Adapt Before Comparison: A New Perspective on Cross-Domain Few-Shot Segmentation

Jonas Herzog
Zhejiang University
jherzog@zju.edu.cn

Abstract

Few-shot segmentation performance declines substantially when facing images from a domain different than the training domain, effectively limiting real-world use cases. To alleviate this, recently cross-domain few-shot segmentation (CD-FSS) has emerged. Works that address this task mainly attempted to learn segmentation on a source domain in a manner that generalizes across domains. Surprisingly, we can outperform these approaches while eliminating the training stage and removing their main segmentation network. We show test-time task-adaption is the key for successful CD-FSS instead. Task-adaption is achieved by appending small networks to the feature pyramid of a conventionally classification-pretrained backbone. To avoid overfitting to the few labeled samples in supervised fine-tuning, consistency across augmented views of input images serves as guidance while learning the parameters of the attached layers. Despite our self-restriction not to use any images other than the few labeled samples at test time, we achieve new state-of-the-art performance in CD-FSS, evidencing the need to rethink approaches for the task. Code is available at https://github.com/Vision-Kek/ABCDFSS.

1. Introduction

With a successful Cross Domain Few Shot Segmentation (CD-FSS) algorithm, segmentation could be deployed on any task, regardless of the type of objects to segment and its environment. This paper studies CD-FSS, a task that has emerged recently motivated by the failure of few-shot segmentation (FSS) when test images are fundamentally different from training images.

In general, both tasks aim to segment novel classes in a test (query) image based on a few labeled (support) images. Given this severe knowledge limitation about the novel class, FSS utilizes a base dataset that can provide a larger number of tasks for training. Since train tasks only provide information about base classes, not the novel classes that will appear at test-time, it is considered crucial that the model can generalize from base to novel classes.

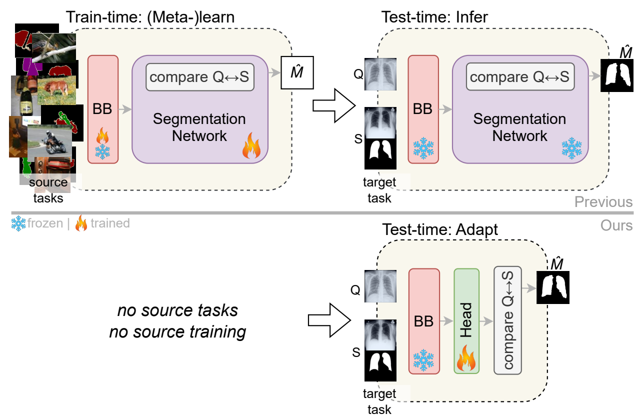

Figure 1. *Top:* Few Shot Segmentation across domains has been addressed by training a deep network on segmentation tasks from a source domain. We demonstrate that its efforts to achieve generalizability during this stage are largely unsuccessful. *Bottom:* In the proposed approach, we entirely forgo such training. Instead, *b*ack*b*one-attached layers (green) adapt features to the target task at test-time.

This becomes substantially more challenging when train and test tasks originate from different domains. Recent approaches for FSS across domains [2, 4, 20, 30, 44] focus on this generalization problem and extend FSS with modules designed to enhance knowledge transfer to unseen target domains. Their learning paradigm and procedure is closely aligned with conventional FSS, not requiring recent popular large models [16]. A single source domain such as PASCAL VOC 2012 [8] supplies source tasks. Learning from the source is either conducted by emulating tasks with episodic meta-learning [4, 20, 30, 44] or by standard supervised learning [2]. A segmentation network is learned on top of a frozen [4, 20] or trainable [44] backbone. Finally, the model is tested on tasks from the target domain. The typical architecture and strategy for this is illustrated in Figure 1. Such approaches rely on similarity-based *comparison* of query and support backbone features in order to locate where the query image matches the support. Inspired by [4], we inspect these similarities. We find that with a significant domain shift, also lower-level intermediate features are not suitable - the discriminability between semantic classes decreases. If the representations for a test task are not discriminative, the subsequent segmentation network is predetermined to fail, regardless its generalization ability acquired during train time. Motivated by this shortcoming, instead of trying to solve the inherently difficult task of learning a generalizable model from a single source domain, we identify *adapting* features to the target task is crucial.

A straightforward solution would be fine-tuning to the test task under utilization of the labeled support set, but it is prone to overfit to the support set [2, 14, 26]. [1] Our solution is a mechanism that relies on embedding consistency within a test task. Different from all previous work, we do not consider any source tasks. We demonstrate that adapting ImageNet pretrained backbone features at test-time is sufficient to achieve superior results. Specifically, we append a small network to each intermediate layer of the backbone. Both query and support images are augmented to obtain multiple views of them. Parameters of the attached layers are found as the optimization of a formulation which enforces both class-agnostic embedding consistency and intra-support class consistency across views. This way we can find features relevant for the current task. After that, we build query-suppport correlation maps by calculating pixel-to-pixel similarities of the task-adapted features. The prediction mask is then simply obtained by a parameter-free aggregation of the multi-layer correlation maps.

- Our research reveals that the current approach of learning a downstream FSS network is still inefficient for CD-FSS. We replace it by tiny adaptors that learn at test-time only, proposing Adapt Before Comparison (ABCDFSS).
- A novel consistency-based contrastive learning scheme can estimate the parameters of our attached layers without overfitting to the support set. Class discriminability in the query feature space improves significantly. Comparing features from shallow and deep layers separately can then provide domain-shift robust prediction masks.
- Our method achieves new state-of-the-art performance on the CD-FSS benchmark and SUIM. Results and experiments highlight the need for our paradigm shift from training a segmentation network to task-adaption.
- Our study points out three issues in current CD-FSS work that must be considered in future work: Source domain conceptualization, evaluation metric and benchmark composition.

[1]Such test-time fine-tuning is not to be confused with train-time fine-tuning [39] and the aforementioned problem to overfit to the base classes [7, 30, 33], which received the primary attention in FSS research.

2. Related Work

Few-Shot Segmentation is mainly addressed by comparing the query feature volume with a representation of support foreground class information. After early approaches with query-support fusion [41, 52], single [7, 54] or multiple [21, 25, 43, 50] prototypes became prevalent for the representation of the support class information. Besides prototype based methods, a more recent branch relies on analyzing pixel-to-pixel correspondences [27, 30, 33, 36, 41], thus avoiding the loss of spatial structure inherently coming with prototyping. The base data structure are the dense correspondences between query and support. Then, either the maximum support correspondence [41], a transformer-style dot product [27, 33, 36, 53], or complex learned schemes [30] are employed to reduce this structure. While most work resorted to the meta-learning scheme, [27] reduced meta-learning to the classifier, [2] trained the classifier at test-time with no meta-learning and [18, 19, 33] combined base- and meta-learning branch. Our test-time learning does not require such strategies. Self-supervised contrastive learning as in [1] has been applied for few-shot segmentation in [35], its dense variants [31, 46] have been proposed for large-scale representation learning. We use the technique for few-shot task adaption instead.

Domain Generalization(DG) and Cross-Domain work under domain shifts where no target domain data is accessible, differentiating them from Domain Adaption (DA). More challenging than DG and DA, in cross-domain few-shot learning (CDFSL) not only the target domain is different from the one seen during training, but also the tasks are novel [45, 49]. While many previous CDFSL methods [9, 10, 29, 42, 44] built upon DG techniques to acquire a task-agnostic network in the base step, our paper makes no DG attempts and focuses on adapting to the novel task in the novel domain instead. Besides fine-tuning [12, 23], incorporating or attaching small task specific adapters to multiple layers of a deep network has been studied for cross-domain classification [22, 38] and object detection [11]. Like in our work, these adapters have also been trained from scratch on the target task in [22, 28]. Unlike these work for classification, we are interested in dense labels and propose attaching tiny networks that can exploit the dense interaction between support and query.

Cross-Domain Few-Shot Segmentation. A few studies in the FSS literature [2, 30, 37] started evaluating their methods under the small domain shift COCO[24]→PASCAL[8]. Subsequently, a small number of work focused explicitly on our task, CD-FSS. RtD [44] employs feature enhancement and stores domain-specific style information which is believed to be domain-specific in a memory which is used to generalize during training and guide during testing. PATNet [20] prepends a transformation module before HSNet [30], leading to more constant prototypes across episodes and domains. The module is suggested to be fine-tunable on the test task, however, from both design and empirical level the focus is stability at train-time, hence the transformation cannot solve the problem of inadequate features of the target task. PMNet [4] proposes a more light-weight architecture based on dense affinity matrices [36] between query and support pixels. One work [26] suggested fine-tuning the backbone on the target task also using the query image, but requires knowledge of target domain unlabeled data. In contrast, we keep the backbone frozen, and assume availability of only one query image as in [2, 20, 30, 44].

Different from all CD-FSS work [4, 20, 26, 30, 44], we do not try to learn a domain-generalizing segmentation network. Our method needs no base-, no meta-learning and no source domain data. There are no learnable parameters other than the task-specific weights learned at test-time.

We adopt CD-FSS as the same problem setting as in RtD [44] and PATNet [20], where access to the target domain is forbidden and classes in the target domain are novel.

3. Method

First, following [20, 30, 36, 41], we use a shared pretrained backbone to extract multi-level features for both support and query images. Secondly, a small network is appended to each intermediate level of the backbone. Keeping the backbone frozen, we train this appended network from scratch on the data available at test-time, i.e. the support and query. Third, given the task adapted features, query pixels that are similar to the support foreground pixels receive a higher foreground score. A coarse prediction map can be obtained this way for each layer. Finally, we fuse the layer-wise prediction maps and threshold and optionally refine to obtain the final segmentation mask.

3.1. Feature Extraction

Following [20, 30, 36], query and support images are fed through a pretrained feature extractor to generate multi-layer feature volumes for each. Due to the structure of the backbone used as feature extractor, the layerwise feature volumes $F^q = \{F_l^q\}_{l=1}^L$ and $F^s = \{F_l^s\}_{l=1}^L$ have different sized dimensions for different l. Deeper layers, indexed with larger l, are smaller in spatial dimensions but larger in the channel dimension. The support mask is bilinearly downsampled to match the corresponding spatial dimension's size, yielding $M^s = \{M_l^s\}_{l=1}^L$.

Our method is based on the consistency across views [48] of the same scene. We augment both query and support geometrically to obtain AUG views of each. The augmented images are forward passed through the feature extractor in the same way as the original images, resulting in their features $\{F^{\tilde{q_a}}, F^{\tilde{s_a}}\}_{a=1}^{AUG}$, where superscripts $\tilde{q_a}$, $\tilde{s_a}$ denote association with the ath augmentation of the original image. After that, we have the augmented features $\{F^{\tilde{q_a}}, F^{\tilde{s_a}}\}_{a=1}^{AUG}$ as well as the original (F^q, F^s).

In our method we want to compare the transformed and original features densely. Therefore, it is required to maintain the pixel-wise correspondences between original and augmented features. We restore the correspondences by backprojecting the augmented features with the inverse of the affine that has been applied during augmentation. For readability it does not receive a new notation. Only the backprojected augmented features are used in the following.

3.2. Attached Adapter

Features from the backbone are meaningful in the domain they have been trained on, e.g. ImageNet. While ImageNet-pretrained weights incorporate a large diversity, for a specific cross-domain few-shot task the embedding space is not optimal. High intra-class distances and low inter-class distances appear to be prominent issue.

We propose to append a small adapter network to the backbone, specifically one to each of its bottlenecks. An image can be forward passed through the backbone, yielding F, and then through our networks $g = (g_1, g_2, ..., g_L)$ to obtain task-adapted features $\hat{F}$:

$$\forall l : \hat{F}_l = g_l(F_l), \tag{1}$$

where F_l represents the intermediate features from the $l-th$ backbone bottleneck.

The small attached networks are trained from scratch on the target query and support set, using self-supervised embedding-alignment and supervised class-alignment. Training is performed independently for each layer l, such that the index l is dropped in the this section readability. Keep in mind, however, that every term is specific to one network g_l.

Reflecting the reduced complexity of the target task compared with ImageNet, the thus generated features are of lower dimensionality, representing distillation of relevant information.

Self-Supervised Embedding Alignment with Dense Contrastive Loss We calculate a contrastive loss between features extracted from augmented and non-augmented images, i.e. views. Dense contrastive learning to match embeddings across views has been proposed for large-scale training [31, 46]. Similar to these works, we have a loss term that enforces dot-product similarity of feature volume $\hat{F}$ and its backprojected view $\hat{F}^{aug}$:

$$\mathcal{L}_{nce} = \frac{1}{HW}\sum_{i=1}^{HW} -\log \frac{\exp(f_i f_i^{aug}/\tau)}{\sum_{j=1}^{HW}\exp(f_i f_j^{aug}/\tau)}, \tag{2}$$

where H, W are the spatial dimensions of both $\hat{F}$ and $\hat{F}^{aug}$, from which respective feature vectors f, f^{aug} are extracted

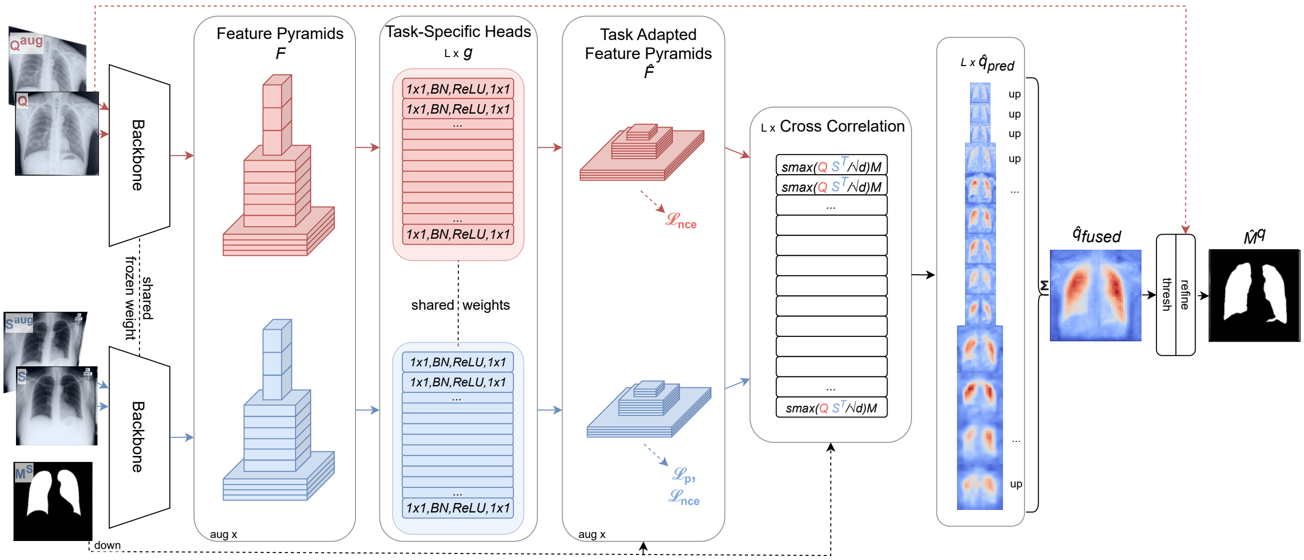

Figure 2. Overview of proposed method: Query (red) and support (blue) images are augmented to generate views of them. Original image and views are fed separately through a frozen backbone as well as our attached task-specific heads to generate a lower-dimensional feature pyramid. The task-specific networks are trained to maximize intra-level consistency across views. Adapted features are then densely compared in the cross-correlation module. Finally, the level-wise prediction maps are aggregated, thresholded and refined to generate a binary query foreground class prediction.

using a position index such as i or j. $\tau = 0.5$ is a temperature as in [48]. The enumerator measures the similarity of a positive pair, while the denominator aggregates the similarities of all possible pairs. Positive pairs are defined as the feature vectors $\in \hat{F} \times \hat{F}^{aug}$ which have the same index. Complementary, negative pair partners for a vector $f_i \in \hat{F}$ are all vectors $f_j^{aug} \in \hat{F}^{aug}, i \neq j$.

For each original image and each of its views, $\mathcal{L}_{nce}$ is calculated independently: Each of the pairs $(\hat{F}^q, F^{\tilde{q}_1}), (\hat{F}^q, F^{\tilde{q}_2}), ..., (\hat{F}^q, F^{qA\tilde{U}G})$ generates one loss value when plugged into Eq. 2 for $(\hat{F}, \hat{F}^{aug})$. Equally, multiple $\mathcal{L}_{nce}$ are obtained for the support set. We average the losses representing embedding discrepancy across views separately for query and support, yielding L^q_{nce} and L^s_{nce}.

Complementing the pairwise-correspondence based L^q_{nce} and L^s_{nce}, a regularizer is added that acts globally on a feature map. It ensures consistent statistics and penalizes differences in mean and variance of a feature map $\hat{F}_c$:

$$\mathcal{L}_{stat} = \frac{1}{C} \sum_{c=1}^{C} |stat(\hat{F}_c) - stat(\hat{F}_c^{aug})|, \tag{3}$$

where C is the number of channels and $stat$ yields the statistic of a feature map as a scalar. This term is calculated for both mean, and variance and then added to the dense contrastive loss, so that we obtain $L^q = \mathcal{L}^q_{nce} + \mathcal{L}^q_{\mu} + \mathcal{L}^q_{var}$ and, by equally but separately calculating $\mathcal{L}_{nce}$ and $\mathcal{L}_{stat}$ for the support features, L^s.

Class Alignment with Global Contrastive Loss While the previous self-supervised loss does not account for class labels, we introduce the class-aware contrastive prototype loss. Intuitively, the class prototypes between different views of the same scene should be identical as semantic information is equal. A contrastive loss motivated by this has been proposed in [44]. In contrast to their usage, our goal is not generalization across domains but generalization *within* the target domain. Hence, we adopt the term as a support-supervised loss. The same image pairs and extracted features from Sec. 3.1 are used. The formulation itself is then

$$\mathcal{L}_p = -\log \frac{\exp(c(p_f, p_f^{aug}))}{\exp(c(p_f, p_f^{aug})) + \exp(c(p_f, p_b^{aug}))}. \tag{4}$$

Foreground prototypes p_f and background prototypes p_b are obtained by global average pooling [7] of the respective feature volume leveraging the support masks. Similarities are calculated by cosine similarity $c(\cdot)$.

Again, we obtain one $\mathcal{L}_p$ for each augmentation, which are subsequently averaged. For k-shot with $k > 1$, prototypes are not calculated k times. Instead, same-class feature vectors are collectively averaged [7] when producing class-prototypes.

Since it receives supervision from task-relevant class labels and penalizes low intra-foreground similarity and high inter-class similarity, $\mathcal{L}_p$ should be the main contributor to learn semantically significant features, while the label agnostic self-consistency based $\mathcal{L}^q$ and $\mathcal{L}^s$ constrain the solution space.

For a specific layer l in the pyramid, our attached network g_l is then optimized on the combined loss

$$\mathcal{L} = L^q + L^s + \mathcal{L}_p. \tag{5}$$

We observe their contribution to the gradients is balanced and hence not introduce weights.

3.3. Dense Comparison

We calculate the similarity between query and support features to predict the foreground probabilities of query pixels. With the task-adapted features from Eq. 1, measuring similarities between the feature representations $\hat{F_l^q}$ and $\hat{F_l^s}$ is now more semantically meaningful.

We observe dense feature comparison is superior over prototyping for our method and hence adopt the transformer-style query-support-cross-attention weighted mask aggregation from [36] to generate a query correlation map $\hat{q}_{pred_l}$ for each layer. Flattening spatial dimensions in query feature, support features and support masks yields $Q = \hat{F_l^q}, K = \hat{F_l^s}, V = M_l^s$ for

$$\hat{q}_{pred_l} = softmax(QK^T/\sqrt{d})V, \tag{6}$$

where d is the size of the channel dimension of Q and K, i.e. the dimension over which the dot product is taken. In [36], positional encoding and a linear projection is used for generating Q and K^T from the feature volumes F_l^q, F_l^s in order to match the transformer architecture. Because we learned our own head g_l to obtain $\hat{F_l}$ from F_l, in Eq. 6 we directly calculate the dot product between adapted query and support feature volumes. The term does allow Q and K to have different spatial dimensions. To extend it to k-shot, we follow [36] to concatenate support images and masks along the spatial dimension, such that Q is typically of shape $(H \cdot W \times C)$ and K of shape $(H \cdot W \cdot k \times C)$.

3.4. Segmentation

Given the layer-wise coarse query prediction maps $\hat{q}_{pred}$, usually [4, 30, 36] a large parametric convolutional downstream segmentation network follows before the final segmentation mask is output. In our approach, we do not attempt to learn any such. Instead, we directly fuse the coarse query predictions to obtain a single prediction mask:

$$\hat{q}_{fused} = \frac{1}{L}\sum_{l=1}^{L} upsample(\hat{q}_{pred_l}), \tag{7}$$

where $upsample$ is bilinear interpolation to match the size of the query image.

Because of the softmax from Eq. 6, resulting maps are in a subrange of $[0, 1]$. The distribution is sample-dependent, however, and cannot be interpreted as probabilities. Therefore the threshold is chosen such that intra-class variance is minimized, or equivalently, inter-class variance is maximized. This can be estimated by k-means on the histogram [32] of $\hat{q}_{fused}$, such that the binary prediction mask for the query could be obtained as

$$\hat{M}^q = \hat{q}_{fused} > thresh(\hat{q}_{fused}). \tag{8}$$

Specifically, $thresh$ calculates [32], and if it cannot find a reasonable solution above $mean(\hat{q}_{fused})$, $mean(\hat{q}_{fused})$ is selected as threshold. See our supplementary for the rationale.

Because of the heavy upsampling ($\times 32$ from the highest-level ResNet50 layer), such a prediction mask is only coarse. A common solution is to skip-connect [4, 36] low-level features and then convolve the concatenated features in a decoder to produce a more fine-grained mask. We observe that using low-level clues in the form of image-appearance and smoothness is sufficient and hence apply [17] as a non-learnable post-processing to obtain the final prediction mask.

4. Experiments

4.1. Experimental Setup

Metrics. Unlike previous CD-FSS, results are reported measuring mIoU *and* FB-IoU. We argue it is crucial for judging the performance and should always be included. Tab. 1 shows how previous SOTA could have been outperformed by a naive predictor when only considering mIoU. The reason is one can get a mIoU boost simply by increasing the predicted foreground ratio. For definitions, formal derivation and more intuition, please see our Supplementary Material.

Datasets. The primary evaluation datasets are given by the CD-FSS benchmark[20]. We align with it and evaluate on Deepglobe[6], ISIC[5], Chest X-ray (Lung)[3], FSS-1000[47] in the same way. Moreover, we compare the results on the underwater dataset SUIM[15], following [4, 44]. Unlike these works, there is no source dataset in this work. As a consequence, PASCAL[8] and COCO[24] have no usage.

Implementation Details. We adopt ResNet50[13] with ImageNet[34] pretrained weights as the backbone and follow [30] to extract features at the end of each bottleneck before ReLU, resulting in 13 layers. Two augmentations per image are generated, using random shearing of maximum absolute 20 degrees. Similar to [2], our layers are trained with SGD for 25 epochs with learning rate 0.01. Each task adaptor is equally defined as $g_l = Conv_{1\times1}(ReLU(BN(Conv_{1\times1}(F)))$ where the 1×1 convolutions have 64 output channels. For postprocessing, we set standard deviations to 1 for spatial Gaussian, 35 for spatial bilateral, 13 for color as well as compatibilities of 2 and 1 for Gaussian and bilateral, respectively and apply it only if it can increase the intersection over union of a pseudoepisode, where the support image functions as pseudoquery and its augments as the pseudosupport.

Table 1. FB-IoU is important to report besides mIoU: One could *naive*ly outperform previous SOTA on mIoU by simply assigning foreground to all query pixels (100%). 1-shot Deepglobe results, true foreground ratio is 43.5%. †: obtained with models trained by ourselves.

Method	mIoU	FB-IoU	% FG
Naive	**43.0**	21.5	100.0
PATNet†[20]	39.4	47.3	41.5
Ours	42.6	**47.7**	48.6

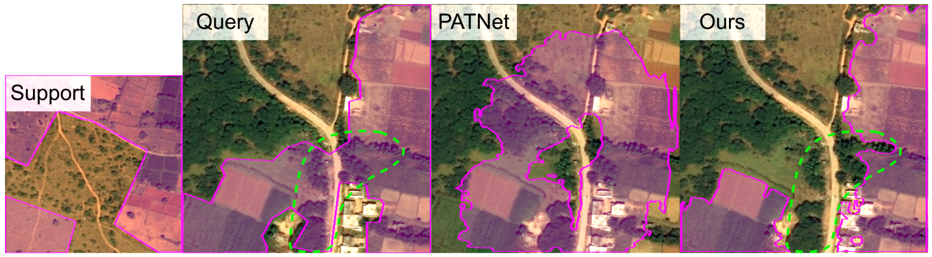

Figure 3. Issue of Deepglobe ground truth annotation. Image row showing an episode featuring the pink overlaid *Agricultural Land* class. Green encircled area contains inaccurate inclusion of *Forest* areas in the ground truth (Query) annotation. Notably, our model appears to segment agricultural land more precise than the ground truth.

Table 2. Results and comparison on the CDFSS benchmark[20] on mIoU. PMNet[4] has not reported class-wise ISIC results.

	1-shot					5-shot				
Method	Deepglobe	ISIC	X-ray	FSS-1000	**Avg.**	Deepglobe	ISIC	X-ray	FSS-1000	**Avg.**
$Linear_{ResNet}$	34.1	20.8	59.1	41.0	38.8	46.5	34.8	64.6	58.7	51.1
$Linear_{Deeplab}$ [20]	33.0	19.4	43.5	40.5	34.1	39.7	30.0	60.3	58.4	47.1
PANet[43](ECCV20)	36.6	25.3	57.8	69.2	47.2	45.4	34.0	69.3	71.7	55.1
RePRI[2](CVPR21)	25.0	23.3	65.1	71.0	46.1	27.4	26.2	65.5	74.2	48.3
HSNet[30] (ICCV21)	29.7	31.2	51.9	77.5	47.6	35.1	35.1	54.4	81.0	51.4
PATNet[20](ECCV22)	37.9	41.2	66.6	78.6	56.1	43.0	**53.6**	70.2	81.2	62.0
HDMNet[33](CVPR23)	25.4	33.0	30.6	75.1	41.0	39.1	35.0	31.3	78.6	46.0
RestNet[14](BMVC23)	22.7	42.3	70.4	81.5	54.2	29.9	51.1	73.7	84.9	59.9
PMNet[4](WACV24)	37.1	-	70.4	**84.6**	-	41.6	-	74.0	**86.3**	-
ABCDFSS (Ours)	**42.6**	**45.7**	**79.8**	74.6	**60.7**	**49.0**	53.3	**81.4**	76.2	**65.0**

Table 3. Results and comparison on SUIM, 1-shot, mIoU.

HS[30]	SCL[51]	RtD[44]	Rest[14]	PM[4]	Ours
28.8	31.8	34.7	25.2	34.8	**35.1**

4.2. Comparison with State-of-the-art

All previous work in our comparisons is trained on PASCAL VOC 2012[8] or, for [33], on the even richer COCO[24]. Our method has seen no dataset.

Table 2 compares our work on the CD-FSS benchmark [20]. In both 1-shot and 5-shot, we surpass PMNet on mIoU by significant margins of 5.5, 7.4 on Deepglobe and by 8.6,7.4 on Chest X-ray, while underperforming on FSS also by a large 10.0 and 10.1. FSS underperformance is due to the character of our approach which does not attempt to learn a segmentation network. As a consequence we can find local semantic similarity well, but a) not learn global semantic clues as a segmentation network's encoder would, and b) not learn spatial accuracy as a segmentation network's decoder would. This impacts our performance on FSS, where finding the object is generally easy, and performance is gained by spatial accuracy. For ISIC, PMNet[4] treats all images as if they belonged to the same semantic class. Hence, they report mIoU by only "averaging" the IoU of one class, which forbids comparison with the CD-FSS benchmark. Nevertheless, we also compare with their ISIC setting, where our work performs better with 51.3(+0.2) on 1-shot and 59.2(+4.7) on 5-shot. Importantly, the CD-FSS benchmark average from previous SOTA[20] is surpassed by our method on average by 4.6 on 1-shot and 3.0 for 5-shot. Tab. 3 shows our method can also surpass PMNet[4] on SUIM by 0.3 against a 0.1 of PM over second-best RtD[44].

In accordance with our findings that FB-IoU needs to be considered as well, Tab. 4 reports our full results compared with PATNet[20] trained by ourselves and the recent FSS-method HDMNet[33]. Instable validation curves during training [20] are observed, causing variations from their reported results, but the overall trend remains stable. HDMNet results are obtained from the meta-branch mask as it was more accurate than the fused mask. Table 4 further documents that even without refinement our results can surpass previous work.

4.3. Computational Efficiency

For a comparison with previous work under equal conditions, the method has been evaluated on episodes consisting of only one query image. In reality, multiple queries might want to be processed subsequently. Since we do test-time adaption utilizing the query, rerunning the task-adaption for every query can be too slow for some application cases. We

Table 4. Our complete results. *m*IoU and *FB*-IoU. *No-pp* reports the performance of the unrefined prediction $\hat{M}^q$ from Eq. 8. †: Results obtained with models trained by ourselves.

Method	Deepglobe				ISIC				X-ray				FSS-1000				**CD-FSS Avg.**				SUIM			
	1-shot		5-shot		1-shot		5-shot		1-shot		5-shot		1-shot		5-shot		1-shot		5-shot		1-shot		5-shot	
	m	FB	m	FB	m	FB	m	FB	m	FB	m	FB	m	FB	m	FB	m	FB	m	FB	m	FB	m	FB
PATNet†[20]	35.4	45.7	41.6	50.2	43.4	62.1	51.8	68.8	63.0	72.6	63.9	73.3	77.7	85.5	79.8	87.2	54.9	66.5	59.3	69.9	32.1	**54.2**	40.2	57.8
HDMNet[33]	25.4	38.4	39.1	46.4	33.0	49.7	35.0	50.4	30.6	25.8	31.3	28.8	75.1	84.1	78.7	86.5	41.0	49.5	46.0	53.0	23.4	49.5	30.9	51.5
Ours (no-pp)	42.3	47.1	48.2	53.4	41.8	57.2	50.8	63.9	80.0	86.2	81.6	87.4	69.3	79.3	73.1	82.3	<u>58.3</u>	<u>67.5</u>	<u>63.4</u>	<u>71.8</u>	<u>35.0</u>	**54.2**	<u>41.1</u>	**58.3**
Ours	42.6	47.7	49.0	54.6	45.7	60.3	53.3	66.1	79.8	86.1	81.4	87.3	74.6	82.7	76.2	84.2	**60.7**	**69.2**	**65.0**	**73.1**	**35.1**	53.5	**41.3**	<u>58.2</u>

Table 5. Task-adapting to a single 1-shot episode and subsequently forward passing other queries. Performance gap to fitting for every episode is reported.

		Deepgl.	ISIC	X-ray	FSS	SUIM	Avg.
quick-infer	mIoU	-0.01	0.01	-0.03	-0.71	-0.01	-0.15
	FB-IoU	-1.08	-0.79	-0.29	0.01	-2.87	-1.00

Table 6. 1-shot performance drop when leaving one loss term out or when using ResNet layers directly as an input for the dense comparison without task-adaption (TA).

		Deepgl.	ISIC	X-ray	FSS	SUIM	Avg.
w/o L^q, L^s	mIoU	-0.06	-8.91	-1.86	-0.18	0.69	-2.06
	FB-IoU	0.31	-6.75	-1.02	0.13	1.27	-1.21
w/o L_p	mIoU	-1.54	-6.52	-4.55	-1.44	-0.95	-3.00
	FB-IoU	-1.52	-6.81	-3.45	-0.66	-1.48	-2.78
w/o TA	mIoU	-3.33	-7.12	-0.03	-15.72	-3.09	-5.86
	FB-IoU	-0.87	-5.68	-0.06	-10.31	0.92	-3.19

therefore evaluate the scenario where task-adaption is only done once, and the thus learned parameters are reutilized for every subsequent query. Under this setting the performance of our method remains stable, with a maximum mIoU drop of 0.7 on FSS-1000 and maximum FB-IoU drop of 2.9 on SUIM, while cutting computational cost to $\sim 1/50$. This highlights that the task-adaption on one query-support pair is able to generalize to other queries. For further speedup, unrefined results are compared. Table 5 reports the results averaged over 200 runs, where a run samples first an episode for training and then infers on further 200 queries.

4.4. Architectural Validation and Ablation

Loss Terms. As shown in Tab. 6, both the unsupervised L^q, L^s and support-supervised L_p contribute to performance enhancement. Interestingly, for FSS-1000, either of them would be sufficient, while for others such as ISIC, only using one term would be harmful. While task adaption is beneficial in all scenarios (compare Fig. 5 c) and Tab. 6), the gap for X-ray is surprisingly small, given that previous work [4, 20, 22] all considered it to have a large domain shift. It implies that intermediate features from the backbone would here be already sufficient. In contrast to the previous segmentation networks which are more harmful (at least -7.4 mIoU) than useful, our multi-layer similarity score aggregation from Eq. 7 proves here to preserve discriminability: Maps with higher confidences receive implicitly higher scores and thus higher weights in the subsequent summation.

Table 7. Intra- and inter-class similarities in the embedding space of (L)ow, (M)iddle and (H)igh-level feature maps before and after Task Adaption. Measure represents averaged cosine similarities of pixel pairs from same and opposite classes, respectively. *Across* SUIM dataset images. A higher delta represents higher discriminability. Full table on all datasets with more extensive measures can be found in supplementary.

Similarities across image pairs ·100	w/o TA L	w/o TA M	w/o TA H	with TA L	with TA M	with TA H
`FG↔FG (INTRA)`	51	42	58	14	20	16
`FG↔BG (INTER)`	50	41	57	-4	-4	-3
delta	**1**	**2**	**1**	**18**	**24**	**18**

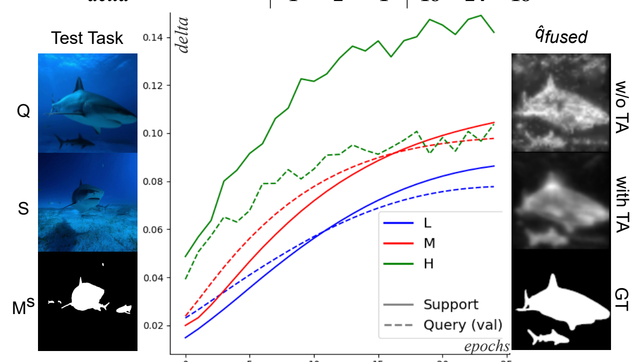

Figure 4. Against common belief, fine-tuning does not lead to overfitting to the support set with our approach. Through learning of consistent embedding spaces, we enhance class discriminability not only for the support (solid lines), but also for the test query (dashed). As a result, irrelevant regions are no longer activated in the coarse query prediction *with TA*.

Discriminabilty in Embedding Space. Lei *et al.* [20] also identify the class distinction as a primary issue for CD-FSS, but measure it using final features of Inception [40] network. However, it is multi-layer ResNet features that are relevant for both their and our network. To obtain a more relevant metric, we reconstruct the dense affinity[4] matrix which is the dot product part of Eq. 6. We measure intra-support through constructing SS^T and across-image through QS^T, where Q and S represent features from a sampled query and support image. Intra-support metrics show how well the model is fit to the relevant class given a single image. Across-image metrics are crucial for generalization

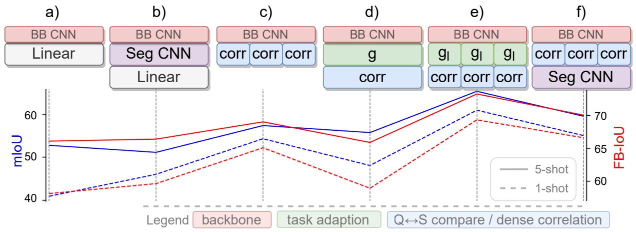

Figure 5. Adapt before comparison e) is the superior approach for CD-FSS. Architectural schemes and their CD-FSS Avg. performance: a) $Linear_{ResNet}$, b) transductive FT [2], c) w/o TA, d) hypercolumn TA, e) proposed f) prev. SOTA [20], also [4, 30, 33, 36].

from support to query. Fig. 4 and Tab. 7 demonstrate the underlying issue of near-zero across-image discriminablilty found initially and how it improves substantially during the learning of our attached layers. This way, the core precondition for the subsequent comparison is restored.

Against other fine-tuning or transfer learning. We validate our proposed architecture by comparing against alternative approaches as shown in Fig. 5. In [35], few-shot segmentation is addressed by appending a small projector network to hypercolumns from a contrastively pretrained encoder. A hypercolumn is the concatenation of features from L layers to shape $H \times W \times (C_0 + C_1 + ... + C_{L-1})$, with upsampling of higher-level features before concatenation, if necessary. This is a viable alternative idea for our level-wise approach. In the *first setting* $Linear_{ResNet}$, Fig. 5 a), we compare with training a linear classifier on the support set, thus mapping backbone hypercolumn directly to foreground probability. Average benchmark performance decreases by mIoU|FBIoU $(-19.6| - 11.1)$ for 1-shot and $(-12.3| - 7.1)$ for 5-shot, proving that simple fine-tuning or transfer learning cannot compete with our method. Compare also $Linear_{Deeplab}$ in Tab. 2 as well as transductive fine-tuning Fig. 5 b). In the *second setting*, Fig. 5 d), we replace our L task adapters with a *single task-adapter network* which takes the ResNet-extracted hypercolumn as input and produces a single feature representation per image. Given query and support representations, the dense comparison module from Sec. 3.3 generates the similarity prediction. No subsequent fusion is required since there is only one prediction map. Average benchmark performance loss is m|FB $(-12.6|-10.3)$, $(-9.4|-7.3)$ for 1- and 5-shot respectively, which suggests that mixing the information from different levels is not as generalizable as comparing them individually. Instead, the simple averaging for fusion proves to be effective for self-regularization and noise suppression.

5. Discussion

Benchmark and Datasets In addition to the demonstrated need to complement mIoU with FB-IoU, there are a few more points to consider. First, we agree with the suggestion [4] to differentiate between *cross-dataset* and *cross-domain* few-shot segmentation. FSS-1000[47] from the CD-FSS benchmark is classified as *cross-dataset* and is hence useful to understand performance in domains where conventional FSS methods also perform well, but the underwater-dataset SUIM[15] used in [44] is more appropriate to consider for pure CD-FSS. Second, the benchmark includes Deepglobe, specifically [20] the Land Cover Classification Dataset. Fig. 3 illustrates that this dataset is improperly annotated, limiting the expressiveness of performance measure. Even though we outperform previous work on Deepglobe, we suggest to find a properly annotated sattelite image segmentation alternative for future work.

Source Domain Our method does not use any dataset for training, yet we refrain from calling it zero-source or source-free to avoid confusion. ImageNet is the source domain of the pretrained backbone. While previous CD-FSS research also uses ImageNet weights, they declare only *their* training domain, i.e. PASCAL or COCO, as the source domain. It might be technically valid since the segmentation network does not see images from ImageNet. However, they suggest misleadingly that the primary challenge lies in bridging the domain gap between e.g. PASCAL and SUIM, neglecting the pivotal role of ImageNet-based features in semantic information transfer. Our results, powered by the backbone features alone, underscore the necessity to acknowledge this across related works.

Limitations and Future Work By removing the learning of a segmentation network, we intend to expose the fundamental problem in cross-domain few-shot segmentation and show that it should be addressed by test-time task adaption. However, we do not see our method as a final solution for CD-FSS. As it adapts features vector-wise, it does not learn the scene-level semantic context which is commonly seen as a key for semantic segmentation. While it is out of scope of this work, we believe that if our findings are addressed appropriately, replacing the heuristic fusion and refinement through reintroducing training on source segmentation tasks could improve performance in future work.

6. Conclusion

Previous sophisticated similarity fusion models are not yet effective for CD-FSS. We presented a cross-domain few-shot segmentation method that outperforms previous approaches with no segmentation network. Averaging similarities across a feature pyramid is simpler and more effective, provided that the features are task-adapted before calculating their similarities. Enforcing contrastive consistency proved to be a strategy that could avoid overfitting to the support set while fine-tuning. Results and experiments suggest task-adaption before comparison is the superior approach for CD-FSS.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 2

[2] Malik Boudiaf, Hoel Kervadec, Imtiaz Masud Ziko, Pablo Piantanida, Ismail Ben Ayed, and José Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13974–13983, 2021. 1, 2, 3, 5, 6, 8

[3] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul Kumar Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Kiran Antani, George R. Thoma, and Clement J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33:577–590, 2014. 5

[4] Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, and Jungong Han. Pixel matching network for cross-domain few-shot segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 978–987, 2024. 1, 2, 3, 5, 6, 7, 8

[5] Noel C. F. Codella, Veronica M Rotemberg, Philipp Tschandl, M. E. Celebi, Stephen W. Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Armando Marchetti, Harald Kittler, and Allan C. Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *ArXiv*, abs/1902.03368, 2019. 5

[6] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–17209, 2018. 5

[7] Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *British Machine Vision Conference*, 2018. 2, 4

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010. 1, 2, 5, 6

[9] Yu Fu, Yu Xie, Yanwei Fu, and Yugang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24575–24584, 2023. 2

[10] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24575–24584, 2023. 2

[11] Yipeng Gao, Lingxiao Yang, Yunmu Huang, Song Xie, Shiyong Li, and Weihao Zheng. Acrofod: An adaptive method for cross-domain few-shot object detection. In *European Conference on Computer Vision*, 2022. 2

[12] Yunhui Guo, Noel C. F. Codella, Leonid Karlinsky, James Codella, John R. Smith, Kate Saenko, Tajana Rosing, and Rogério S. Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, 2019. 2

[13] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2015. 5

[14] Xinyang Huang, Chuanglu Zhu, and Wenkai Chen. Restnet: Boosting cross-domain few-shot segmentation with residual transformation network. In *British Machine Vision Conference*, 2023. 2, 6

[15] Md. Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1769–1776, 2020. 5, 8

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *ArXiv*, abs/2304.02643, 2023. 1

[17] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 5

[18] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8047–8057, 2022. 2

[19] Chunbo Lang, Gong Cheng, Binfei Tu, Chao Li, and Junwei Han. Base and meta: A new perspective on few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10669–10686, 2023. 2

[20] Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *European Conference on Computer Vision*, 2022. 1, 2, 3, 5, 6, 7, 8

[21] Gen Li, V. Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8330–8339, 2021. 2

[22] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7151–7160, 2021. 2, 7

[23] Hanwen Liang, Qiong Zhang, Peng Dai, and Juwei Lu. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. *IEEE/CVF International Conference on Computer Vision*, pages 9404–9414, 2021. 2

[24] Tsung-Yi Lin, Michael Maire, Serge J. 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*, 2014. 2, 5, 6

[25] 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, 2020. 2

[26] Yuhang Lu, Xinyi Wu, Zhenyao Wu, and Song Wang. Cross-domain few-shot segmentation with transductive fine-tuning. *ArXiv*, abs/2211.14745, 2022. 2, 3

[27] 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. *IEEE/CVF International Conference on Computer Vision*, pages 8721–8730, 2021. 2

[28] Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. In *International Conference on Machine Learning*, 2023. 2

[29] Tianyi Ma, Yifan Sun, Zongxin Yang, and Yi Yang. Prod: Prompting-to-disentangle domain knowledge for cross-domain few-shot image classification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19754–19763, 2023. 2

[30] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmenation. *IEEE/CVF International Conference on Computer Vision*, pages 6921–6932, 2021. 1, 2, 3, 5, 6, 8

[31] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020. 2, 3

[32] Nobuyuki Otsu. A threshold selection method from gray level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62–66, 1979. 5

[33] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chenyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 2, 6, 7, 8

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014. 5

[35] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. Ganorcon: Are generative models useful for few-shot segmentation? *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9981–9990, 2021. 2, 8

[36] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision*, pages 151–168, 2022. 2, 3, 5, 8

[37] Chen Shuai, Meng Fanman, Zhang Runtong, Qiu Heqian, Li Hongliang, Wu Qingbo, and Xu Linfeng. Visual and textual prior guided mask assemble for few-shot segmentation and beyond. *arXiv preprint arXiv:2308.07539*, 2023. 2

[38] Manogna Sreenivas and Soma Biswas. Similar class style augmentation for efficient cross-domain few-shot learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4590–4598, 2023. 2

[39] Yanpeng Sun, Qiang Chen, Xiangyu He, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Jian Cheng, Zechao Li, and Jingdong Wang. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. *Advances in Neural Information Processing Systems*, 35:37484–37496, 2022. 2

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2018. 7

[41] 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*, 44:1050–1065, 2020. 2, 3

[42] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. *ArXiv*, abs/2001.08735, 2020. 2

[43] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. *IEEE/CVF International Conference on Computer Vision*, pages 9196–9205, 2020. 2, 6

[44] Wenjian Wang, Lijuan Duan, Yuxi Wang, Qing En, Junsong Fan, and Zhaoxiang Zhang. Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7055–7064, 2022. 1, 2, 3, 4, 5, 6, 8

[45] Wenjian Wang, Lijuan Duan, Yuxi Wang, Junsong Fan, Zhi Gong, and Zhaoxiang Zhang. A survey of deep visual cross-domain few-shot learning. *ArXiv*, abs/2303.09253, 2023. 2

[46] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3023–3032, 2021. 2, 3

[47] Tianhan Wei, Xiang Li, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2866–2875, 2019. 5, 8

[48] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3, 4

[49] Huali Xu, Shuaifeng Zhi, Shuzhou Sun, Vishal M Patel, and Li Liu. Deep learning for cross-domain few-shot visual recognition: A survey. *arXiv preprint arXiv:2303.08557*, 2023. 2

[50] 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 151–168, 2020. 2

[51] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8308–8317, 2021. 6

[52] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5212–5221, 2019. 2

[53] Gengwei Zhang, Guoliang Kang, Yunchao Wei, and Yi Yang. Few-shot segmentation via cycle-consistent transformer. In *Neural Information Processing Systems*, 2021. 2

[54] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50:3855–3865, 2018. 2