Boosting Few-Shot Segmentation via Instance-Aware Data Augmentation and Local Consensus Guided Cross Attention

Anonymous CVPR submission

Paper ID 8870

Abstract

Few-shot segmentation aims to train a segmentation model that can fast adapt to a novel task for which only a few annotated images are provided. Fine-tuning the classification layer of a deep segmentation network pre-trained on diverse base classes is a strong baseline for few-shot segmentation. However, the classification layer optimized with sparsely annotated samples is often biased and exhibits poor generalization capacity. This paper proposes a straightforward solution to alleviate this problem. Specifically, we introduce an instance-aware data augmentation (IDA) strategy which augments the support images based on relative sizes of the target objects. The proposed IDA effectively increases the support set's diversity and promotes the distribution consistency between support and query images. On the other hand, the large visual difference between query and support images may hinder the knowledge transfer and cripple the segmentation performance. To cope with this challenge, we introduce the local consensus guided cross attention (LCCA) to align the query feature with support features based on their dense correlation, further improving the model's generalizability to the query image. The significant performance improvements on the standard few-shot segmentation benchmarks PASCAL-5^i and COCO-20^i verify the efficacy of our proposed method.

1. Introduction

Semantic segmentation has achieved tremendous success in recent years, thanks to the rapid development of deep learning algorithms. Despite the effectiveness of the deep learning models, they rely heavily on large amounts of annotated samples from well-established datasets. However, collecting sufficient dense annotated samples is both time-consuming and costly, especially for dense prediction tasks such as semantic segmentation and instance segmentation. To cope with this challenge, few-shot segmentation (FSS) aims to learn a generic segmentation model that can quickly adapt to novel classes in low-data regimes.

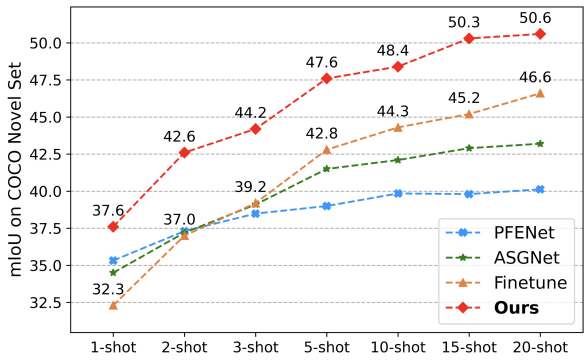

Figure 1. FSS performance (mIoU) on COCO-20^i as shot number increases. All models adopt a ResNet-50 backbone.

Even though the literature on FSS exhibits great diversity, they broadly fall into two categories. The first line of work adopts a prototype learning paradigm, which uses masked global average pooling [31] or clustering [12, 29] to generate one or more prototypes and perform query object inference based on the dense comparison with the prototypes. On the other hand, several other works [2, 16] have adopted a two-stage fine-tuning based training strategy, which only fine-tunes the last layer *i.e.*, the classification layer of a segmentation model pre-trained on the data abundant base classes. Figure 1 compares some popular few-shot segmentation algorithms under different few-shot settings. As is shown, the prototype-based approaches [12, 26] achieve great performance in the extreme low-shot case, while their performance saturates quickly beyond the standard 1- or 5-shot settings. On the contrary, fine-tuning based learning paradigm can sufficiently utilize the increasing number of support samples and reveals superiority in 5- or 10-shot cases, while it is significantly outperformed by its counterpart in the 1-shot case due to over-fitting.

Compared to the fine-tuning-based methods [2, 16], which leverage the overall support set to optimize the classifier for separating pixels of different categories, the

prototype-based approaches [15, 21, 26, 28] perform correlation learning between the query and support images, fully utilizing the feature similarity of each support-query pair. We conjecture that the direct correlation learning between the support-query image pair is the key ingredient that helps the prototype based approaches to excel at extreme low-shot settings. It is, therefore, natural to ask whether it is possible to improve the fine-tuning based approaches by incorporating the direct correlation between query and support images. In this paper, we answer this question affirmatively.

In particular, we recognize that the classification layer fine-tuned on a few annotated samples inevitably overfits the support images. Therefore, instead of directly making inferences on the query object based on the query feature, we exploit the dense correlation between support and query image pairs to align the query feature with the support feature and perform classification based on the aligned query feature. However, the pixel-wise correlation between the two images can be very noisy [23] due to the large visual difference, which undermines the reliability of the cross attention module. Inspired by recent work on semantic correspondence [8, 13, 17, 23], we refine the dense correlation between the two images using local consensus constraints. Then we perform cross attention (Local Consensus guided Cross Attention) based on the enhanced correlation map to achieve feature alignment between the image pair. Incidentally, by integrating the local consensus guided cross attention module into the two-stage fine-tuning based training framework, our proposed method, dubbed Local Consensus guided Cross Attention Network (LC-CAN), exhibits strong generalizability to the query images.

In addition, we introduce a novel data augmentation mechanism to further alleviate the model over-fitting in fine-tuning based approaches. In ordinary supervised learning with sufficient labeled samples, the training data contains objects of different sizes and scales, whose distribution is relatively consistent with the testing data. Notably, under few-shot settings, the model is fine-tuned based on just one or few support images. Therefore the support set tends to form a biased representation of the ground truth distribution from which the test cases are sampled during evaluation. We propose Instance-aware Data Augmentation (IDA) to remedy this problem. As illustrated in Figure 2, IDA is implemented in an instance-aware manner: we first examine the relative sizes of the target objects in the support image, based on which an appropriate augmentation method is chosen to crop or downsize the image. The key idea behind IDA is that we make the model exposed to support images of different scales while promoting distribution consistency between support and query images.

The proposed IDA augmentation strategy and LC-CAN framework cooperate in a synergistic fashion to further boost the model's segmentation performance on query images. The improvements on standard few-shot segmentation benchmarks of PASCAL-5^i and COCO-20^i verify the efficacy of our proposed method.

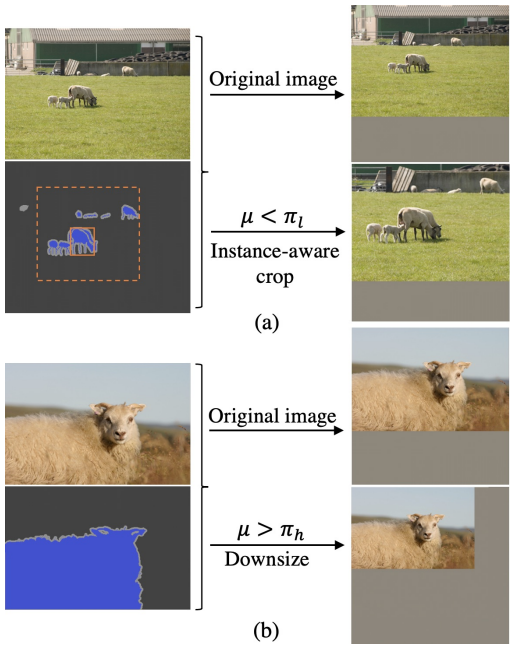

Figure 2. Visualization of instance-aware data augmentation. Original images are resized (while preserving the aspect ratio) and padded with grey pixels to the input image size. Given the foreground ratio μ, we perform (a) instance-aware cropping when $\mu < \pi_l$ and (b) image downsizing when $\mu > \pi_h$. The solid line in (a) is the bounding box of the largest target object, and the dashed line represents the cropping window.

2. Related Work

Few-shot Classification. Few-shot learning (FSL) is a challenging research problem that aims to learn transferable knowledge that can be generalized to new novel classes with a few labeled samples. Existing approaches for FSL can be loosely organized into the following two families. Gradient-based approaches [5, 22] aim to optimize the gradient descent procedure to equip the learner with a good initialization, update direction, and learning rate for fast adaption to a new task. Metric-learning approaches [9, 25, 27, 30] have focused on learning generalizable feature embeddings and use these embeddings on simple classifiers such as nearest neighbor rules. Several recent works [3, 10] suggest that feature extractors trained with standard cross entropy loss on base classes can generate powerful embeddings for new downstream tasks and often outperform its counterpart trained with a meta-learning paradigm. In the same spirit, our work adopts a pretraining stage to learn a task-independent feature embedding network.

Few-shot Segmentation. Few-shot segmentation (FSS) is a natural application of FSL in dense prediction tasks, and it has attracted considerable attention after the pioneering work OSLSM [24]. Most of the recent works in FSS adopt the prototype learning paradigm, which utilizes the prototype extracted from support samples to facilitate query object inference. Notably, CANet [31] applies masked global average pooling on support images for prototype learning and conducts dense comparisons between the prototype and the query features. Instead of using a single prototype, ASGNet [12] and PMMs [29] cluster foreground pixels of the support image to multiple prototypes to account for the intra-class variation and provide more accurate guidance on query image inference. Several recent works [2, 16] argue that the feature extractor of a deep segmentation model pre-trained on base classes is sufficiently generalizable to unseen classes. Instead of meta-training the feature extractor, they freeze the feature extractor pre-trained on base classes and focus on fine-tuning the classification layer for adaptation to new tasks. Our work adopts a similar fine-tuning based training scheme. In addition, we introduce a novel data augmentation strategy to alleviate the over-fitting problem when training the classifier in a low-data regime.

Semantic Correspondence. Semantic correspondence aims to find correspondences between semantically similar images under challenging degrees of variations [1, 6]. Recent approaches build the correlation map based upon feature representations extracted from convolutional neural networks pre-trained on image classification tasks. An emerging trend is to employ 4D convolutions [13, 17, 23] on the dense correlation map to identify spatially consistent matches with the local match-to-match consensus constraint. In addition, some recent approaches [19, 20, 32] for semantic correspondence show that combining features at different semantic levels can help to generate reliable features representation and further improve the matching accuracy. Inspired by existing works in semantic correspondence, we also exploit high dimensional convolution and multi-level features to construct and refine the affinity map between support and query features, which is utilized to guide the feature alignment between the image pair.

3. Problem Formulation

We follow the standard setup and annotations in few-shot semantic segmentation [24]. Specifically, we are given two datasets $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ with disjoint category sets $\mathcal{C}_{base}$ and $\mathcal{C}_{novel}$ respectively, where $\mathcal{C}_{base} \cap \mathcal{C}_{test} = \emptyset$. The goal is to learn a segmentation model from the base dataset $\mathcal{D}_{base}$ with sufficient annotated samples so that the model can generalize well on new tasks sampled from the novel classes. Following the common episodic training protocol, we sample a series of episodes from $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ to simulate the few-shot scenario. Particularly, under K-shot setting, each episode is composed of a small support set $\mathcal{S} = \{(x_k^s, m_k^s)\}_{k=1}^K$ and a query set $\mathcal{Q} = \{(x^q, m^q)\}$, where $x^{\cdot}$ and $m^{\cdot}$ represent a raw image and its corresponding binary mask for a specific category. During training, the model is optimized based on the training episodes sampled from $\mathcal{D}_{base}$ to learn a mapping from $\mathcal{S}$ and x^q to a prediction m^q. At inference, we evaluate the few-shot segmentation performance on test episodes sampled from $\mathcal{D}_{novel}$.

4. Methodology

In this section, we first present our overall framework, which adopts a two-stage fine-tuning training paradigm. Then we introduce the instance-aware data augmentation (IDA) strategy and local consensus guided cross attention (LCCA) for improving the model's generalizability.

4.1. Overall Framework

A semantic segmentation model typically comprises a CNN encoder, a CNN decoder, and a simple classifier. Given an input image x, the CNN encoder $f_x = E_\phi(x)$ gradually reduces the feature map resolution and captures higher semantic information. The decoder module $z_x = D_\psi(f_x)$ aggregates multi-scale features and recovers the spatial information. Then the output embedding z_x from the decoder module is directly passed to a pixel-wise classifier $\hat{m} = p_\theta(z_x)$ to separate pixels from different categories.

The key objective in meta-learning is to learn a transferable feature embedding network that generalizes to any new task. Several works [3,10] in few-shot classification showed that feature extractor pre-trained with standard cross entropy loss on base classes generates powerful embeddings for downstream tasks and often outperforms its counterpart trained with meta-learning paradigm. Following these findings, we separate the feature representation learning and classifier training and adopt a two-stage training scheme.

Model Pre-training. In the first stage, we train the feature extraction network (*i.e.*, the encoder and decoder) on the whole base dataset $\mathcal{D}_{base}$. Specifically, we use PSPNet [33] as our backbone segmentation model, which is trained with standard cross entropy supervision. The training details are given in section 5.2.

New Task Adaptation. In the second stage, we fine-tune the pre-trained segmentation model for adaptation to new tasks. Specifically, we keep the encoder and decoder backbone frozen and train the classifier only.

To reduce the inductive bias of the classifier optimized on sparsely annotated samples, we present instance-aware data augmentation in section 4.2. In addition, we introduce the local consensus guided cross attention module in section 4.3 to align the query feature with support features and improve the model's generalizability to the query image.

4.2. Instance-Aware Data Augmentation

Data augmentation is a simple way to increase the number of training samples and alleviate model over-fitting. Random resize and crop are effective data augmentation methods in regular segmentation tasks and can help improve the model's generalizability to test images of different scales. Appropriate data augmentations can be even more beneficial under the setting of FSS since the distribution of the limited support images tends to be biased. To alleviate the distribution bias of the support set, the proposed Instance-aware Data Augmentation (IDA) adaptively augments the support image based on the relative size of its target objects.

Given a support image, we first compute the proportion of its foreground area against the overall image size based on its ground truth mask, and we denote the foreground proportion as μ. We compare the relative size of the target object μ with pre-set hyperparameters π_l and π_h to adaptively determine the augmentation method applied to the support image. In our experiments, we set the thresholds π_l and π_h to 0.15 and 0.3 respectively.

When the foreground object is relatively small, *i.e.*, $\mu < \pi_l$, we generate the augmented image using instance-aware crop as illustrated in Figure 2(a). When there are multiple foreground objects, we first identify the largest object which resides in the largest connected component of the foreground area. Then we crop the image to cover the largest foreground object. Particularly, for the foreground object with the bounding box represented by the coordinates of its top-left corner and bottom-right corner (x_0, y_0, x_1, y_1), we create a rectangular cropping window with coordinates $(\frac{x_0}{2}, \frac{y_0}{2}, \frac{x_1+W}{2}, \frac{y_1+H}{2})$, where W and H are the width and height of the original image. The cropped patch is then resized to the input image size while keeping its aspect ratio.

When the foreground object is relatively large in size, *i.e.*, $\mu > \pi_h$, the support image is downsized to create its augmented version. As illustrated in Figure 2(b), we downsize the image using bilinear interpolation with a resizing factor of 0.7. The resulting image is then padded with gray pixel values to the input image size.

On the other hand, if $\pi_l < \mu < \pi_h$, IDA is not triggered, and only the original support image is included in the final support set.

Why not use random augmentation? Random data augmentation has proved effective in regular semantic segmentation tasks. However, only a few support images are available under the setting of FSS. Therefore, augmenting the support set using random data augmentation can drastically change the overall distribution of the support set and may create a bigger distribution discrepancy between support and query images, which will hinder the model's generalization ability to the query image. On the other hand, IDA creates the augmented image adaptively based on the relative size of the target objects. It not only improves the support set's diversity but also helps correct the distribution bias of the support images.

4.3. Local Consensus Guided Cross Attention

Before introducing LCCA, we first revisit the fine-tuning based learning framework. Given a pair of support and query images, $x^s, x^q \in \mathbb{R}^{3\times H\times W}$, we use the pre-trained backbone encoder to generate intermediate feature maps:

$$f_1^s, \cdots, f_L^s = E_\phi(x^s), \tag{1}$$

$$f_1^q, \cdots, f_L^q = E_\phi(x^q), \tag{2}$$

with $f_l^\bullet \in \mathcal{R}^{c_l\times h_l\times w_l}(l = 1, \cdots, L)$ being the intermediate output of the l^{th} encoder block. And we use the pre-trained backbone decoder to extract the final feature embeddings:

$$z^s = D_\psi(f_L^s) \in \mathbb{R}^{c\times h\times w}, \tag{3}$$

$$z^q = D_\psi(f_L^q) \in \mathbb{R}^{c\times h\times w}. \tag{4}$$

The feature embedding $z^\bullet$ is directly passed to a binary classifier $\hat{m} = p_\theta(z^\bullet)$ to separate foreground and background pixels. The classifier p_θ is trained with standard cross-entropy loss on the support set.

Under few-shot settings, the learned classifier p_θ tends to overfit support images. To improve the model's generalizability to the query image, our proposed Local Consensus guided Cross Attention Network (LC-CAN) aligns the query feature z^q with support feature z^s and makes the final prediction based on the aligned query feature. Its overall diagram is illustrated in Figure 3. The Local Consensus Guided Cross Attention (LCCA) is the key component in LC-CAN to perform feature alignment between support and query images using their dense correlation. Particularly, the proposed LCCA module consists of the following three components.

Local Self-Attention. Before constructing the cross-affinity between the support and query image pair, we first enhance their feature representation using local self-attention to capture more reliable local contextual information. Unlike global attentions, we only compute self-attention within the local window of each pixel. Particularly, given a pixel $f_{ij} \in \mathbb{R}^c$, we first extract its $k \times k$ neighborhood region $\mathcal{N}_k(i, j)$ which is centered around f_{ij}. We pass pixel f_{ij} and its neighborhood pixels $f_{ab}\big(a, b \in \mathcal{N}_k(i, j)\big)$ to linear transformations to get the query $q_{ij} = W_Q f_{ij}$, keys $k_{ab} = W_K f_{ab}$, and values $v_{ab} = W_V f_{ab}$. The local self-attention computes the output at location ij as:

$$\tilde{f}_{ij} = f_{ij} + G\Big(\sum_{a,b\in\mathcal{N}_k(i,j)} \text{softmax}_{ab}(q_{ij}^T k_{ab}) v_{ab}\Big) \tag{5}$$

where softmax_{ab} denotes the softmax computed across all pixels in neighborhood $\mathcal{N}_k(i, j)$ and G is a transformation

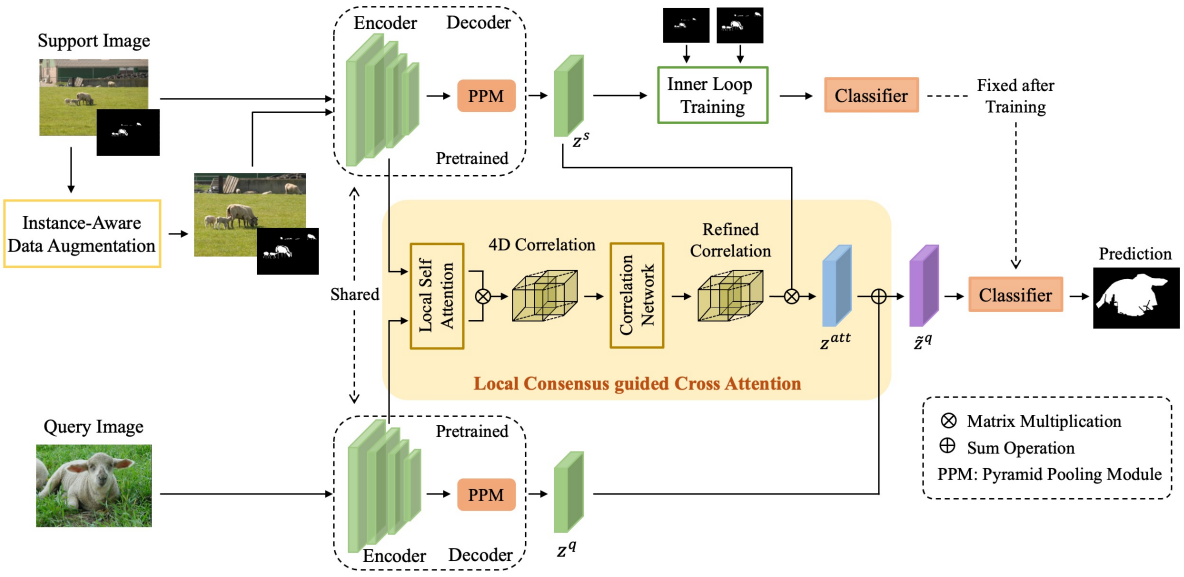

Figure 3. Overview of LC-CAN, which is trained in two stages. In the first stage, the backbone encoder and decoder are pre-trained on base classes. In the second stage, we meta-learn the LCCA module in an episodic manner. At inference time, we train the classifier with the IDA-augmented support set and then pass the LCCA-aligned query feature to the learned classifier for query mask prediction. Note that LCCA module is based on features from multiple intermediate layers and the diagram only illustrates one layer for simplicity.

function implemented with 1×1 convolution. We set k to 3 in our experiments, meaning each pixel only attends to the pixels in its 3×3 neighborhood area.

Correlation Network. We then build the correlation map between support and query images based on the refined feature maps. Particularly, a pair of intermediate features f_l^q and f_l^s is first passed to the local self-attention to get refined features $\tilde{f}_l^q$ and $\tilde{f}_l^s$. Then we compute the pixel-wise cosine similarities between $\tilde{f}_l^q$ and $\tilde{f}_l^s$, and obtain a 4D correlation map $c_l \in \mathbb{R}^{h_l \times w_l \times h_l \times w_l}$ with

$$c_l(i,j,a,b) = \frac{\langle \tilde{f}_l^q(i,j), \tilde{f}_l^s(a,b) \rangle}{\|\tilde{f}_l^q(i,j)\| \cdot \|\tilde{f}_l^s(a,b)\|}, \tag{6}$$

where (i,j) and (a,b) denote 2-dimensional spatial positions of the query and support features respectively. The correlation maps obtained from different layers $\{c_l\}_{l=1}^L$ are then stacked together along the channel dimension after bilinear interpolation to the size of $h \times w \times h \times w$, resulting in the final multi-channel correlation map $c \in \mathbb{R}^{L \times h \times w \times h \times w}$.

Due to the strong appearance difference between the image pair, great majority of the information in the correlation map corresponds to noisy matching. Inspired from [23, 32], we adopt the neighborhood consensus module to refine the correlation map:

$$\tilde{c} = H(c) \in \mathbb{R}^{h \times w \times h \times w}, \tag{7}$$

where the correlation network H is composed of a sequence of multi-channel 4D convolution units to refine the correlation map using local consensus constraints. We adopt the same model structure for H as [23] except that our model takes a multi-channel correlation map as input to exploit diverse levels of feature representations. We refer the reader to [23] for more details. To reduce the computational burden caused by high-dimensional convolutions, we replace the original 4D convolution with a lightweight center-pivot 4D convolution [18].

Attentive Feature Alignment. With the refined affinity map between the image pair, we align the query feature with support feature using cross attention and compute the attention feature at position ij as:

$$z_{ij}^{att} = \sum_{ab} \text{softmax}_{ab}(\tau \tilde{c}_{ijab}) z_{ab}^s \tag{8}$$

where $\tilde{c}_{ijab}$ is the refined correlation score between query pixel at position ij and support pixel at position ab, and z_{ab}^s is the support feature obtained from the decoder backbone. The temperature τ is a hyper-parameter to control the softness of the attention, and it is set to 10 in our experiments.

The final aligned query feature is then computed as:

$$\tilde{z}^q = (1-\gamma) z^q + \gamma z^{att}, \tag{9}$$

where γ is a hyper-parameter balancing the two terms and is set to 0.1 in all experiments. The aligned query feature $\tilde{z}^q$ is then passed to the classifier p_θ to get the final mask prediction $\hat{m}^q$.

To provide direct supervision on LCCA, we pass the intermediate feature z^{att} directly to the classifier p_θ. The dice loss between $p_\theta(z^{att})$ and the ground-truth query mask m^q is used as the meta-training objective.

4.4. K-Shot Setting

The proposed LC-CAN can be easily extended to K-shot setting. Particularly, given K support image-mask pairs $\mathcal{S} = \{(x_k^s, m_k^s)\}_{k=1}^K$ and a query image x^q, we first train the classifier p_θ based on the support set. We then align the query feature with K support features through the proposed LCCA module to provide K aligned query features, which are averaged along the pixel dimension to get the final query feature. This query feature is then passed to the classifier p_θ to obtain the final query mask prediction.

4.5. Incorporating IDA and LC-CAN

It's worth mentioning that the proposed LC-CAN, combined with IDA, is synergistic in boosting the model performance. Remarkably, by incorporating IDA, the classifier p_θ is optimized based on the support set augmented by IDA. And instead of aligning the query feature with the original support image, we align it with the augmented version, which is more balanced in terms of the foreground proportion μ. In Section 5.4, we conduct extensive experiments to verify the benefit of incorporating IDA and LC-CAN.

5. Experiments

5.1. Datasets

We evaluate the performance of our approach on two widely-used FSS benchmarks, namely PASCAL-5^i [24] and COCO-20^i [21]. PASCAL-5^i is built from PASCAL VOC 2012 [4] and contains 20 object categories that are evenly divided into 4 folds. COCO-20^i is constructed from MS-COCO [14] and consists of mask-annotated images from 80 object classes divided into 4 folds. For both datasets, the model is trained on 3 folds and tested on the remaining one in a cross-validation manner.

5.2. Implementation Details

Pre-training. We build our model based on PSPNet [33] with ResNet-50 and ResNet-101 [7] as backbones. We adopt the standard supervised learning to train the feature extractor (*i.e.*, the encoder and decoder) on each fold of the FSS dataset, which consists of 16/61 classes (including background) for PASCAL-5^i/COCO-20^i. We train the model for 100 epochs on PASCAL-5^i and 20 epochs for COCO-20^i with cross-entropy loss as the objective function. To update the parameters, we use SGD optimizer with an initial learning rate of $2.5\text{e}{-3}$ and cosine learning rate decay, and the momentum is set to 0.9, and weight decay to $1\text{e}{-4}$. We set the batch size to 12 and the input image size to 473. Label smoothing is used with the smoothing parameter $\epsilon = 0.1$. As for data augmentations, we only use random mirror flipping.

Episodic Training. After the pre-training stage, we adopt an episodic training procedure to meta-learn the LCCA module. Note that with ResNet as the encoder backbone, LCCA computes 2 layer-wise correlation maps from the output features of block3 and block4, which are further refined to get the final matching score.

In this stage, we organize the training data of base classes $\mathcal{D}_{base}$ into episodes, each including a support set and query set from a randomly sampled class. We conduct the meta-training in a two-loop manner [25]. The inner loop trains a classifier p_θ for the selected class for 100 iterations on the support set with cross-entropy supervision. SGD optimizer with learning rate $1\text{e}{-1}$ is used. The outer loop trains the proposed LCCA with dice loss supervision on the query images. The outer loop is trained with SGD optimizer on PASCAL-5^i for 5 epochs and COCO-20^i for 1 epoch. The learning rate is set to $1\text{e}{-3}$ on both datasets.

Evaluation Metrics. For evaluation metrics, we adopt the widely used mean Intersection over Union (mIoU), which is computed by averaging the IoU values of all classes in a fold. Following previous works, for each fold, the model is validated on 1000 randomly sampled episodes.

5.3. Experiment results

In Table 1 and 2, we evaluate LC-CAN under standard 1-shot and 5-shot settings on PASCAL-5^i and COCO-20^i datasets. Our final implementation of LC-CAN has integrated IDA for support set augmentation. We present the performance of LA-CAN along with a two-stage fine-tuning baseline. Particularly, after pre-training, the baseline approach goes straight to the testing stage and trains the classifier on the given support set *without* using IDA for data augmentation or LCCA for query feature alignment. Our proposed method shows a significant performance gain over the baseline, which verifies the effectiveness of the proposed IDA and LC-CAN. It also outperforms existing state-of-the-art approaches by a sizable margin, especially under 5-shot settings. For example, with a ResNet-50 backbone, the 5-shot mIoU gains on COCO-20^i are 4.7% compared to the baseline and 0.7% compared to the best competitor HSNet [18].

5.4. Ablation study

We conduct extensive ablation studies to investigate the impact of major components in our model. The experiments in this section are performed on COCO-20^i dataset using the ResNet-50 backbone unless specified otherwise.

Ablation on IDA. To inspect the impact of the proposed IDA, we compare the IDA augmented fine-tuning baseline, denoted by `Baseline+IDA`, where we use

Backbone	Method	1-shot					5-shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	mean
ResNet-50	CANet [31]	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
	PFENet [26]	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
	CWT [16]	56.3	62.0	59.9	47.2	56.4	61.3	68.5	68.5	56.6	63.7
	RePRI [2]	60.2	67.0	**61.7**	47.5	59.1	64.5	70.8	71.7	60.3	66.8
	DCP [11]	63.8	70.5	61.2	55.7	62.8	67.2	**73.2**	66.4	64.5	67.8
	HSNet [18]	**64.3**	**70.7**	60.3	**60.5**	**64.0**	**70.3**	**73.2**	67.4	**67.1**	69.5
	Baseline	55.0	62.5	60.6	47.5	56.4	61.5	70.7	72.5	59.7	66.1
	LC-CAN (ours)	60.0	65.0	**61.7**	52.8	59.9	67.1	72.8	**74.3**	64.0	**69.6**
ResNet-101	PFENet [26]	60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
	CWT [16]	56.9	65.2	61.2	48.8	58.0	62.6	70.2	68.8	57.2	64.7
	RePRI [2]	59.6	68.6	**62.2**	47.2	59.4	66.2	71.4	67.0	57.7	65.6
	HSNet [18]	**67.3**	**72.3**	62.0	**63.1**	**66.2**	**71.8**	**74.4**	67.0	**68.3**	**70.4**
	Baseline	56.3	63.1	59.2	47.7	56.6	63.1	70.5	70.0	59.0	65.7
	LC-CAN (ours)	62.3	67.3	61.2	53.1	61.0	69.3	73.5	**72.8**	64.7	70.1

Table 1. Comparison with state-of-the-art methods on PASCAL-5^i in terms of mIoU. The 1st / 2nd methods are **bold** / underlined.

Backbone	Method	1-shot					5-shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	mean
ResNet-50	CWT [16]	32.2	36.0	31.6	31.6	32.9	40.1	43.8	39.0	42.4	41.3
	RePRI [2]	31.2	38.1	33.3	33.0	34.0	38.5	46.2	40.0	43.6	42.1
	DCP [11]	**40.9**	**43.8**	**42.6**	38.3	**41.4**	45.8	49.7	43.7	46.6	46.5
	HSNet [18]	36.3	43.1	38.7	**38.7**	39.2	43.3	**51.3**	**48.2**	45.0	46.9
	Baseline	30.5	34.8	30.6	33.2	32.3	42.6	45.3	40.4	43.1	42.9
	LC-CAN (ours)	35.2	39.6	37.2	38.3	37.6	**46.0**	50.8	45.4	**48.1**	**47.6**
ResNet-101	CWT [16]	30.3	36.6	30.5	32.2	32.4	38.5	46.7	39.4	43.2	42.0
	PFENet [26]	36.8	41.8	38.7	36.7	38.5	40.4	46.8	43.2	40.5	42.7
	HSNet [18]	37.2	**44.1**	**42.4**	**41.3**	**41.2**	45.9	53.0	**51.8**	47.1	49.5
	Baseline	33.0	38.7	32.2	34.7	34.7	42.2	49.3	43.7	43.6	44.7
	LC-CAN (ours)	**37.7**	42.7	40.0	39.8	40.1	**47.1**	**54.4**	48.6	**50.1**	**50.0**

Table 2. Comparison with state-of-the-art methods on COCO-20^i in terms of mIoU. The 1st / 2nd methods are **bold** / underlined.

Variants	1-shot mIoU				
	Fold-0	Fold-1	Fold-2	Fold-3	Mean
Baseline	30.5	34.8	30.6	33.2	32.3
Baseline+RDA	31.3	34.6	29.8	32.3	32.0
Baseline+IDA	**32.5**	**35.7**	**31.9**	**33.6**	**33.4**

Table 3. Ablation studies of IDA under 1-shot setting.

Variants	1-shot mIoU				
	Fold-0	Fold-1	Fold-2	Fold-3	Mean
Baseline	30.5	34.8	30.6	33.2	32.3
LC-CAN	**33.4**	**37.3**	**34.5**	**36.6**	**35.5**
LC-CAN w/o LSA	32.7	36.6	34.0	35.6	34.7

Table 4. Ablation studies of LC-CAN. 'LSA' denotes the local self-attention module in LCCA used for feature enhancement.

IDA to augment the support set in the testing stage and train the classifier based on the augmented support set, and the naive random augmentation approach, denoted by `Baseline+RDA`, where we use random crop and resize to create the augmented support image. As shown in Table 3, our proposed IDA strategy brings 1.1% improvement in mIoU compared to the baseline model, while random augmentation may hurt the model's performance on the query set. As explained in Section 4.2, the potential reason is that our adaptive augmentation strategy would correct the distribution bias in the support set. In contrast, random augmentation may create a bigger distribution discrepancy between the support and query sets. Finally, we guide the readers to the supplementary material for additional analyses on IDA. In Figure 4, we qualitatively analyze the impact of our proposed IDA. Compared to the baseline method, IDA effectively improves the model's generalizability to the query image, especially when the support image is unbalanced in terms of the target object size.

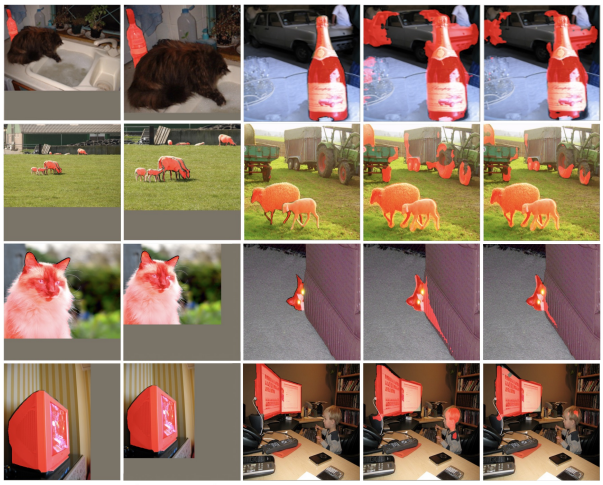

Figure 4. Ablation on instance-aware data augmentation.

Variants	1-shot mIoU				
	Fold-0	Fold-1	Fold-2	Fold-3	Mean
Baseline	30.5	34.8	30.6	33.2	32.3
LC-CAN/both	35.5	39.5	37.1	38.1	37.5
LC-CAN/org	34.5	38.5	36.3	37.1	36.6
LC-CAN/aug	35.2	39.6	37.2	38.3	37.6

Table 5. Incorporating IDA and LC-CAN. The approach adopted in the final implementation is highlighted in gray.

Ablation on LCCA. To study the effectiveness of LCCA, we train LC-CAN *without* any data augmentation. Particularly, in the testing stage, we optimize the classifier p_θ based on the original support set *without* IDA for data augmentation. Then with the learned classifier p_θ, we predict the query mask based on the refined query feature, which is aligned with support images through LCCA. As shown in Table 4, our proposed LC-CAN outperforms the baseline model by a large margin. If we remove the local self-attention in LCCA which is used for feature enhancement, the performance drops by 0.8% as presented in the 3$^{\text{rd}}$ row of Table 4, which indicates that the contextual information around the neighborhood pixels is very important for refining the cross affinity. In addition, LCCA exploits correlation maps $\{c_l\}_{l=1}^{L}$ obtained from multiple layers of the encoder backbone to get more accurate matching correspondence. We provide more details about the impact of individual layer-wise correlation c_l in the supplementary materials.

We visualize the results of LC-CAN compared to the baseline model in Figure 5. When there are large visual differences between the support and query images, LC-CAN can effectively suppress the noisy background area, which is falsely activated in the baseline model.

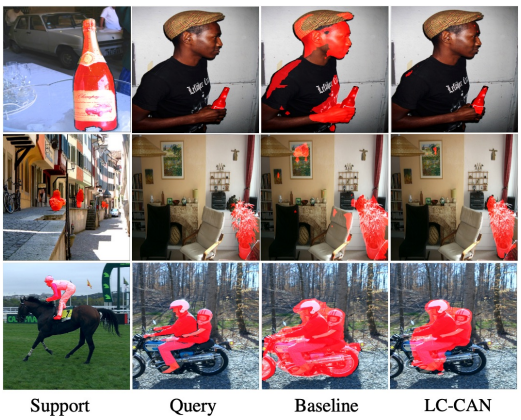

Figure 5. Ablation on local consensus guided cross attention.

Incorporating IDA and LC-CAN. An intuitive way to incorporate IDA with LC-CAN is to align the query feature with both the original support image and its augmented version through LCCA to produce two aligned query features, the average of which is the final query embedding used for classification. We denote this method as `LC-CAN/both`. However, this approach is costly regarding computation overhead and memory consumption. In Table 5, we compare two other strategies with the intuitive method mentioned above. The first strategy is to align the query feature with the original support image only, which is denoted as `LC-CAN/org`. The second strategy is to align the query feature with the augmented support image only, which is denoted as `LC-CAN/aug`. As shown in Table 5, `LC-CAN/aug` achieves the best result despite its lower computation cost compared to `LC-CAN/both`. Aligning the query image with the augmented support image is more beneficial than the other two approaches since the augmented image is more balanced in foreground proportion.

IDA augmentation strategy and the proposed LC-CAN framework cooperate in a synergistic fashion to improve the model's performance. For example, with a ResNet-50 backbone, IDA and LC-CAN bring 1.1% and 3.2% improvement on COCO-20^i in 1-shot mIoU compared to the baseline. And incorporating both methods the final model achieves 5.3% improvement over the baseline.

6. Conclusion

In this paper, we re-evaluate fine-tuning based FSS approaches compared to the prototype learning paradigm. In particular, we incorporate the dense correlation between support and query images to improve the performance of fine-tuning based approaches, especially in extreme low-shot settings. Furthermore, by highlighting the advantages and disadvantages of fine-tuning based methods, we hope to shed some light on the development of new algorithms.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017. 3

[2] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13979–13988, 2021. 1, 3, 7

[3] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019. 2, 3

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2

[6] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1711–1725, 2017. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[8] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8707, 2022. 2

[9] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille, 2015. 2

[10] Steinar Laenen and Luca Bertinetto. On episodes, prototypical networks, and few-shot learning. *Advances in Neural Information Processing Systems*, 34:24581–24592, 2021. 2, 3

[11] Chunbo Lang, Binfei Tu, Gong Cheng, and Junwei Han. Beyond the prototype: Divide-and-conquer proxies for few-shot segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. 7

[12] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8334–8343, 2021. 1, 3

[13] Shuda Li, Kai Han, Theo W Costain, Henry Howard-Jenkins, and Victor Prisacariu. Correspondence networks with adaptive neighbourhood consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10196–10205, 2020. 2, 3

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[15] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer, 2020. 2

[16] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8741–8750, 2021. 1, 3, 7

[17] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2950, 2021. 2, 3

[18] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6941–6952, 2021. 5, 6, 7

[19] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3395–3404, 2019. 3

[20] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020. 3

[21] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 622–631, 2019. 2, 6

[22] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017. 2

[23] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018. 2, 3, 5

[24] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 3, 6

[25] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2, 6

[26] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2, 7

[27] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 2

[28] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *European Conference on Computer Vision*, pages 730–746. Springer, 2020. 2

[29] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 763–778. Springer, 2020. 1, 3

[30] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020. 2

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019. 1, 3, 7

[32] Dongyang Zhao, Ziyang Song, Zhenghao Ji, Gangming Zhao, Weifeng Ge, and Yizhou Yu. Multi-scale matching networks for semantic correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3354–3364, 2021. 3, 5

[33] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3, 6