Learning Robust Correlation with Foundation Model for Weakly-Supervised Few-Shot Segmentation

Xinyang Huang^a, Chuang Zhu^{a,*}, Kebin Liu^a, Ruiying Ren^a, Shengjie Liu^a

^aSchool of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

Abstract

Existing few-shot segmentation (FSS) only considers learning support-query correlation and segmenting unseen categories under the precise pixel masks. However, the cost of a large number of pixel masks during training is expensive. This paper considers a more challenging scenario, weakly-supervised few-shot segmentation (WS-FSS), which only provides category (i.e. imagelevel) labels. It requires the model to learn robust support-query information when the generated mask is inaccurate. In this work, we design a Correlation Enhancement Network (CORENet) with foundation model, which utilizes multi-information guidance to learn robust correlation. Specifically, correlation-guided transformer (CGT) utilizes self-supervised ViT tokens to learn robust correlation from both local and global perspectives. From the perspective of semantic categories, the class-guided module (CGM) guides the model to locate valuable correlations through the pre-trained CLIP. Finally, the embedding-guided module (EGM) implicitly guides the model to supplement the inevitable information loss during the correlation learning by the original appearance embedding and finally generates the query mask. Extensive experiments on PASCAL- 5^i and COCO- 20^i have shown that CORENet exhibits excellent performance compared to existing methods. Our code will be available soon after acceptance.

^{*}Corresponding author

Email: hsinyanghuang7@gmail.com; czhu@bupt.edu.cn



Fig. 1. Comparison between (a) few-shot segmentation (FSS) task [7], (b) weaklysupervised few-shot classification and segmentation (WS-FCS) task [15], and (c) our weakly-supervised few-shot segmentation (WS-FSS) task settings. (a) The FSS task requires many support-query masks during training. (b) The classification and segmentation tasks are decoupled in the WS-FCS task. It provides supervisory information on whether images belong to the same category without providing specific category assistance for segmentation. (c) The WS-FSS task assists the model in segmentation through specific categories of supervised information in the presence of noise in the mask generated by the model.

1. Introduction

Few-shot learning [1, 2, 3, 4, 5, 6] is a machine learning method that uses very little labeled data to help the model quickly adapt to new tasks or categories. It is crucial in applications where data collection is costly or requires intensive annotation, such as image segmentation. Consequently, few-shot segmentation (FSS) has been proposed and extensively studied [7, 8, 9, 10, 11, 12, 13, 14].

Existing FSS methods are typically trained based on the meta-learning paradigm [16, 17, 18, 19, 20]. They often assume the presence of a large amount of accurately annotated data for model training and learn the supportquery correlation by abundant support and query masks, as shown in Fig. 1a. Likewise, during testing, several ground-truth (GT) support masks are required during the reference of the model. However, the cost of obtaining the segmentation masks required for these models is very expensive and cumbersome.

Although related works [21, 22] have explored the setting of few-shot segmentation in weakly-supervised scenarios, they are unable to generate supervised masks for unseen categories during the testing phase or require additional training in a mask generation module. CST [15] solved this problem and proposed the weakly-supervised few-shot classification and segmentation (WS-FCS), as shown in Fig. 1b. However, there are two issues when directly applying it to WS-FSS tasks: firstly, it simply considers the correlation information of the support-query pair in the presence of the GT mask, which may introduce a lot of matching noise in the case of inaccurate masks; secondly, the provided category information is whether the two images belong to the same category, ignoring the benefits of semantic level information for segmentation. [23] is closest to the problem setting of this paper, but it ignores the exploration of robust correlation and the contribution and role of the foundation model in the WS-FSS task. Similar to [23], this paper focuses on a weakly-supervised few-shot segmentation (WS-FSS) scenario where the model should learn robust support-query matching information and perform segmentation on query images with only image-level category information and no access to GT masks, as depicted in Fig. 1c.

To solve the WS-FSS, this paper introduces a Correlation Enhancement Network (CORENet) with foundation model assistance that helps the model learn robust correlation from multiple perspectives, even in the presence of inaccurate masks generated by the model. Specifically, we first design a Correlation-Guided Transformer (CGT), which takes high-quality tokens obtained from a self-supervised Vision Transformer (ViT) [24] as input. It fuses information from local and global perspectives to guide the model in better utilizing correlation information. Therefore, CGT can be relatively robust in the face of generated imprecise masks. Furthermore, a well-designed self-distillation loss helps CGT generate higher-quality correlation maps in the early stages. However, when the model generates inaccurate masks, the effect of segmenting the query from the perspective of correlation is limited. To address the above issues, the Class-Guided Module (CGM) helps the model to roughly locate specific objects from inaccurate masks using prior knowledge by using the provided class information. Although existing works [25, 15] utilize category supervision information by classifying support and query during FSS. They provide category information to support and query whether the images belong to the same category (0/1 la)

bel) without providing specific category semantic information assistance for segmentation. By using pre-trained CLIP [26] to generate coarse attention, CGM utilizes existing correlation features to filter out background features unrelated to the query foreground, helping the model roughly locate valuable correlation information. Finally, to further reduce potential information loss during correlation processing and implicitly guide the model in refining matching information, we propose an Embedding-Guided Module (EGM). EGM uses efficient tokens generated by ViT to supplement the information of the original embeddings, resulting in the final masks.

To generate supervised masks, inspired by previous works [24, 15], the paper utilizes attention maps generated by pre-trained self-supervised ViT to create pseudo-masks. Furthermore, we leverage pixel relationships within the image to generate more accurate pseudo-masks through the Pixel-Adaptive Refinement (PAR) module [27], which helps the model learn robust correlations from the perspective of mask enhancement. Even when encountering unseen categories during testing, the model can provide relatively accurate pseudo-masks. Our main contributions are summarized as follows:

- We propose a Correlation Enhancement Network (CORENet) with foundation model assistance to guide models from multiple perspectives to learn robust correlation in WS-FSS.
- We propose a Correlation-Guided Transformer (CGT) that learns to support-query knowledge from a knowledge aggregation perspective and apply the Pixel Adaptive Refinement (PAR) module in a few-shot scenario for the first time.
- We propose a Class-Guided Module (CGM) and an Embedding-Guided Module (EGM) to mine and supplement target information in correlation features from category semantics and appearance embedding perspective.
- Our CORENet achieved state-of-the-art results compared to the latest FSS and WS-FSS methods in two WS-FSS scenarios (*i.e.* PASCAL- 5^i and COCO- 20^i).

The remainder of this paper is as follows: Section 2 reviews recent work related to WS-FSS. Sections 3 and 4 elaborate on the entire process of our proposed CORENet. Then, comprehensive quantitative and qualitative results are reported in Section 5, followed by a series of ablation studies. Finally, Section 6 gives the conclusion of this work.

2. Related Work

Few-Shot Semantic Segmentation. Few-shot semantic segmentation (FSS) aims to segment new semantic objects in images, with only a few densely labeled examples available. The current methods mainly focus on the improvement of the meta-learning stage. They can be classified as prototype-based methods and relational-based methods. The intuition behind the prototype-based methods [28, 8, 9, 29, 30, 31, 14] is to extract representative foreground or background prototypes from the supporting samples using the method, and then use different strategies to interact between different prototypes or between prototypes and query features. Relational-based methods [10, 11, 12, 25, 15, 13] have also achieved great success in the fewshot semantic segmentation. However, these methods only focus on learning to support and query matching information between images under precise supervision. This paper considers a more challenging weak supervision version of FSS, which completes the segmentation of query images without providing any mask information, only providing support images and category information.

Weakly-Supervised Few-Shot Segmentation. Due to the severe challenge of data scarcity, many works currently study few-shot segmentation in a weakly-supervised environment. However, the definition of weakly supervised few-shot segmentation (WS-FSS) in existing methods is still flawed and inconsistent. WS Co-FCN [32] generated a pseudo-mask to support the image by retaining pixels not classified as background. However, it cannot handle supporting images that contain multiple new classes. Some methods [21, 33] use supervision information such as bounding boxes. WRCAM [22] requires pre-training of a mask generation module for all image categories in advance, including test image categories that have not been seen during the training phase, which does not follow the training paradigm of few-shot learning during the training stage. The problem setting of CST [15] is similar to that of this paper. However, the provided category information is whether the two images belong to the same category and does not provide specific category assistance for segmentation. [23] is closest to the problem setting of this paper, but this paper focuses on exploring the contribution

and role of the foundation model in the WS-FSS task. This paper focuses on the few-shot segmentation in a weakly-supervised scenario, where no GT mask information is provided at any stage. It provides category information assistance to complete the segmentation of the query image.

3. Problem Definition

Similar to the few-shot segmentation [7, 8, 9, 10, 11, 12, 13], in order to avoid overfitting risks caused by insufficient training data, we adopted a widely used meta-learning method called episodic training [34]. In weaklysupervised few-shot segmentation, we define two datasets, D_{train} and D_{test} , with category sets C_{train} and C_{test} respectively, where $C_{train} \cap C_{test} = \emptyset$. The model trained on D_{train} is directly transferred to D_{test} for evaluation and testing. We train the model in an episode manner [34]. Under the weak-supervised setting, each episode only comprises support set $S = \{I_s\}$, query set $Q = \{I_q\}$, and their corresponding category c. Unlike few-shot segmentation, we do not provide mask information at any stage. Under the K-shot setting, it includes the support set $S = \{I_s^i\}_{i=1}^K$, query set $Q = \{I_q\}$ and the corresponding category c. Training set D_{train} and test set D_{test} means $D_{train} = \{I_s^i, I_q^i, c\}_{i=1}^{N_{train}}$ and $D_{test} = \{I_s^i, I_q^i, c\}_{i=1}^{N_{test}}$, where N_{train} and N_{test} is a series of quantitative training and testing. During training, the model iteratively samples an episode from D_{train} to generate a pseudo-mask using limited information and to learn segmentation knowledge through the generated pseudo-masks. During the testing, the model changed from D_{test} randomly samples $\{I_s^i, I_a^i, c\}$ to predict the query mask.

4. Methodology

4.1. Overview

As shown in Fig. 2, the Correlation Enhancement Network (CORENet) is composed of three key modules, namely, correlation-guided transformer (CGT), class-guided module (CGM), and embedded-guided module (EGM). Precisely, we extract high-quality features through pretrained DINO ViT [24] and calculate the correlation between the token of the supporting image pair and the query image pair. Then, the robust cross-correlation information is learned from the local and global perspectives through CGT. With the assistance of CLIP [26], CGM uses category information to guide the generation of a coarse attention map and filters out the irrelevant information in the



Fig. 2. The overall architecture of our Correlation Enhancement Network (CORENet). Firstly, the **Correlation-Guided Transformer (CGT)** is introduced to generate robust correlation features using the local and global similarity calculations of ViT tokens. Then, with the assistance of CLIP, the **Class-Guided Module (CGM)** transforms the category information into coarse attention and further refines them to filter irrelevant information in the relevant features. Meanwhile, the **Embedding-Guided Module (EGM)** combines the support query appearance of each layer with the enhanced correlation features, further reducing the potential information loss of the model in correlation-enhanced learning under weakly-supervised settings and obtaining the final query mask.

query features through the generated cross-correlation features. To reduce the potential information loss of the model in correlation reinforcement, we propose EGM, which further aggregates the matching information by using the embedded information obtained from the feature graph to guide the model to learn the matching information implicitly. Then, the model sends the learned robust features into the segmentation header to predict the final segmentation mask \tilde{M}_q of the query image. Next, each module will be described in detail in the following paragraphs.

4.2. Correlation-Guided Transformer

The correlation between support and query plays a crucial role in FSS. The existing methods [12, 25, 15] help the model segment the query image on the existing support foreground information by using the similarity between the support and query image pixels. However, due to the lack of a GT mask, it is not comprehensive to only consider the correlation information of this local-to-local matching. In this paper, the correlation-guided transformer (CGT) is proposed. From the perspective of local-to-local and local-to-global, CGT uses the features extracted by self-supervised pretrained ViT to learn the multi-view robust correlation information between support images and query images.



Fig. 3. Illustration of Multi-kernel information fusion in CGT.

Local-to-local correlation. Specifically, CGT uses DINO [24] as the backbone of pretrained frozen ViT. It gets K-layers patch tokens f_q , f_s and class tokens $f_{q,cls}$, $f_{s,cls}$ by inputting support images and query images and through multi-head attention. Then, we calculate the local-to-local (*i.e.* pixel-to-pixel) correlation between the query and support patch tokens in each layer and preserve the semantic diversity of the M heads of the ViT, *i.e.*, we calculate the $M \times K$ cosine similarities of the query to support tokens and concatenate them along the new dimension:

$$C_{local} = \frac{(f_q)^T f_s}{||f_q||||f_s||} \in \mathbb{R}^{MK \times h_q w_q \times h_s w_s},\tag{1}$$

where $h_s w_s$ and $h_q w_q$ represent the product of length and width of supports and query images, $|| \cdot ||$ means l_2 regularization.

Local-to-global correlation. From the global view, we use the support mask to cut out the foreground and background regions from f_s . Unlike the foreground area, which is cut off as a whole area, the background area is divided into N local areas because the background may not be uniform. To this end, we use the Voronoi-based method [35, 36] to divide the background into N different regions. Then, the global features of foreground and background

are obtained by mask average pooling:

$$p_{f} = \frac{1}{|M_{s}|} \sum_{i=1}^{h_{s}w_{s}} f_{s,i}M_{s,i},$$

$$p_{b,n} = \frac{1}{|B_{s}^{n}|} \sum_{i=1}^{h_{s}w_{s}} f_{s,i}B_{s,i}^{n},$$
(2)

where M_s is the pseudo-mask for support images, its generation will be introduced in Section 4.5. $B_s^n = 1 - M_s$ is the *n*-th background mask for the support mask. Similar to Eq. 1, the local-to-global correlation between the query and support token is calculated as follows:

$$C_{global}^{f} = \frac{(f_q)^T p_f}{||f_q||||p_f||} \in \mathbb{R}^{MK \times h_q w_q \times 1},$$

$$C_{global,n}^{b} = \frac{(f_q)^T p_{b,n}}{||f_q||||p_{b,n}||} \in \mathbb{R}^{MK \times h_q w_q \times N},$$
(3)

where N = 5 is the number of the background. We further concatenate the features to obtain the correlation token $\mathbf{C}_i^0 \in \mathbb{R}^{(1+N+h_sw_s)\times ML}$, where $i \in [1, \dots, h_qw_q]$ is an index over the query token and L is the number of the transformer layers. The correlation token refers to the token obtained after feeding the correlation map into the transformer. Then following CST [15], it takes \mathbf{C}_i^0 and support mask M_s as input and returns three types of token: foreground, background, and local correlation token through a twolayer transformer [37]. Each transformer layer can be described as follows:

$$\mathbf{C}_{i}^{l\prime} = \mathrm{LN}_{l}(\mathrm{MHSA}_{l}(\mathbf{C}_{i}^{l}, M_{s,i}) + \mathbf{C}_{i}^{l}),$$

$$\mathbf{C}_{i}^{l+1} = \mathrm{LN}_{l}(\mathrm{MLP}_{l}(\mathbf{C}_{l}^{l\prime}) + \mathbf{C}_{i}^{l\prime}) \in \mathbb{R}^{C_{l} \times h_{q}w_{q} \times h_{l}w_{l}},$$

(4)

where l means the transformer layer index, C_l means its dimension, and LN_l , MHSA_l , MLP_l correspond to a multi-head self-attention (MHSA) [37], a group normalization [38], and a linear layer, respectively. Similar to related works [39, 15], in each MHSA layer, the generated query is embedded into a spatial pool, and the output size changes from $h_s w_s$ to 1. Then, we split the tensor **C** into foreground, background, and local correlation token along the second dimension, *i.e.* $\mathbf{C}_{global}^f, \mathbf{C}_{global,n}^b, \mathbf{C}_{local}$.

Fore-background fusion. After obtaining the global correlation tokens, we propose an adaptive fusion method for different global foreground and

background features. For different backgrounds, we select them with adaptive weighting, which consists of a simple linear layer.

$$\mathbf{C}^{b}_{global} = w_1 \mathbf{C}^{b}_{global,1} + \dots + w_n \mathbf{C}^{b}_{global,n} + \beta, \tag{5}$$

where w_n means the *n*-th weight of the linear layer, and β means the bias. Then, we fuse the merged background and foreground correlation features by a convolutional layer:

$$\mathbf{C}_{global} = \operatorname{Conv}_1(\operatorname{Cat}(\mathbf{C}^f_{global}, \mathbf{C}^b_{global})), \tag{6}$$

where $Cat(\cdot, \cdot)$ is the concatenation operation. This method can help the model integrate necessary support knowledge from different backgrounds and foregrounds.

Multi-kernel information fusion. We use the multi-kernel information fusion mechanism after obtaining the local and global features. Multikernel information fusion uses different receptive field convolution kernels to fuse the local and global correlation information, reducing the noise of different matching information due to the lack of GT masks. We process the features by concatenating two parts of the features, utilizing convolutional kernels of different receptive fields, and helping the model learn robust knowledge:

$$\mathbf{C}_{i} = \operatorname{Conv}_{i}(\operatorname{Cat}(\mathbf{C}_{local}, \mathbf{C}_{global})), \tag{7}$$

where Conv_i means the $i \times i$ convolutional operation and $i \in \{1, 3, 5, 7\}$. Then, we will integrate the obtained feature knowledge of different receptive fields:

$$\mathbf{C}' = \operatorname{Conv}_1(\operatorname{Cat}(\mathbf{C}_1, \mathbf{C}_3, \mathbf{C}_5, \mathbf{C}_7)) + \mathbf{C}_{local}.$$
(8)

Finally, the final correlation token is obtained through the residual connection layer [40]:

$$\mathbf{C}_{fusion} = \operatorname{Conv}_3(\mathbf{C}') + \mathbf{C}' \in \mathbb{R}^{C \times h_q \times w_q}.$$
(9)

Self-distillation loss. We propose a self-distillation loss for our CGT to help the model generate higher-quality robust correlation diagrams in the early stage. We average the feature dimensions for the correlation map of each layer to get the correlation map $\hat{\mathbf{C}} \in \mathbb{R}^{h_q \times w_q}$, and use the high-level correlation map to guide the low-level correlation feature map, as follows:

$$L_{distill} = \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{h_l w_l} \zeta_l(\hat{\mathbf{C}}_{local,i}^{l+1}) \cdot \log \frac{\zeta_l(\hat{\mathbf{C}}_{local,i}^{l+1})}{\hat{\mathbf{C}}_{local,i}^l}, \tag{10}$$

where $\zeta_l(\cdot)$ is the resize function of the *l*-th layer, *L* is the number of the layers and $h_l w_l$ is the the product of length and width of the *l*-th layer. The guidance of the high-level correlation graph to the low-level feature graph helps the model retain the fine-grained segmentation quality, reduces the impact of noise, and does not discard the context information [41], which can help learn robust correlations.

4.3. Class-Guided Module

The knowledge learned by the model from the correlation between support and query is limited, especially in the case of imprecise support masks in the WS-FSS scenarios. To further assist the model in filtering potential noise in correlation features, we use additional category information from the perspective of category semantics to help the model locate more valuable correlation information. Pre-trained CLIP [26] has been proven to generate relatively coarse CAM based on category information by using Grad-CAM [42, 43]. After large-scale pre-training, CLIP already has powerful zero-shot learning capabilities. Even without seeing specific supervision labels during training, CLIP is able to understand and generate output for tasks for which it was not explicitly trained [44, 45]. This paper utilizes this to construct the CGM that helps the model roughly locate the approximate positions of the objects that need to be segmented.

To simplify our method, this paper will not discuss obtaining more accurate masks for CLIP. Instead, we will choose a simple mask generation method and discuss utilizing the generated coarse masks. CGM can also be seen as a simple zero-shot method, and it does not rely on various complex cue engineering and other zero-shot models but can still achieve satisfactory performance.

We first input the query image and its category prompt "a photo of [class]", where class represents its corresponding category c, into the pretrained CLIP and then use Grad-CAM to obtain a coarse attention \mathbf{A}_c . Next, we multiply \mathbf{A}_c by the obtained correlation token \mathbf{C}_{fusion} and use \mathcal{F}_{CGM} to refine the attention:

$$\mathbf{A}_{r} = \mathcal{F}_{CGM}(\mathbf{C}_{fusion} \otimes \zeta(\mathbf{A}_{c})) \in \mathbb{R}^{h_{q} \times w_{q}}, \tag{11}$$

where \mathcal{F}_{CGM} consists of two convolutional layers and a sigmoid function, \otimes is the Hadamard product and $\zeta(\cdot)$ is the resize function. Finally, we combine the features with the attention \mathbf{A}_r to obtain filtered correlation features that

discard irrelevant background information:

$$\mathbf{C} = (\mathbf{C}_{fusion} \otimes \mathbf{A}_r) \oplus \mathbf{C}_{fusion}, \tag{12}$$

where \oplus stands for the element-wise sum. Combined with backpropagation, the parameters of \mathcal{F}_{CGM} in CGM are updated. Therefore, the refined \mathbf{A}_r can focus more on the objects that need to be segmented based on the \mathbf{A}_c generated by CLIP. Through the coarse-to-fine training strategy, when the model encounters unfamiliar categories, even without the precise support of mask supervision, it can combine the powerful zero-shot capability of CLIP to capture the approximate location of the segmented object.

4.4. Embedding-Guided Module

Towards the goal of reducing the potential information loss of the model in correlation-enhanced learning under weakly-supervised segmentation settings, we suggest embedding the original appearance of each layer obtained from support and query feature maps into the decoder for further aggregation to implicitly guide the model in utilizing the learned robust support-query matching information.

First, add the features of each layer and project them:

$$F_{s} = \mathcal{F}_{Proj.}(\sum_{k=1}^{K} f_{s}^{k}),$$

$$F_{q} = \mathcal{F}_{Proj.}(\sum_{k=1}^{K} f_{q}^{k}),$$
(13)

where $\mathcal{F}_{Proj.}$ denotes 1×1 convolution and K means the layer number of the backbone ViT. Then they are concatenated to the similarity feature $\tilde{\mathbf{C}}$, and the final prediction mask \tilde{M}_q is obtained through the EGM composed of two layers of transformers [46] and a segmentation header:

$$\tilde{M}_q = EGM(Cat(\tilde{\mathbf{C}}, F_s, F_q)).$$
(14)

The original appearance information implicitly helps the model reduce information loss in learning robust correlation. Meanwhile, the appearance embedding information is an effective guide for filtering noise in matching scores [47, 48, 49], while self-supervised pre-trained ViT can provide an efficient multi-layer feature. It helps the model learn the relevant information obtained in the presence of certain mismatches through embedding guidance at each layer.

4.5. Training Objective

Pseudo-mask. As demonstrated in previous studies [50, 24, 15], query key attention maps can capture semantically significant foreground objects. Inspired by this, we generate a pseudo-GT mask for dynamic queries and image support by calculating the cross attention of the last ViT layer:

$$M_{s,i}^{m} = \frac{(f_{s,i}^{m})^{T} f_{q,cls}^{m}}{||f_{s,i}^{m}||| |f_{q,cls}^{m}||} \in \mathbb{R}^{h_{s}w_{s} \times 1},$$

$$M_{q,i}^{m} = \frac{(f_{q,i}^{m})^{T} f_{s,cls}^{m}}{||f_{q,i}^{m}||| |f_{s,cls}^{m}||} \in \mathbb{R}^{h_{q}w_{q} \times 1},$$
(15)

where $f_{q,cls}^m$, $f_{s,cls}^m$ means *m*-th head query or support class token. Meanwhile, we use the Pixel-Adaptive Refinement (PAR) module [27] to generate pseudomasks based on the relationship information between various pixels within the image, generating more accurate supervision information:

$$M_{s,i} = \mathbb{1}(PAR(\zeta(\frac{1}{M}\sum_{m=1}^{M}M_{s,i}^{m}) > \alpha)),$$

$$M_{q,i} = \mathbb{1}(PAR(\zeta(\frac{1}{M}\sum_{m=1}^{M}M_{q,i}^{m}) > \alpha)),$$
(16)

where $\alpha = 0.4$ is the prediction threshold and $\mathbb{1}(\cdot)$ is the indicator function. Unlike existing WS-FSS methods [22], our mask generation module also applies to unseen categories without additional training stages.

Training loss. There are two parts of training loss: segmentation loss and self-distillation loss. The segmentation loss L_{seg} is calculated by the final prediction \tilde{M}_q and M_s using the cross-entropy function. The self-distillation loss is obtained from Eq. 10. The final loss is:

$$L = L_{seg} + \lambda_{distill} L_{distill}, \tag{17}$$

where $\lambda_{distill}$ is the balance parameter set to 0.5. The whole training process for CORENet is summarized in Algorithm 1.

5. Experiments

In this section, we evaluate the proposed method, compare it with recent state-of-the-art, and provide in-depth analyses of the results of the ablation study.

Algorithm 1: Training Process for CORENet.

Input: A training set D_{train} and a training category set C_{train} . **Output:** The final trained model ϕ . for each episode $(S, Q) \in D_{train}$ and category $c \in C_{train}$ do Extract features by pretrained DINO ViT. Generate the pseudo-mask using Eq.15 and 16. # Correlation-Guided Transformer Compute the local-to-local and local-to-global correlation using Eq. 1, 2 and 3. Obtain foreground, background, and local correlation tokens using Eq. 4. Obtain the final correlation token C_{fusion} using Eq. 5, 6, 7, 8 and 9. Compute the self-distillation loss $L_{distill}$ as in Eq. 10. # Class-Guided Module Obtain filtered correlation feature C using Eq. 11 and 12. # Embedding-Guided Module Predict the query mask M_q using Eq. 13 and 14. Compute the final loss L as in Eq. 17. Compute gradients and optimize via SGD. end **Return** the final trained model ϕ .

5.1. Experimental Settings

Datasets. To evaluate our method, experiments are conducted on two commonly used few-shot segmentation datasets, PASCAL-5^{*i*} and COCO-20^{*i*}. PASCAL-5^{*i*} is created according to PASCAL VOC 2012 [51] with additional notes of SBD [52]. A total of 20 classes in the dataset are evenly divided into four folds $i \in \{0, 1, 2, 3\}$, and each fold contains five classes. COCO-20^{*i*} is proposed by [53] and is based on MSCOCO [54]. Similar to PASCAL-5^{*i*}, the 80 classes in COCO-20^{*i*} are divided into four folds, and each fold contains 20 classes.

Evaluation metrics. We use union average intersection (mIoU) as our evaluation indicators. The mood indicator averages the IoU values of all classes in the fold: $\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \text{IoU}_c$, where C is the number of classes in the target fold and IoU_c is the intersection on the union of class c. Because mIoU better reflects the generalization ability and prediction quality of the

Performance of PASCAL-5^{*i*} [51] in mIoU. The superscript * indicates that the model is trained on the pseudo-mask generated by CST [15]. Bold numbers indicate the best performance, and <u>underlined</u> numbers indicate the second best.

			1-shot					5-shot		
Methods	5^{0}	5^{1}	5^{2}	5^3	mean	5^{0}	5^{1}	5^{2}	5^{3}	mean
HSNet^* [12]	47.6	45.4	41.0	<u>37.0</u>	42.8	48.0	46.1	41.6	<u>37.3</u>	43.3
$ASNet^*$ [25]	<u>49.0</u>	44.6	43.8	35.2	43.2	<u>50.1</u>	45.6	45.0	35.8	<u>44.1</u>
$MIANet^*$ [13]	44.9	34.3	41.2	35.9	39.1	46.4	45.1	46.5	36.4	43.6
CST [15]	48.2	45.1	42.4	34.6	42.5	49.5	45.5	42.8	35.1	43.2
CORENet (Ours)	50.3	51.6	47.6	39.4	47.2	50.7	51.8	47.8	39.6	47.5

model, we mainly focus on mIoU in our experiments.

Implementation details. To compare with previous works based on ResNet50 [12, 25, 13], we use the ViT-small backbone [55]. The feature extraction backbone network conducts self-monitoring and pre-training on ImageNet 1K [56] through DINO [24]. Following the CST [15], the reason for choosing this ViT is that its training data size and the number of model parameters are similar to ResNet50 [40]. It is also trained on ImageNet 1K but uses class labels as supervision. The backbone of CLIP is ResNet101. However, our framework based on DINO and CLIP can easily replace the backbone network with a foundation model such as ViT-G/14 [57] with a huge parameter amount (2.5B), which distinguishes our method from existing methods. As in the previous works [12, 25, 15], the backbone is frozen during training. The learning rate is initialized to 0.0005, the batch size is 16, and the additional layer is trained using Adam [58]. The loss balance parameter $\lambda_{distill}$ is set to 0.5, and the number of backgrounds N is set to 5. Consistent with CST [15], our CORENet uses a 1-way 1-shot segment for training and any N-way K-shot inference.

5.2. Comparison with State-of-the-Arts.

Due to the lack of supervision masks, the existing FSS model cannot be directly migrated to the WS-FSS scenario. To better compare existing FSS methods, we combine them with the mask generation method in CST [15] to generate supervised information and guide the model in predicting the final query mask. We label them with the superscript *.

PASCAL-5ⁱ. Table 1 compares mIoU performance between our method and existing representative models. From this, it can be seen that: (i)

Performance of COCO- 20^i [53] in mIoU. The superscript * indicates that the model is trained on the pseudo-mask generated by CST [15]. **Bold** numbers indicate the best performance, and <u>underlined</u> numbers indicate the second best.

			1-shot					5-shot		
Methods	20^{0}	20^{1}	20^{2}	20^{3}	mean	20^{0}	20^{1}	20^{2}	20^{3}	mean
HSNet^* [12]	19.9	22.5	22.1	<u>23.0</u>	21.9	21.0	24.2	22.7	23.8	22.9
$MIANet^*$ [13]	20.5	22.8	21.6	22.3	21.8	<u>21.5</u>	23.2	21.8	22.5	22.3
CST [15]	<u>21.0</u>	21.9	22.4	22.5	<u>22.0</u>	21.3	22.1	22.7	22.6	22.1
CORENet (Ours)	22.1	22.8	22.3	23.4	22.7	22.3	24.7	22.6	24.0	23.4



Fig. 4. Qualitative results of our CORENet on PASCAL- 5^i and COCO- 20^i benchmarks. Zoom in for details.

CORENet achieved state-of-the-art performance in both 1-shot and 5-shot settings. Compared to the recent FSS model MIANet [13] and the weaklysupervised classification & segmentation model CST [15], we have improved by 8.1%, 4.7% (1-shot), and 3.9%, 4.3% (5-shot), respectively. (ii) MIANet has performed poorly in certain situations, which holds the previous stateof-the-art results of FSS. This is because in WS-FSS scenarios, the generated mask contains noise, and excessive dependence on correlated features with noise can lead to a decrease in model performance. This also confirms that the method proposed in this paper can effectively handle scenarios with mask noise in weakly-supervised few-shot segmentation.

 $COCO-20^i$. $COCO-20^i$ is a more challenging dataset containing multiple objects and more significant variance. Table 2 shows the performance

Baseline	CGT	CGM	EGM	PAR	1-shot	5-shot
\checkmark					42.5	43.2
\checkmark	\checkmark				43.1	43.9
\checkmark	\checkmark	\checkmark			44.0	44.7
\checkmark	\checkmark	\checkmark	\checkmark		45.1	45.8
\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	47.2	47.5

Ablation studies of main model components. Bold numbers indicate the best performance.

comparison of mIoU. Overall, the mean mIoU of MIANet in the 1-shot and 5-shot settings surpasses all previous methods. Under the 1-shot setting, our CORENet exceeded MIANet and CST by 0.9% and 0.7%. This proves the superiority of our method despite the challenging scenarios.

Qualitative results. Fig. 4 reports quantitative results from CORENet and baseline models based on PASCAL-5^{*i*} and COCO-20^{*i*} benchmark tests. We can see that CORENet performs well in capturing object details. For example, more subtle details are retained in segmenting dogs and cars.

5.3. Ablation Study

We conducted extensive ablation studies of PASCAL- 5^i to verify the effectiveness of the critical modules (CGT, CGM, and EGM) we proposed. In addition, we provide experimental details and additional experiments in the supplementary materials.

Components analysis. Our CORENet consists of four key components: Correlation-Guided Transformer (CGT), Class-Guided Module (CGM), Embedding-Guided Module (EGM), and Pixel-Adaptive Refinement (PAR). Table 3 shows our validation of the effectiveness of each component. PAR can further reduce imprecise noise in masks and achieve a performance improvement of 2.1% in 1-shot by fully utilizing the information of surrounding pixels. EGM is an essential component of our model, which increases mIoU by 1.1% in 1-shot. CGT and CGM are also indispensable. By combining all three modules, CORENet achieves state-of-the-art performance.

Main components in CGT. CGT constructs local-to-global correlations to help the model fully understand matching information. Table 4 shows the impact of each element in CGT on model performance. "FBC" means the fore-background concatenation, "FBF" means the fore-background fusion, "SIF" denotes the single-kernel information fusion, and "MIF" denotes the multi-kernel information fusion. We can see that the fusion using adaptive

Ablation studies of main components in CGT. **Bold** numbers indicate the best performance.

Table 8	5
---------	---

Analysis of background regions in CGT.

1-shot

47.0

47.2

47.1

47.2

Background regions

N = 4

N = 5

N = 6

N = 7

FBC	FBF	SIF	MIF	1-shot	5-shot
\checkmark		\checkmark		46.6	46.9
\checkmark			\checkmark	46.9	47.1
	\checkmark	\checkmark		46.7	47.2
	\checkmark		\checkmark	47.2	47.5

Table 6			
Ablation	studies	of	CGM.

CLIP	Refinement	1-shot
		46.5
\checkmark		46.8
\checkmark	\checkmark	47.2

Table 7

Analysis of projection dimension in EGM.

Dimension	1-shot
32	46.8
64	47.2
128	46.9
384	47.1

weights is 0.3% better than the method of directly concatenating background features along the channel. In addition, by establishing path information between different receptive fields, the proposed multi-kernel information fusion method is compared to the single-kernel fusion method (using only 3×3 convolutions), which can further reduce the impact of mismatches on the model and achieve better performance. Through the proposed fusion mechanism, CGT can help the model learn more robust correlation with more information while support mask inaccuracies.

Number of background regions for CGT. In Section 4.2, we mentioned that the Voronoi-based method [35, 36] helps the model learn complex background correlation knowledge by dividing different background regions. We further discuss the impact of this region on the final results of the model, as shown in Table 5. It can be seen from the results in the table that the model results are relatively robust for different numbers of regions. This shows that our CGT can perform robust correlation modeling for different complex background knowledge, which can help the model still learn valuable correlation knowledge when facing the generated imprecise masks.

Table 9

Analysis of differences in projection operations of EGM.

Performance of	differences	with	related	methods	[23]
in PASCAL-5	<i>i</i> .				

Dimension	1-shot	Method	1-shot	5-shot
Concatenation	46.9	Pixel-level meta-learner [23]	42.4	45.5
Sum	47.2	CORENet (Ours)	47.2	47.5

Refinement of CGM. In Section 4.3, we mentioned using pre-trained CLIP [26] assisted models for segmentation by constructing CGM. We further discuss the necessity of pre-trained CLIP and the proposed enhancement module, as shown in Table 6. When no additional components are added, the model directly feeds the correlation features obtained through CGT into the EGM module. For the different approaches to fusing attention map of CLIP, we compare the approach of simply fusing it into the original correlation features (denoted as "CLIP") with the approach of fusing it by designing additional learnable refinement modules (denoted as "Refinement"). Due to the zero-shot capability of CLIP, the performance of the model can be improved to a certain extent when only CLIP is used to weight features. The experimental results show that after the refinement module, the model can be tirrelevant background areas and achieve the best results.

Differenct backbone of CLIP in CGM. We use the GradCAM of the pre-trained CLIP in the CGM module to generate initialization attention maps. As shown in Fig. 5, we visualized the thermal maps obtained by CLIP for different backbones. When using a deeper network as the backbone of CLIP, the initial attention map obtained is visually better. A better initial attention map can help the model focus on more critical correlation information. Therefore, we chose ResNet101 [40] as the backbone of our CGM module.

Projection dimension of EGM. In Section 4.4, we mentioned projecting the original feature representation onto a particular dimension and concatenating it into the enhanced correlation features to reduce potential information loss during enhancement. We conducted experiments on different projection dimensions, and the results are shown in Table 7. Different projection dimensions have little impact on the experimental results, and the best effect is achieved when the dimension is 64.



Fig. 5. Visualization of GradCAM obtained from different backbones of CLIP in CGM.

Differences in projection operations of EGM. In Eq. 13, we mentioned the operation of adding all features during the feature projection process. A feasible alternative is to concatenate all features and feed them into a dimensionally reduced convolutional layer. The features obtained in this way are of the same size as those obtained by direct addition, where we note this scheme as "concatenation". The comparison results of the two schemes are shown in Table 8. It can be seen that directly adding features can obtain better results than concatenation without the need for additional convolution operations.

Comparison with existing similar work. It is worth noting that the recent related work [23] also considers a similar problem setting. Different from [23], our method focuses more on exploring the performance capabilities of the foundation model in WS-FSS. To further demonstrate the difference between the two methods, we compare the differences between the two methods on PASCAL-5^{*i*}, and the results are shown in Table 9. It can be seen from the results that our method can help the model perform better in the WS-FSS scenario due to its strong generalization ability based on the foundation model and the robustness of the proposed method.

Parameter sensitivity. For our CGT, we have designed a self-distillation loss aid model to generate higher-quality correlation maps. We conducted sensitivity experiments on different loss balance parameters $\lambda_{distill}$, as shown in Fig. 6. Under different $\lambda_{distill}$, the mIoU variation of the model is relatively robust and reaches its optimal value at 0.5. However, without using self-distillation loss, *i.e.* $\lambda_{distill} = 0$, the model performance decreases by 1.1%, further proving the advantage of our proposed loss.



Fig. 6. Results with loss balance parameter $\lambda_{distill}$ on the 1-shot setting.

The difference between related work and weakly-supervised few-shot segmentation.

Related task settings	Difference
Few-shot semantic segmentation [7]	Dependence on accurate ground-truth mask.
Weakly-supervised semantic segmentation [60]	The model can only segment seen categories.
Weakly-supervised few-shot classification & segmentation [15]	The provided category information is whether the two images belong to the same category $(0/1 \text{ label})$ and does not provide specific category assistance for segmentation.

5.4. Discussions

Discussions about related settings. We demonstrate the differences between WS-FSS and related work settings in Table 10. Compared to WS-FSS, few-shot segmentation methods [49, 13, 8, 12, 29, 28, 10] rely more on precise ground-truth masks and learn to support and query the correlation of images based on this. Due to differences in application scenarios, weakly-supervised segmentation methods [59, 60, 27] cannot segment new classes that have not been seen before. Most relevant to WS-FSS, the weakly-supervised few-shot classification & segmentation method [15] not only performs weakly-supervised segmentation on query images containing unseen categories but also allows the model to output whether they belong to the same category as the support images. However, the category supervision information it provides is whether the query image is in the same category as the support image. It does not provide specific category information to assist the model in FSS.

Feature work. Compared to state-of-the-art methods on relatively simple datasets, our method has succeeded considerably. However, model performance can continue to improve when faced with more complex datasets (such as $COCO-20^i$). We will explore in the future how to better learn correlations between more complex images. On the other hand, our random division of background region in CGT also considers this problem. More complex correlations are learned by setting a more significant background number N, and the problem becomes part of parameter selection. In the future, we will have a more in-depth discussion on this issue to help the model learn more robust correlation knowledge without GT masks.

6. Conclusion

This paper proposes a framework to address the issue of requiring precise masks for existing FSS tasks, which address weakly-supervised few-shot segmentation tasks with only category information. To better mine the robust correlation between support queries, this paper proposes that CGT calculate similarity information from global and local perspectives. Then, from the perspective of category semantics, we designed CGM to help the model roughly locate targets using the pre-trained CLIP. In addition, the EGM module was designed to implicitly guide the model in filtering noise in correlation from the perspective of appearance embedding. Extensive experiments have shown that our CORENet has achieved state-of-the-art results in our weakly-supervised few-shot segmentation tasks.

References

- L. Fei-Fei, R. Fergus, P. Perona, One-shot learning of object categories, IEEE transactions on pattern analysis and machine intelligence 28 (4) (2006) 594–611.
- [2] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM computing surveys (csur) 53 (3) (2020) 1–34.
- [3] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, Advances in neural information processing systems 30 (2017).
- [4] Y. Xie, H. Wang, B. Yu, C. Zhang, Secure collaborative few-shot learning, Knowledge-Based Systems 203 (2020) 106157.

- [5] Y. Qin, W. Zhang, C. Zhao, Z. Wang, X. Zhu, J. Shi, G. Qi, Z. Lei, Prior-knowledge and attention based meta-learning for few-shot learning, Knowledge-Based Systems 213 (2021) 106609.
- [6] Y. Zhang, M. Gong, J. Li, K. Feng, M. Zhang, Autonomous perception and adaptive standardization for few-shot learning, Knowledge-Based Systems 277 (2023) 110746.
- [7] A. Shaban, S. Bansal, Z. Liu, I. Essa, B. Boots, One-shot learning for semantic segmentation, arXiv preprint arXiv:1709.03410 (2017).
- [8] K. Wang, J. H. Liew, Y. Zou, D. Zhou, J. Feng, Panet: Few-shot image semantic segmentation with prototype alignment, in: proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9197–9206.
- [9] C. Zhang, G. Lin, F. Liu, R. Yao, C. 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, 2019, pp. 5217–5226.
- [10] W. Liu, C. Zhang, G. Lin, F. Liu, Crnet: Cross-reference networks for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4165–4173.
- [11] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, J. Jia, Prior guided feature enrichment network for few-shot segmentation, IEEE transactions on pattern analysis and machine intelligence 44 (2) (2020) 1050–1065.
- [12] J. Min, D. Kang, M. Cho, Hypercorrelation squeeze for few-shot segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6941–6952.
- [13] Y. Yang, Q. Chen, Y. Feng, T. Huang, Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7131–7140.
- [14] Q. Li, B. Sun, B. Bhanu, Lite-fenet: Lightweight multi-scale feature enrichment network for few-shot segmentation, Knowledge-Based Systems 278 (2023) 110887.

- [15] D. Kang, P. Koniusz, M. Cho, N. Murray, Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19627–19638.
- [16] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: International conference on machine learning, PMLR, 2017, pp. 1126–1135.
- [17] J. Schmidhuber, Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook, Ph.D. thesis, Technische Universität München (1987).
- [18] A. Nichol, J. Schulman, Reptile: a scalable metalearning algorithm, arXiv preprint arXiv:1803.02999 2 (3) (2018) 4.
- [19] A. Rivolli, L. P. Garcia, C. Soares, J. Vanschoren, A. C. de Carvalho, Meta-features for meta-learning, Knowledge-Based Systems 240 (2022) 108101.
- [20] Y. Feng, J. Chen, J. Xie, T. Zhang, H. Lv, T. Pan, Meta-learning as a promising approach for few-shot cross-domain fault diagnosis: Algorithms, applications, and prospects, Knowledge-Based Systems 235 (2022) 107646.
- [21] P. H. T. Gama, H. N. Oliveira, J. Marcato, J. Dos Santos, Weakly supervised few-shot segmentation via meta-learning, IEEE Transactions on Multimedia (2022).
- [22] M. Zhang, Y. Zhou, B. Liu, J. Zhao, R. Yao, Z. Shao, H. Zhu, Weakly supervised few-shot semantic segmentation via pseudo mask enhancement and meta learning, IEEE Transactions on Multimedia (2022).
- [23] Y.-H. Lee, F.-E. Yang, Y.-C. F. Wang, A pixel-level meta-learner for weakly supervised few-shot semantic segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2170–2180.
- [24] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers,

in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.

- [25] D. Kang, M. Cho, Integrative few-shot learning for classification and segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9979–9990.
- [26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
- [27] L. Ru, Y. Zhan, B. Yu, B. Du, Learning affinity from attention: Endto-end weakly-supervised semantic segmentation with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16846–16855.
- [28] M. Siam, B. N. Oreshkin, M. Jagersand, Amp: Adaptive masked proxies for few-shot segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5249–5258.
- [29] L. Yang, W. Zhuo, L. Qi, Y. Shi, Y. Gao, Mining latent classes for few-shot segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 8721–8730.
- [30] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, J. Kim, Adaptive prototype learning and allocation for few-shot segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8334–8343.
- [31] A. Okazawa, Interclass prototype relation for few-shot segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 362–378.
- [32] H. Raza, M. Ravanbakhsh, T. Klein, M. Nabi, Weakly supervised one shot segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
- [33] O. Saha, Z. Cheng, S. Maji, Improving few-shot part segmentation using coarse supervision, in: European Conference on Computer Vision, Springer, 2022, pp. 283–299.

- [34] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, Advances in neural information processing systems 29 (2016).
- [35] F. Aurenhammer, Voronoi diagrams—a survey of a fundamental geometric data structure, ACM Computing Surveys (CSUR) 23 (3) (1991) 345–405.
- [36] J.-W. Zhang, Y. Sun, Y. Yang, W. Chen, Feature-proxy transformer for few-shot segmentation, Advances in Neural Information Processing Systems 35 (2022) 6575–6588.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
- [38] Y. Wu, K. He, Group normalization, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
- [39] Z. Dai, G. Lai, Y. Yang, Q. Le, Funnel-transformer: Filtering out sequential redundancy for efficient language processing, Advances in neural information processing systems 33 (2020) 4271–4282.
- [40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [41] B. Peng, Z. Tian, X. Wu, C. Wang, S. Liu, J. Su, J. Jia, Hierarchical dense correlation distillation for few-shot segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23641–23651.
- [42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.
- [43] H. Wang, L. Liu, W. Zhang, J. Zhang, Z. Gan, Y. Wang, C. Wang, H. Wang, Iterative few-shot semantic segmentation from image label text, arXiv preprint arXiv:2303.05646 (2023).

- [44] Z. Zhou, Y. Lei, B. Zhang, L. Liu, Y. Liu, Zegclip: Towards adapting clip for zero-shot semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11175–11185.
- [45] S. Jiao, Y. Wei, Y. Wang, Y. Zhao, H. Shi, Learning mask-aware clip representations for zero-shot segmentation, Advances in Neural Information Processing Systems 36 (2023) 35631–35653.
- [46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
- [47] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, M. Gelautz, Fast costvolume filtering for visual correspondence and beyond, IEEE transactions on pattern analysis and machine intelligence 35 (2) (2012) 504–511.
- [48] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8934– 8943.
- [49] S. Hong, S. Cho, J. Nam, S. Lin, S. Kim, Cost aggregation with 4d convolutional swin transformer for few-shot segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 108–126.
- [50] S. Amir, Y. Gandelsman, S. Bagon, T. Dekel, Deep vit features as dense visual descriptors, arXiv preprint arXiv:2112.05814 2 (3) (2021) 4.
- [51] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2010) 303–338.
- [52] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13, Springer, 2014, pp. 297–312.

- [53] K. Nguyen, S. Todorovic, Feature weighting and boosting for few-shot segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 622–631.
- [54] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
- [55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
- [56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (2015) 211–252.
- [57] X. Zhai, A. Kolesnikov, N. Houlsby, L. Beyer, Scaling vision transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12104–12113.
- [58] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- [59] D. Li, J.-B. Huang, Y. Li, S. Wang, M.-H. Yang, Weakly supervised object localization with progressive domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3512–3520.
- [60] H. Bilen, A. Vedaldi, Weakly supervised deep detection networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2846–2854.