
<tr>
<td>✔</td>
<td>✔</td>
<td>58.6</td>
<td>68.7</td>
<td>63.1</td>
<td>55.3</td>
<td>61.4</td>
</tr>
</tbody>
</table>

4.3 Deep Learnable Meta Learner

Inspired by the Relation Network, we apply a deep non-linear metric to measure the similarity between the feature map of query image and the descriptor generated by base learner. As in [Rakelly et al., 2018], we also use the deep learnable metric with late fusion as the main component of meta learner. First, we multiply the features by the downsampling mask and aggregate it to obtain the latent features of the foreground. This feature is then tiled to the original spatial scale, so that each dimension of the query feature is aligned with the representative feature ($R^r_s$):

$$R^r_s = tile(pool(R^r_s \cdot T^l_S))$$ (12)

Through the Relation Network comparator, the response area of the query image is obtained:

$$M = Relation(R_q, R^r_s)$$ (13)

Finally, we feed it into the segmentation decoder, refining and restoring the original image size to get accurate segmentation results.

4.4 Loss and Generalization to 5-Shot Setting

In addition to the cross-entropy loss function of segmentation obtained by query-support set commonly used in other methods (main loss), we also include the cross-entropy loss of support-support segmentation from base learner (auxiliary loss) into the final training loss as an auxiliary. In our method, the auxiliary loss itself is created as an intermediate process, so there is no much extra computation required. It can be seen from the experiment part that the auxiliary loss can accelerate the convergence and improve the performance.

When generalizing to 5-shot segmentation, the main difference is the base learner part. Considering that the number of samples for 5-shot segmentation is still small, if we calculate the gradient optimization together, it is easy to produce large discrepancy. Therefore, we calculate the gradient value separately, getting 5 separate response areas, and then take the weighted summation according to the self-supervised scores to get the final result, that is:

$$M_{weighted} = \sum_{i=1}^{5} f_{IoU}(S^i_{sup}, T^i_S) \cdot f_m(R^i_q, R^i_s; \theta_m)$$ (14)

where $f_{IoU}$ represents the function which is used to calculate the IoU scores.

5 Experiment

5.1 Dataset and Settings

Dataset. To evaluate the performance of our model, we experiment on PASCAL-5$^i$ and COCO-20$^i$ datasets. The for-
The backbone of the existing methods are different in each row.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Results (mean-IoU %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>59.4</td>
</tr>
<tr>
<td>Average</td>
<td>61.2</td>
</tr>
<tr>
<td>Weighted</td>
<td>61.4</td>
</tr>
</tbody>
</table>

Table 2: Comparison among different fusion methods in 5-shot setting under the metric of mean-IoU.

is set to 0.0005 and the attenuation rate is set to 0.0005. The model stops training after 200 epochs. All images are resized to 321 × 321 size and the batch size is set to 16.

5.2 Ablation Study

To prove the effectiveness of our architecture, we conduct several ablation experiments on PASCAL-5\textsuperscript{i} dataset as shown in Table 1. To improve efficiency, we only choose ResNet50 as backbone for all models in the ablation study for fair comparison. The performance of our network is mainly attributed to two prominent components: SSM and auxiliary loss. Note that we progressively add addition components to the baseline, which enables us to gauge the performance improvement obtained by each of them. Due to the fact that the self-supervised tuning mechanism dynamically adjusts the marginalized distributions of latent features at each stage, SSM brings about a 1.2 and 2.9 percent mean-IoU increase in 1-shot and 5-shot settings, respectively. At the same time, we can see that auxiliary loss boosts the overall performance, resulting in a 2.3 and 3.8 percent improvement.

5.3 Analysis

As PASCAL-5\textsuperscript{i} is the most commonly used dataset by all few-shot segmentation methods, the main analysis parts of our experiment are accomplished on PASCAL-5\textsuperscript{i} dataset.

Effect of the Self-Supervised Module

To clarify the actual function of the self-supervised module for few-shot segmentation task, we show the following visualization results. As shown in Fig. 3, we feed the feature representations generated before and after the tuning process to Relation Network to obtain two sets of segmentation results. Note that the original network often segments other objects by mistakes that are easily confused in the background. After adding the SSM module, the category-specific semantic constraint is introduced to help query images to correct the
Comparison among Different 5-Shot Fusion Methods

To prove the superiority of our weighted fusion in 5-shot setting, we compare the 5-shot segmentation results with the average and maximum fusion methods, in which the average and maximum segmentation results of 5 support images are computed, respectively. It can be seen in Table 2 and Fig. 6 that our weighted fusion strategy achieves the best. The samples in the PASCAL-5 dataset are relatively simple, and most of the weights are close to one fifth, so the result of average fusion method is similar to weighted fusion in general.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>i=0</th>
<th>i=1</th>
<th>i=2</th>
<th>i=3</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSLSM</td>
<td>Vgg16</td>
<td>33.6</td>
<td>55.3</td>
<td>40.9</td>
<td>33.5</td>
<td>43.9</td>
</tr>
<tr>
<td>SG-One</td>
<td>Vgg16</td>
<td>40.2</td>
<td>58.4</td>
<td>48.4</td>
<td>38.4</td>
<td>46.3</td>
</tr>
<tr>
<td>PAnet</td>
<td>Vgg16</td>
<td>42.3</td>
<td>58.0</td>
<td>51.1</td>
<td>41.2</td>
<td>48.1</td>
</tr>
<tr>
<td>FWBFS</td>
<td>Vgg16</td>
<td>47.0</td>
<td>59.6</td>
<td>52.6</td>
<td>48.3</td>
<td>51.9</td>
</tr>
<tr>
<td>Ours</td>
<td>Vgg16</td>
<td>50.9</td>
<td>63.0</td>
<td>53.6</td>
<td>49.6</td>
<td>54.3</td>
</tr>
</tbody>
</table>

Table 3: Comparison with SOTA for 1-shot segmentation under the mean-IoU metric on PASCAL-5 dataset. Bold fonts represent the best results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>i=0</th>
<th>i=1</th>
<th>i=2</th>
<th>i=3</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSLSM</td>
<td>Vgg16</td>
<td>35.9</td>
<td>58.1</td>
<td>42.7</td>
<td>39.1</td>
<td>43.8</td>
</tr>
<tr>
<td>SG-One</td>
<td>Vgg16</td>
<td>41.9</td>
<td>58.6</td>
<td>48.6</td>
<td>39.4</td>
<td>47.1</td>
</tr>
<tr>
<td>PAnet</td>
<td>Vgg16</td>
<td>51.8</td>
<td>64.6</td>
<td>59.8</td>
<td>46.5</td>
<td>55.7</td>
</tr>
<tr>
<td>FWBFS</td>
<td>Vgg16</td>
<td>50.9</td>
<td>62.9</td>
<td>56.5</td>
<td>50.1</td>
<td>55.1</td>
</tr>
<tr>
<td>Ours</td>
<td>Vgg16</td>
<td>52.5</td>
<td>64.8</td>
<td>59.5</td>
<td>51.3</td>
<td>57.0</td>
</tr>
</tbody>
</table>

Table 4: Comparison with SOTA for 5-shot segmentation under the mean-IoU metric on PASCAL-5 dataset. Bold fonts represent the best results.

5.4 Comparison with SOTA

To better assess the overall performance of our network, we compare it to other methods (OSLSM [Shaban et al., 2017], SG-One [Zhang et al., 2018], PAnet [Wang et al., 2019], FWBFS [Nguyen and Todorovic, 2019] and CAnet [Zhang et al., 2019a]) on PASCAL-5 and COCO-20 datasets.

PASCAL-5

We train two types of SST (Self-Supervised Tuning) models with VGG-16 and ResNet-50 backbones (we call them SST-vgg and SST-res model) on PASCAL-5 dataset. It can be seen in Table 3 that our SST-vgg model surpasses the best existing method over two percentage points in 1-shot setting, and the SST-res model yields 2.2 points improvement, which is even 1.4 points higher than the method with ResNet-101 backbone. Under the setting of 5-shot, our SST-vgg and SST-res models significantly increase by 1.9 and 1.5 points as shown in Table 4, respectively. These comparisons indicate that our method boosts the recognition performance of few-shot segmentation.

COCO-20

To prove that our method also has good generalization performance on larger datasets, we compare our method with others which recently report results on the COCO-20 dataset as shown in Table 5. Obviously, the average results of our method surpass other best methods by 1 and 1.6 points under 1-shot and 5-shot setting, respectively.

6 Conclusion

In this paper, a self-supervised tuning framework is proposed for few-shot segmentation. The category-specific semantic constraint is provided by the self-supervised inner loop and utilized to adjust the distribution of latent features across different episodes. The resulting auxiliary loss is also introduced into the outer loop of training process, achieving faster convergence and higher scores. Extensive experiments on benchmarks show that our model is superior in both performance and adaptability compared with existing methods.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB130092, the National Natural Science Foundation of China (NSFC) under Grants 61872327, the Fundamental Research Funds for the Central Universities under Grant WK2380000001 as well as Huawei Technologies Co., Ltd.
References


[Dong and Xing, 2018] Nanqing Dong and Eric P Xing.


