Rethinking the Correlation in Few-Shot Segmentation: A Buoys View

Yuan Wang* Rui Sun* Tianzhu Zhang†
University of Science and Technology of China
{wy2016, issunrui}@mail.ustc.edu.cn, {tzzhang}@ustc.edu.cn

Abstract

Few-shot segmentation (FSS) aims to segment novel objects in a given query image with only a few annotated support images. However, most previous best-performing methods, whether prototypical learning methods or affinity learning methods, neglect to alleviate false matches caused by their own pixel-level correlation. In this work, we rethink how to mitigate the false matches from the perspective of representative reference features (referred to as buoys), and propose a novel adaptive buoys correlation (ABC) network to rectify direct pairwise pixel-level correlation, including a buoys mining module and an adaptive correlation module. The proposed ABC enjoys several merits. First, to learn the buoys well without any correspondence supervision, we customize the buoys mining module according to the three characteristics of representativeness, task awareness and resilience. Second, the proposed adaptive correlation module is responsible for further endowing buoy-correlation-based pixel matching with an adaptive ability. Extensive experimental results with two different backbones on two challenging benchmarks demonstrate that our ABC, as a general plugin, achieves consistent improvements over several leading methods on both 1-shot and 5-shot settings.

1. Introduction

Semantic segmentation has achieved conspicuous achievements attributed to the recent advances in deep neural network [20]. However, its data-driven nature makes it heavily dependent on massive pixel-level training data, which is labor-intensive and time-consuming to collect. To imitate the human learning habits which can recognize new classes with only a glance, few-shot segmentation [25] (FSS) has attracted increasing interest in recent years, which aims at segmenting novel objects in the given query image with a few annotated support images.

In previous literature, superior prototypical learning methods [15, 28, 37] and affinity learning methods [11, 26, 30, 40] are almost all equipped with pixel-level correlation. In specific, for prototypical learning methods, pixel-level correlation is implicitly endowed with the expectation to generate the foreground prior mask [28] for guiding the query pixel classification. For affinity learning methods, pixel-level correlation directly serves to aggregate support information and convey it to the query image [40].

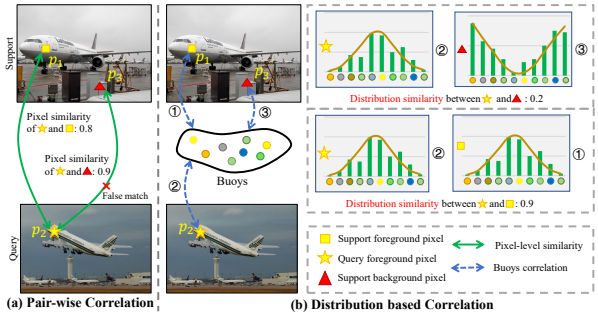

Figure 1. The motivation of our proposed method. False matches tend to occur in the pixel-level correlation due to large intra-class variations. We introduce a series of representative features (buoys) as references and calculate the buoys-level correlation to suppress false matches.

Despite their promising results, these methods neglect the fact that there may exist cluttered background and inherent large intra-class variations between support and query images. In this case, directly employing pairwise pixel correlation may lead to considerable false matches. To make matters worse, the negative impact is inevitably amplified by inbuilt low-data regimes of FSS, leading to sub-optimal results. As shown in Fig. 1 (a), due to the significant pose difference of the object *plane* in the support-query image pair, p_2 in the query image located in the *plane hatch* is erroneously closer to p_3 situated on the *ground* than counterpart p_1 in the support image. Therefore, it is highly desirable to rectify these false matches caused by the direct pairwise pixel-level correlation.

In this paper, we aim to mitigate the false matches in previous FSS methods from the perspective of representative reference features (referred to as *buoys*). Specifically, we design a novel **A**daptive **B**uoys **C**orrelation (**ABC**) network that can be applied as a generic plugin, including a buoy mining module and an adaptive correlation module to rec-

*Equal contribution
†Corresponding author

tify direct pairwise pixel-level correlation for robust FSS. The main idea is, for each pixel from the support or query image, we can obtain the buoy-level correlation (*i.e.*, a likelihood vector) by comparing this pixel with a set of buoys. In essence, the buoy-level correlation reflects the consensus among representative buoys with a broader receptive field, thus it encodes the relative semantic comparability of the buoys that can be relied upon. Intuitively, each pair of true pixel correlation (*e.g.*, the p_1-p_2 pair in Fig. 1 (b)) derived from the query and support images should be not only visually similar to each other (*i.e.*, high pairwise pixel-level correlation), but also similar to any other buoys (*i.e.*, similar buoy-level correlation pair). Based on this *correlation consistency* in ABC, false matches caused by similar vision but dissimilar buoy correlations will be suppressed (*e.g.*, the point p_3-p_2 pair in Fig. 1 (b)), ensuring that true pixel correlations enjoy higher weights to safely extract support information.

However, it is non-trivial to learn the buoys well without any correspondence supervision for training. **In the buoys mining module** (BMM), we carefully design this module customized for the following three characteristics. (1) Representativeness. Intuitively, the buoys should have the ability to represent the diverse semantic clues from both support and query pixels with a broader semantic contrast descriptive. In other words, the matching between support-query pixels based on buoy-level correlation should preserve as much critical information as possible in the correspondence based on pixel-level correlation. In specific, we take advantage of Singular Value Decomposition (SVD) in pursuit of controllable information decay. Besides, a representation decay loss is devised to prevent the degradation of buoys. (2) Task awareness. Since tasks are randomly sampled during FSS training, each individual task consists of unique categories with large distribution differences. Therefore, to enable the buoys to perceive the current task and generalize to novel classes well, we employ the cross-aggregation mechanism to flexibly adjust buoys to meet the expectations of any tasks, even the tasks with unseen classes. (3) Resilience. Considering the large gap between support and query images caused by large intra-class variations and cluttered background, it is necessary for buoys to bridge this gap and become more referable. In specific, we prepend the self-aggregation mechanism to amend buoys by reconciling the intrinsic resilience between support and query images.

Moreover, we observe that not all buoys are profitable when calculating pixel pair matching based on buoy-level correlation, and comprehensive consideration of the relationships between helpful buoys with potential intersections can assist in the final matching score. **In the adaptive correlation module** (ACM), we endow buoy-correlation-based pixel matching with an *adaptive* ability, which can flexibly assign less weight to irrelevant buoys conditioned to different pixel pairs and focus on the structural similarity of related buoys. In specific, given the corresponding buoy pair for each pixel pair as the initial marginal distribution of the optimal transport(OT) algorithm, we can attain the optimal transport plan which can be regarded as the structural buoy contribution adaptive to the current pixel pair, and the corresponding OT distance is adopted for scoring matches.

In this work, our contributions can be concluded as follows: (1) We propose an Adaptive Buoys Correlation (ABC) network to rectify the widely used pairwise pixel correlation in FSS. To the best of our knowledge, this is the first work to mitigate the false matches in FSS methods, from the perspective of representative reference features (buoys). (2) We introduce two novel modules, namely Buoys Mining Module (BMM) and Adaptive Correlation Module (ACM), for representative buoys construction and adaptive matching respectively. They can cooperate well to achieve effective false match suppression. (3) Extensive experimental results with two different backbones on two challenging benchmarks demonstrate that our ABC, as a general plugin module, achieves consistent improvements over several leading methods on both 1-shot and 5-shot settings.

2. Related Work

In this section, we briefly overview several lines of research in semantic segmentation, few-shot segmentation.

Semantic Segmentation. Semantic Segmentation is a fundamental computer vision task that aims to achieve pixel-level classification of the given images on predefined categories. Benefitting from the advantages of the DNN [21], remarkable progress has been achieved in the past decade. Based on the Fully Convolutional Network(FCN) [20], many remarkable modules designs have been proposed, such as feature pyramid modules [13, 16, 44], context absorbing modules [8, 12, 23] and dilated convolution modules [2, 3]. In addition to CNN-based models, many researchers have turned their attention to transformer-based semantic segmentation models [4, 27, 34, 43]. Though achieving promising results, these methods cannot generalize to novel classes in the low-data regime. This paper tackles the semantic segmentation problem in a few-shot setting.

Few-Shot Segmentation. Few-shot segmentation [25] aims to perform pixel-wise classification on images of previously unseen categories with only a handful of labeled images available. Owing to the scarcity of annotated samples, fully mining category information from support images is crucial for FSS tasks. Mainstream FSS methods can be roughly divided into two categories according to the paradigm to excavate support information: prototypical feature learning methods and affinity learning methods. For the former, most methods [15, 19, 32, 36, 37, 39, 42] condense the masked support features into a single or multiple prototypes for feature comparison or aggregation. For example, SG-

One [42] exploits cosine similarity to compare the prototype obtained by mask average pooling with the query features for segmentation results. However, relying only on heavily compressed prototypes will inevitably lead to information loss, resulting in degraded segmentation performance. While for the affinity learning methods [11, 30, 35, 38, 40], fine-grained pairwise relationships between support and query features are further considered to retain the details. For instance, CyCTR [40] introduces the cycle-consistent attention mechanism to selectively aggregate support features. Shi *et al.* [26] exploit the attention weight between support and query features to conduct additive aggregation of support masks. Though achieving promising results, pixel-level correlation tends to suffer from false matches caused by intro-class variations and cluttered backgrounds.

Pixel-level Correlation in FSS. Pixel-level correlation is a powerful manner of extracting fine-grained support information and is thus widely applied in FSS methods. According to recent top-performing researches, the following two paradigms are the most frequently employed. (1) *Prior Mask Guidance* (PMG) is first proposed by [28] and widely adopted in numerous subsequent methods [14, 15, 32, 33, 37, 40] due to its simplicity and effectiveness. To be concrete, PMG calculates the maximum one-to-one correlation responses between the high-level support foreground features and query features to tell the probability of query pixels belonging to the foreground. (2)*Multi-head Attention* (MHA) is receiving increasing interest in visual tasks since the advent of ViT [6] and DETR [1]. In terms of FSS, MHA is also exploited to transport the support information to facilitate the query segmentation [18, 26, 40, 41]. As described before, Non-ignorable false matches usually occur in pixel-level correlation. In this paper, we propose to mitigate false matches by introducing a series of reliable buoys. We deploy our approach on three models with PMG [28] or MHA [26, 40] and obtain significant performance gains with a slight increase in the number of parameters.

3. Method

3.1. Problem Formulation

The aim of FSS is to perform segmentation on novel classes with only a handful of densely-labeled samples. Following previous works, we adopt the widely used episodic meta-training paradigm. Specifically, given the training set $\mathcal{D}_{train}$ and the testing set $\mathcal{D}_{test}$ that are disjoint in the target categories($\mathbb{C}_{train} \cap \mathbb{C}_{test} = \emptyset$), a set of subtasks are sampled from the $\mathcal{D}_{train}$ to train the model. Each subtask contains a support set $\mathcal{S} = \{(I_S^k, M_S^k)\}_{k=1}^K$ and a query set $\mathcal{Q} = \{(I_Q, M_Q)\}$, where the I and M denote the RGB image and the corresponding ground-truth. During training, the model learns to predict on the I_Q conditioned on the $\mathcal{S}$ under the supervision of M_Q. For testing, the trained model is evaluated with the tasks sampled from the $\mathcal{D}_{test}$.

3.2. Overview

As illustrated in Fig. 2, the proposed ABCNet mainly includes two components, i.e., the buoys mining module (BMM) and the adaptive correlation module (ACM). BMM is responsible for constructing buoys with representativeness, task awareness and resilience through SVD-based initialization, cross-aggregation and self-aggregation, respectively. The ACM module further enables the well-constructed buoys to adaptively score the support-query matches via the specially designed OT distance. The ultimate goal of our method is to suppress the false matches in the original correlation matrix $\mathbf{W}$:

$$\mathbf{W} = \boldsymbol{Sim}(\mathbf{F}_Q, \mathbf{F}_S), \tag{1}$$

where the $\mathbf{F}_Q \in \mathbb{R}^{hw \times c}$ and $\mathbf{F}_S \in \mathbb{R}^{hw \times c}$ are the reshaped query and support features extracted from shared ImageNet [24] pretrained backbones, respectively. Here the h, w and c mean the spatial size and the channel size of the features. Note that $\mathbf{F}_Q$ and $\mathbf{F}_S$ can be high-level or middle-level features depending on baseline methods, and the $\boldsymbol{Sim}$ denotes the pixel-wise similarity measurement such as cosine similarity or attention weights in MHA. The details are as follows.

3.3. Buoys Mining Module

It is non-trivial to construct buoys that are well-tailored for corresponding tasks. The buoys mining process can be divided into three procedures, i.e., SVD-based initialization, cross-aggregation and self-attention which endow the buoys with representativeness, task awareness and resilience, respectively.

SVD-based Initialization. In order to enable buoys to represent the diverse semantic clues from both support and query pixels, we resort to Singular Value Decomposition(SVD) to implement principal component analysis on the basis of original pixel wise matching matrix $\mathbf{W} \in \mathbb{R}^{hw \times hw}$ in both the support and the query dimensions. Specifically, we decompose the original correlation matrix $\mathbf{W}$ via SVD and only keep the largest K singular values:

$$\mathbf{W} \xrightarrow[Top-K]{SVD} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{hw \times K}$, $\mathbf{\Sigma} \in \mathbb{R}^{K \times K}$, $\mathbf{V}^T \in \mathbb{R}^{K \times hw}$. Since the singular value decreases rapidly, it is enough to keep the largest first few singular values for retaining correlation information. For the left singular matrix $\mathbf{U} \in \mathbb{R}^{hw \times K}$ after SVD decomposition, it can be seen as K orthogonal bases in the space of query feature number, and then we explicitly map the query features $\mathbf{F}_Q$ in the form of linear transformation (multiplied with the orthogonal bases) to attain the initial buoys. The same procedure is conducted for buoy

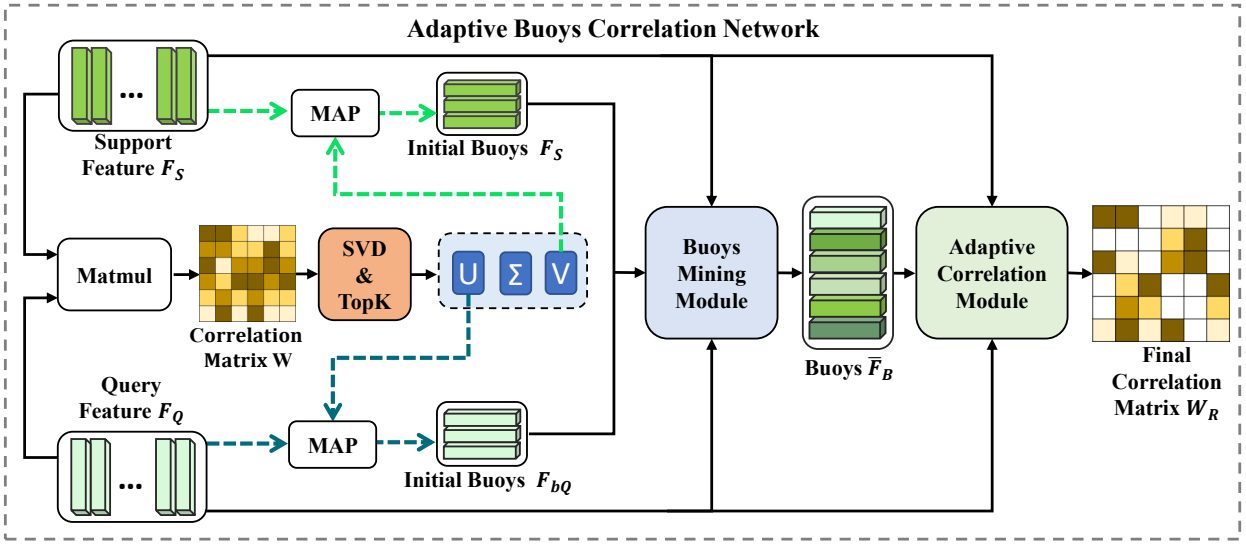

Figure 2. Framework of our proposed Adaptive Buoys Correlation Network (ABCNet). There are two main modules in ABCNet, i.e., the buoys mining module for establishing representative buoys (including the buoys initialization) and the adaptive correlation module for adaptive matching. The ultimate goal of our method is to suppress the false matches that occur in the original pixel-level correlation matrix

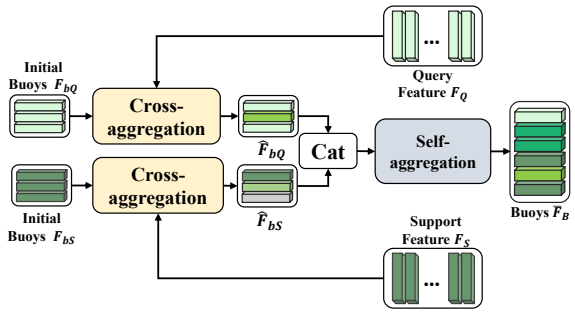

Figure 3. Illustration of the buoys mining module (BMM). There are two submodules in the BMM, i.e., the cross-aggregation module and the cross-aggregation module. The former is responsible for absorbing task-aware context and the latter is used for reducing the gap between the support and the query features.

initialization in support features. Specifically,

$$\mathbf{F}_{bq}^T = \mathbf{F}_Q^T\mathbf{U}, \quad \mathbf{F}_{bs}^T = \mathbf{F}_S^T\mathbf{V}, \tag{3}$$

the resulting features $\mathbf{F}_{bq} \in \mathbb{R}^{K\times C}$ and $\mathbf{F}_{bs} \in \mathbb{R}^{K\times C}$ are the combination of the original features which retain the most of the information included in the $\mathbf{W}$. In this way, the well-initialized buoys not only enable controllable information decay but also speed up the convergence of the training process. In order to further prevent buoys from degradation, we propose a representation decay loss in form of:

$$\boldsymbol{Loss}_{rd} = \|\mathbf{W}_R - \mathbf{W}\|_F, \tag{4}$$

where the $\mathbf{W}_R$ is buoys-correlation based matching matrix in Sec. 3.4. The effect of this loss is analyzed in the Sec. 4.

Cross-aggregation Mechanism. Task-specific contextual information not only enables the buoys to perceive the current task but also endows the buoy with resistance to noisy pixels. To achieve task awareness, we introduce the cross-aggregation mechanism to condense the task-specific information into corresponding buoys. As illustrated in Fig. 3, given the initial buoys features $\mathbf{F}_{b\star} = [\mathbf{b}_{1,\star}, \mathbf{b}_{2,\star}, ..., , \mathbf{b}_{K,\star}]$ ($\mathbf{b}_{i,\star} \in \mathbb{R}^{1\times C}$ indicates the i-th buoy feature), the support and query features $\mathbf{F}_\star = [\mathbf{f}_{1,\star}, \mathbf{f}_{2,\star}, ..., , \mathbf{f}_{hw,\star}]$ ($\mathbf{f}_{j,\star} \in \mathbb{R}^{1\times C}$ indicates the image feature of the j-th spatial position). Note that we denote the corner mark as $\star$ and $\star \in \{S, Q\}$ for brevity. As done in [1], we obtain the *queries* ($\mathbf{Q}_\star$, we denote as *qry* to distinguish it from the query image) arise from buoys features $\mathbf{F}_{b\star}$ and *keys* ($\mathbf{K}_\star$) from support or query features $\mathbf{F}_\star$ for calculating aggregation weights, and *values* ($\mathbf{V}_\star$) from $\mathbf{F}_\star$ for feature aggregation, in concrete:

$$\mathbf{Q}_{i,\star} = \mathbf{b}_{i,\star}\mathbf{W}_\star^{\mathcal{Q}}, \ \mathbf{K}_{j,\star} = \mathbf{f}_{j,\star}\mathbf{W}_\star^{\mathcal{K}}, \ \mathbf{V}_{j,\star} = \mathbf{f}_{j,\star}\mathbf{W}_\star^{\mathcal{V}}, \tag{5}$$

among which, $\mathbf{W}_\star^{\mathcal{Q}} \in \mathbb{R}^{C\times Ck}$, $\mathbf{W}_\star^{\mathcal{K}} \in \mathbb{R}^{C\times Ck}$,$\mathbf{W}_\star^{\mathcal{V}} \in \mathbb{R}^{C\times Cv}$ are linear projections. For the i-th *qry* $\mathbf{Q}_{i\star}$, we calculate the attention weights via dot-product between $\mathbf{Q}_{i\star}$ and all other *keys*:

$$s_{i,j} = \frac{\exp(\beta_{i,j})}{\sum_{j=1}^{hw}\exp(\beta_{i,j})}, \beta_{i,j} = \frac{\mathbf{Q}_{i,\star}\mathbf{K}_{j,\star}^T}{\sqrt{d_k}}, \tag{6}$$

where $\sqrt{d_k}$ is a scaling factor. The context-aware buoys features are further obtained via the weighted sum over all values:

$$\hat{\mathbf{b}}_{i,\star} = \sum_{j=1}^{hw} s_{i,j}\mathbf{V}_{j\star}, \tag{7}$$

The Eq. (6) and Eq. (7) are implemented with the multi-head paradigm following the standard operation [29]. Then

a feed-forward network (FFN) is further applied to obtain the support-related $\hat{\mathbf{F}}_{bS} = [\hat{\mathbf{b}}_{1,s}, \hat{\mathbf{b}}_{2,s}, ..., , \hat{\mathbf{b}}_{K,s}]$ and query-related buoys $\hat{\mathbf{F}}_{bQ} = [\hat{\mathbf{b}}_{1,q}, \hat{\mathbf{b}}_{2,q}, ..., , \hat{\mathbf{b}}_{K,q}]$, which contain abundant corresponding contextual information incorporated.

Self-aggregation Mechanism. The obtained $\hat{\mathbf{F}}_{bS}$ and $\hat{\mathbf{F}}_{bQ}$ in Eq. (7) are respectively originated from the support and query features, which usually exist significant information gap due to large intra-class variations and cluttered background. To bridge this gap, we propose the self-aggregation mechanism to further absorb the full task contextual information, especially to capture the co-occurring target object. Specifically, we first concatenate the $\hat{\mathbf{F}}_{bs}$ and $\hat{\mathbf{F}}_{bq}$ in the dimension of number:

$$\hat{\mathbf{F}}_B = \boldsymbol{Concatenate}(\hat{\mathbf{F}}_{bs}, \hat{\mathbf{F}}_{bq}), \tag{8}$$

then two multi-head self-attention layers are implemented on the concatenated buoys $\hat{\mathbf{F}}_B \in \mathbb{R}^{2K\times c}$, for the sake of brevity, we omit the element ordinal index in the formula, formally:

$$\mathbf{Q}_B = \hat{\mathbf{F}}_B\mathbf{W}_B^{\mathcal{Q}}, \quad \mathbf{K}_B = \hat{\mathbf{F}}_B\mathbf{W}_B^{\mathcal{K}}, \quad \mathbf{V}_B = \hat{\mathbf{F}}_B\mathbf{W}_B^{\mathcal{V}}, \tag{9}$$

where the linear projections $\mathbf{W}_B^{\mathcal{Q}} \in \mathbb{R}^{C\times Ck}$, $\mathbf{W}_B^{\mathcal{K}} \in \mathbb{R}^{C\times Ck}$,$\mathbf{W}_B^{\mathcal{V}} \in \mathbb{R}^{C\times Cv}$. Then the attention weight matrix $\mathbf{S} \in \mathbf{R}^{2K\times 2K}$ is calculated with the scaled dot-product to selectively aggregate the beneficial values:

$$\bar{F}_B = \boldsymbol{S}_{QK}\mathbf{V}_B = \boldsymbol{Softmax}(\frac{\mathbf{Q}_B\mathbf{K}_B^T}{\sqrt{d_{k1}}})\mathbf{V}_B. \tag{10}$$

Like in Eq. (7), a FFN is further applied. In this way, buoys are amended by absorbing complementary information. Through the above three procedures, the buoys that have excellent esthesia of the current task can be well established.

3.4. Adaptive Correlation Module

With the attendance of buoys from BMM as references, it is intuitive to measure the similarity of buoy-level correlation via dot product or L_2 distance. We argue that these solutions are suboptimal as they separately and equally consider the association between different pixel features and different buoys. But not all buoys are profitable when comparing the similarity of buoy-level correlation of a specific pair of pixels. We declare that the association between pixel features and buoys, as well as the association between buoys and buoys, should be taken into account when measuring buoy-based similarity. To achieve that, as shown in Fig. 4, we employ the **optimal transport (OT)** algorithm with a specially designed cost matrix to calculate the matching based on buoys-level correlation more comprehensively. The

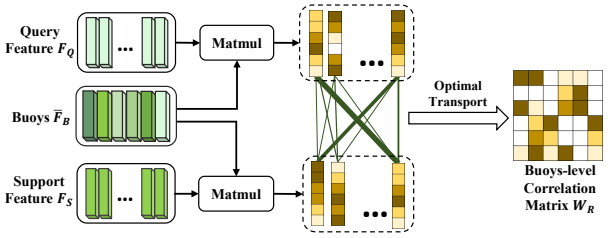

Figure 4. Illustration of the adaptive correlation module(ACM), which endows buoys-correlation-based pixel matching with an adaptive ability via optimal transport (OT) algorithm.

OT algorithm aims to find the minimum transmission cost between probability distributions on the given cost matrix, which can be solved elegantly by the Sinkhorn algorithm with linear programming [5]. In our solution, the cost matrix is defined as $(1 - \mathbf{S}^{BB})$, where the $\mathbf{S}^{BB} \in \mathbb{R}^{K\times K}$ is the similarity matrix between the buoys(for brevity we take the K as the number of all buoys from BMM). The transport plan is denoted as $\mathbf{T} \in \mathbb{R}^{K\times K}$ and the optimization function is as follows:

$$\min_{\mathbf{T}\in\mathcal{T}} \mathrm{Tr}\left(\mathbf{T}^{\mathsf{T}}(1 - \mathbf{S}^{BB})\right) + \epsilon H(\mathbf{T}), \tag{11}$$

where the $H(\mathbf{T}) = -\sum_{ij}\mathbf{T}_{ij}\log\mathbf{T}_{ij}$ measures the entropy regularization and the ϵ is the weight for the entropy term that controls the smoothness of the mapping. The transmission matrix needs to satisfy the following distribution constraints:

$$\mathcal{T} = \left\{\mathbf{T} \in \mathbb{R}_+^{K\times K} \mid \mathbf{T}1 = \boldsymbol{\mu}, \mathbf{T}^{\mathsf{T}}1 = \boldsymbol{\nu}\right\}, \tag{12}$$

where the $\boldsymbol{\mu} \in \mathbb{R}^K$ and $\boldsymbol{\nu} \in \mathbb{R}^K$ represent the similarity distribution of the i-th support feature $\mathbf{f}_{i,S}$ and j-th query feature $\mathbf{f}_{j,Q}$ on the buoys, respectively. The optimal transportation matrix $\mathbf{T}_{i,j}^*$ can be obtained effectively via several sinkhorn iterations, then the final OT-based similarity is calculated as:

$$\mathbf{W}_R(i,j) = \mathbf{T}_{i,j}^* \odot \mathbf{S}^{BB}. \tag{13}$$

The cost matrix subtly injects the structural information inherent in buoys into the calculation of matching weights, and the constraint of pixel-specific marginal distribution adaptively makes irrelevant buoys assigned less weight.

4. Experiments

4.1. Dataset and Evaluation Metric

Dataset. We evaluate the proposed ABCNet on two widely used FSS datasets, namely PASCAL-5^i [25] and COCO-20^i. Among them, PASCAL-5^i is built based on PASCAL VOC 2012 dataset [7] with extra annotations from SBD [9]. The 20 semantic categories are divided into 4 splits for cross-validation as done in OSLSM [25]. COCO-20^i is a larger dataset built from MSCOCO [17]. As done in [22], all 80 categories are separated into 4 splits, 3 of them are

Table 1. Performance on Pascal-5^i [7] in terms of mIoU and FBIoU for 1-shot and 5-shot segmentation. The best mean results are shown in **bold**. The combination of the proposed ABCNet consistently improves the results.

Method	backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FBIoU	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FBIoU
PFENet[TPAMI2020] [28]	ResNet50	61.7	69.5	55.4	56.3	60.7	73.3	63.1	70.7	55.8	57.9	61.9	73.9
PFENet w/ ABCNet		62.5	70.8	57.2	58.1	**62.2**$_{(\uparrow 1.5)}$	**74.1**$_{(\uparrow 0.8)}$	64.7	73.0	57.1	59.5	**63.6**$_{(\uparrow 1.7)}$	**74.2**$_{(\uparrow 0.3)}$
CyCTR[NIPS2021] [40]		67.8	72.8	58.0	58.0	64.2	–	71.1	73.2	60.5	57.5	65.6	–
CyCTR w/ ABCNet		67.8	74.3	59.2	59.4	**65.2**$_{(\uparrow 1.0)}$	**73.8**$_{(\uparrow --)}$	72.6	74.4	61.3	59.0	**66.8**$_{(\uparrow 1.2)}$	**76.2**$_{(\uparrow --)}$
DCAMA[ECCV2022] [26]		67.5	72.3	59.6	59.0	64.6	75.7	70.5	73.9	63.7	65.8	68.5	79.5
DCAMA w/ ABCNet		68.8	73.4	62.3	59.5	**66.0**$_{(\uparrow 1.4)}$	**76.0**$_{(\uparrow 0.3)}$	71.7	74.2	65.4	67	**69.6**$_{(\uparrow 1.1)}$	**80.0**$_{(\uparrow 0.5)}$
PFENet[TPAMI2020] [28]	ResNet101	60.5	69.4	54.4	55.9	60.1	72.9	62.8	70.4	54.9	57.6	61.4	73.5
PFENet w/ ABCNet		62.7	70.0	55.1	57.5	**61.3**$_{(\uparrow 1.2)}$	**73.7**$_{(\uparrow 0.8)}$	63.4	71.8	56.4	57.7	**62.3**$_{(\uparrow 0.9)}$	**74.0**$_{(\uparrow 0.5)}$
CyCTR[NIPS2021] [40]		69.3	72.7	56.5	58.6	64.3	72.9	73.5	74.0	58.6	60.2	66.6	75.0
CyCTR w/ ABCNet		71.2	73.0	57.9	60.2	**65.6**$_{(\uparrow 1.3)}$	**74.6**$_{(\uparrow 1.7)}$	74.2	73.0	60.2	62.1	**67.4**$_{(\uparrow 0.8)}$	**76.6**$_{(\uparrow 1.6)}$
DCAMA[ECCV2022] [26]		65.4	71.4	63.2	58.3	64.6	77.6	70.7	73.7	66.8	61.9	68.3	80.8
DCAMA w/ ABCNet		65.3	72.9	65.0	59.3	**65.6**$_{(\uparrow 1.0)}$	**78.5**$_{(\uparrow 0.9)}$	71.4	75.0	68.2	63.1	**69.4**$_{(\uparrow 1.1)}$	**80.8**$_{(\uparrow 0.0)}$

Table 2. Performance on COCO-20^i [17] in terms of mIoU and FBIoU for 1-shot and 5-shot segmentation. The best mean results are shown in **bold**. The combination of the proposed ABCNet consistently improves the results.

Method	backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FBIoU	Fold-0	Fold-1	Fold-2	Fold-3	mIoU	FBIoU
PFENet[TPAMI2020] [28]	ResNet101	34.3	33.0	32.3	30.1	32.4	58.6	38.5	38.6	38.2	34.3	37.4	61.9
PFENet w/ ABCNet		36.5	35.7	34.7	31.4	**34.6**$_{(\uparrow 2.2)}$	**59.2**$_{(\uparrow 0.6)}$	40.1	40.1	39.0	35.9	**38.8**$_{(\uparrow 1.4)}$	**62.8**$_{(\uparrow 0.9)}$
CyCTR[NIPS2021] [40]	ResNet50	38.9	43.0	39.6	40.3	40.5	–	41.1	48.9	45.2	47.0	45.6	–
CyCTR w/ ABCNet		40.7	45.9	41.6	40.6	**42.2**$_{(\uparrow 1.8)}$	**66.7**$_{(\uparrow --)}$	43.2	50.8	45.8	47.1	**46.7**$_{(\uparrow 1.1)}$	**62.8**$_{(\uparrow --)}$
DCAMA[ECCV2022] [26]		41.9	45.1	44.4	41.7	43.3	69.5	45.9	50.5	50.7	46.0	48.3	71.7
DCAMA w/ ABCNet		42.3	46.2	46.0	42.0	**44.1**$_{(\uparrow 0.8)}$	**69.9**$_{(\uparrow 0.4)}$	45.5	51.7	52.6	46.4	**49.1**$_{(\uparrow 0.8)}$	**72.7**$_{(\uparrow 1.0)}$

BMM			ACM	mIOU
RD-loss	Cross	Self		
				63.0
	✓			64.1
	✓	✓		64.9
✓	✓	✓		65.3
✓	✓	✓	✓	**66.0**

Table 3. Ablation studies of the proposed BMM and ACM modules.

Init	mIoU
Rand	64.3
Top-k	65.2
Learnable	65.4
SVD	**66.0**

Table 4. Ablation studies on different approaches for buoys initialization.

used for training and the rest one for testing. We randomly sample 1000 episodes for evaluation.

Evaluation Metric. Following previous works [28, 31, 38, 42], two evaluation metrics, i.e., mean intersection-over-union (mIoU) and foreground-background intersection-over-union (FBIoU) are adopted as our evaluation metrics for experiments. As mIou reflects the average results over all classes thus we mainly focus on the performance on the mIoU.

4.2. Implementation Details

Our proposed ABCNet can be easily implanted to amount of existing FSS models, and we mainly evaluate the effectiveness of our model on three baselines: PFENet [28], CyCTR [40] and DCAMA [26]. To be concrete, in DCAMA [26] and CyCTR [40], we adopt the proposed method to improve the pixel-wise cross-attention modules, while in PFENet [28], ABCNet is exploited to ameliorate the prior mask guidance (PMG). We use the middle-level features from the ImageNet [24] pretrained ResNet50 and ResNet101 [10] backbones as done in the original correlation module when using CyCTR [40] as our baseline. Different from that implementation, ABCNet is intergraded into the high-level support-query interaction when taking PFENet [28] and DCAMA [26] as the baseline.

All training settings are the same as the baseline methods, please refer to the **Supplementary Material** for more details. Two attention layers are adopted in our self-aggregation and cross-aggregation modules and the channel size in both of them is 512. The ϵ is set to be 0.1 and we limit the max iteration number to 3 which is sufficient to obtain good performance. The weight of representation decay loss λ_{rd} is set to be 0.1. All experiments are run on four NVIDIA Tesla V100 GPUs.

4.3. Results and Analysis

Quantitative results. Tabs. 1 and 2 present the performance comparison with and without ABCNet on three baseline methods and two benchmarks. It can be found that our ABCNet can consistently boost the performance of all three baseline methods with a considerable margin under all settings. Specifically, as shown in Tab. 1 on the ResNet-50, our approach improves PFENet by 1.5 and 1.7 mIoU on 1-shot and 5-shot, respectively. It shows that our ABCNet can significantly improve the quality of prior masks thus better guiding the query pixel classification. Besides, ABCNet brings CyCTR the performance gain of 1.0 & 1.2 mIoU and brings DCAMA the gain of 1.4 & 1.1 mIoU, respectively. It demonstrates that our ABCNet can ameliorate the pair-wise

Figure 5. Qualitative comparison with the baseline. Results with ABCNet can achieve more accurate segmentation

correlation in these methods thus obtaining better support information aggregation. In addition, on the deeper ResNet-101 backbone, all three methods also can achieve nearly 1% improvement. As for the more complex dataset COCO-20^i, the 1-shot and 5-shot results of models with the assistance of ABCNet respectively surpass the baselines. The performance boost is even more pronounced which proves the applicability of our method in complex scenarios. Please refer to Tab. 2 for details.

Qualitative results. To further analyze and understand the proposed method, We visualize a series of segmentation results on the most challenging split of PASCAL-5^i, as shown in Fig. 5. It can be noticed that the baseline model is prone to incorrectly segmenting the background or is unable to activate the whole object. We deem the main reason is that the false matches in the pair-wise pixel-level correlation lead to erroneous support information transfer, which interferes with the classification of query pixels. While the model with ABCNet generates more accurate prediction masks, which demonstrates the effectiveness of the proposed approach.

4.4. Ablation Study

A series of ablation studies are conducted to investigate the impact of our proposed ABCNet thoroughly. We mainly implement our experiments on PASCAL-5^i, the baseline method is DCAMA [26] and ResNet-50 is used as the backbone. All results are average mIoU across 4 splits unless otherwise specified.

Ablation study on BMM. Diagnostic experiments are conducted progressively to demonstrate the effectiveness of each submodule in the BMM as shown in Tab. 3. Note that the first line is the result of our ablation baseline, where we directly compare the buoys-level correlation via dot product, and the buoys are initialized by SVD. We can observe that the baseline performs worse than the original DCAMA. We deem the reason is that the buoys correlation based on the buoys not tailored for the current task is not beneficial or even damaging to the matching process. The introduction of the cross-aggregation module achieves a significant performance lift compared with the baseline, i.e., 1.1% mIoU. The improvements can be mainly ascribed to the abundant task-specific contextual information absorbed during the process of cross-attention. By further integrating the self-aggregation module, we observe a performance advance of 0.8% mIoU. We attribute this performance gain to the amended buoys via the self-aggregation module, which takes into consideration both the support and query information thus bridging the gap caused by large intra-class variance. Finally, when the representation decay loss (RD-loss) is adopted, the performance improves by 0.4% mIoU. This proves that the additional RD-loss can effectively prevent buoy degradation, thereby avoiding excessive deviation of the buoy correlation. In fact, the sum of the squares of the singular values after singular value decomposition (SVD) can be regarded as the energy of the matrix (*i.e.*, the representative information contained in the original correlation matrix), which is equivalent to its squared Frobenius norm. Therefore, the RD-loss could prevent information loss to guarantee the diversity of buoys.

Ablation study on ACM. As shown in Tab. 3, in the absence of ACM, the performance of the model is 0.7% worse than the complete ABCNet. This is because the dot product considers the similarity of each buoy individually and equally, ignoring the association between the buoys and pixels and the structural information inherent in the buoys. In contrast, the proposed ACM takes both aspects into account adaptively through a specially designed OT algorithm. In cooperation with well-established buoys, the ACM module can effectively suppress false matches on the premise of ensuring correct ones.

Ablation study on SVD-based initialization. To explore the effectiveness of different ways to initialize the buoys, we compare the performance of several intuitive initialization methods as shown in Tab. 4. Among these methods, *Rand* is conducted by randomly selecting k pixel features as buoys from the support and query features, respectively. In *Top-k*, the cumulative affinities based on the original similarity matrix are calculated on both support and query dimension, and only the top-k entries are kept as the buoys. *Learnable* means that all 2k buoys are randomly initialized but trainable and shared across tasks. Our SVD-based consistently outperforms these schemes by a considerable margin credited to the characteristics of SVD. Compared with other methods, SVD-based buoy initialization can better reduce information decay on the premise of keeping the same number of buoys.

Hyperparameter Evaluations. Quantitative experiments are conducted to clearly find a suitable number of buoys k. As shown in Fig. 7 (b), the performance continues to grow until $k = 24$ and then begins to decline if k keeps increasing. We deem the main reason is too few buoys cannot represent diverse semantic information, while too many buoys will produce undesirable redundancy. We then explore how λ_{rd} affects our model learning. It can be observed from Fig. 7 (c) that our model achieves much better performance when

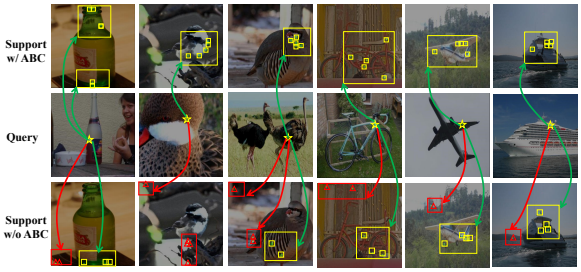

Figure 6. Visualization of pixel correspondence. The red arrows indicate the false matches occur in the model without ABCNet.

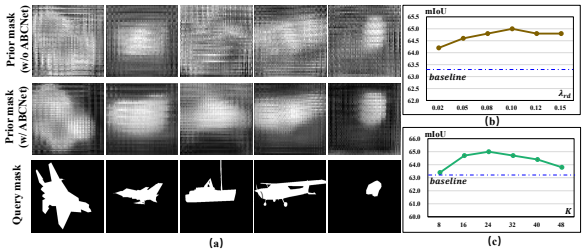

Figure 7. (a) Visualization of prior masks with and without ABCNet. (b) Hyperparameter experiments on the number of buoys (K). (c) Hyperparameter experiments on the weight of representative decay loss (λ_{rd}). Note that the hyperparameter experiments are conducted using the DCAMA on the most challenging split-2 of Pascal-5^i with ResNet-101 as the backbone.

$\lambda_{rd} = 0.1$.

4.5. Visualizations

Visualization of the Prior Masks. We visualize the prior masks originated from PFENet with and without ABCNet to evaluate the effect of the proposed method. As shown in Fig. 7(a), it can be observed that the prior masks from the baseline model cannot clearly indicate the target region as numerous false matches arise. While thanks to the adaptive suppression capability of ABCNet, the support background pixels are less likely to match with foreground ones in the query. The resulting prior masks are capable of indicating the location of the target more accurately so as to guide the classification of query pixels better.

Visualization of the Pixel-level Correspondence. To more intuitively demonstrate the noise suppression ability of our method, we visualize the pixel-wise correspondences using the models with and without our ABCNet, respectively. As shown in Fig. 6, we highlight the five most similar pixels of a foreground query pixel in the corresponding support image. We observe that in the absence of ABCNet, pixel features with distinct semantics sometimes share a high similarity, and these ambiguous support pixel features tend to transfer information to the query with a larger weight. Different from that, the top five pixels tend to line in the foreground of support images when we integrate the ABCNet into the baseline. This is in line with the core intention of ABCNet,i.e., filtering out as many false matches as possible.

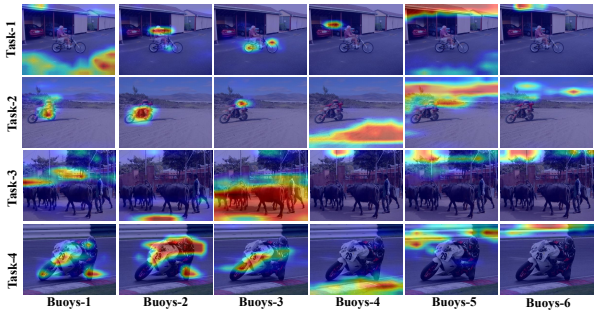

Figure 8. Visualization of activation regions of different buoys. As we can see, different buoys focus on diverse areas. This is because abundant semantic clues are retained in buoys thanks to BMM.

Visualization of Constructed Buoys. We further visualize the activation area of different buoys. As shown in Fig. 8, where we randomly select 5 from 64 buoys activation maps in different tasks. We can observe that diverse semantic regions are highlighted by different buoys. Another observation is that the key semantic clues in the current task scene can be well captured. For example, in the first-row of Fig. 8, important scene components such as human, bicycle, car, ground and roof are activated by various buoys. It should be noted that in some cases, different buoys focus on similar semantic regions as in the last two columns in Fig. 8, which further justifies the necessity of considering the structural similarity between related buoys in our proposed ACM.

5. Conclusion

In this paper, we propose to mitigate the false matches encountered in the pixel-level correlation modules in FSS by introducing a series of representative reference buoys. We design a novel plug-and-play ABCNet which includes a Buoys Mining Module (BMM) and an Adaptive Correlation Module (ACM). BMM is responsible for constructing multiple buoys that are tailored for specific tasks and ACM aims to conduct adaptive matching via the optimal transport algorithm. We conduct extensive experiments on three baseline methods and achieve consistent performance gains on two benchmarks in both 1-shot and 5-shot settings.

6. Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (Grant 62022078), National Defense Basic Scientific Research Program of China (Grant JCKY20200903B002).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-

end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3, 4

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[4] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 5

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] 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. 5, 6

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 2

[9] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European conference on computer vision*, pages 297–312. Springer, 2014. 5

[10] 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

[11] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 1, 3

[12] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019. 2

[13] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 2

[14] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8057–8067, 2022. 3

[15] 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, 2, 3

[16] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 2

[17] 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. 5, 6

[18] Yuanwei Liu, Nian Liu, Xiwen Yao, and Junwei Han. Intermediate prototype mining transformer for few-shot semantic segmentation. *arXiv preprint arXiv:2210.06780*, 2022. 3

[19] 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

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[21] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 2

[22] 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. 5

[23] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 2

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3, 6

[25] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 1, 2, 5

[26] 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. Springer, 2022. 1, 3, 6, 7

[27] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 2

[28] 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 & Machine Intelligence*, (01):1–1, 2020. 1, 3, 6

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[30] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 730–746. Springer, 2020. 1, 3

[31] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019. 6

[32] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022. 2, 3

[33] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 517–526, 2021. 3

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[35] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5475–5484, 2021. 3

[36] 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. 2

[37] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8312–8321, 2021. 1, 2, 3

[38] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019. 3, 6

[39] 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. 2

[40] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems*, 34:21984–21996, 2021. 1, 3, 6

[41] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *arXiv preprint arXiv:2210.06908*, 2022. 3

[42] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 50(9):3855–3865, 2020. 2, 3, 6

[43] Yifan Zhang, Bo Pang, and Cewu Lu. Semantic segmentation by early region proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1258–1268, 2022. 2

[44] 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. 2