Unlocking the Potential of Pre-trained Vision Transformers for Few-Shot Semantic Segmentation through Relationship Descriptors

Ziqin Zhou Hai-Ming Xu Yangyang Shu Lingqiao Liu*
The University of Adelaide, Australia
{ziqin.zhou,hai-ming.xu,yangyang.shu,lingqiao.liu}@adelaide.edu.au

Abstract

The recent advent of pre-trained vision transformers has unveiled a promising property: their inherent capability to group semantically related visual concepts. In this paper, we explore to harnesses this emergent feature to tackle few-shot semantic segmentation, a task focused on classifying pixels in a test image with a few example data. A critical hurdle in this endeavor is preventing overfitting to the limited classes seen during training the few-shot segmentation model. As our main discovery, we find that the concept of "relationship descriptors", initially conceived for enhancing the CLIP model for zero-shot semantic segmentation, offers a potential solution. We adapt and refine this concept to craft a relationship descriptor construction tailored for few-shot semantic segmentation, extending its application across multiple layers to enhance performance. Building upon this adaptation, we proposed a few-shot semantic segmentation framework that is not only easy to implement and train but also effectively scales with the number of support examples and categories. Through rigorous experimentation across various datasets, including PASCAL-5^i and COCO-20^i, we demonstrate a clear advantage of our method in diverse few-shot semantic segmentation scenarios, and a range of pre-trained vision transformer models. The findings clearly show that our method significantly outperforms current state-of-the-art techniques, highlighting the effectiveness of harnessing the emerging capabilities of vision transformers for few-shot semantic segmentation. We release the code at https://github.com/ZiqinZhou66/FewSegwithRD.git.

1. Introduction

The emergence of pre-trained vision transformers constitutes a pivotal advancement within the computer vision field. The pioneering model ViT [8] demonstrated competitive performance on various image classification tasks.

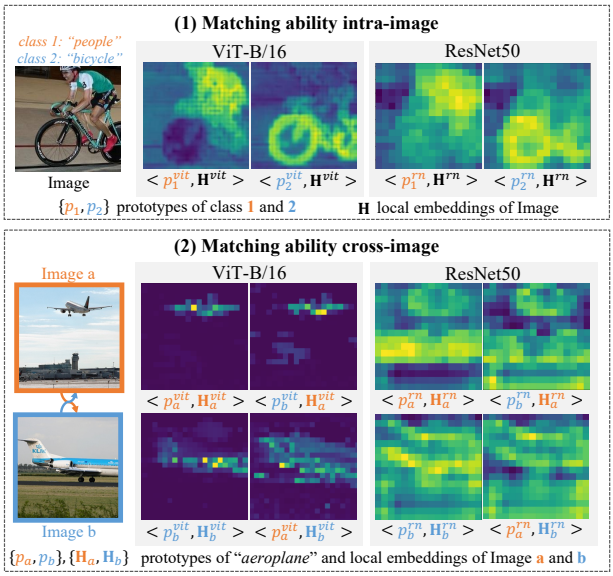

Figure 1. Visualized hidden (1) intra- and (2) cross- image matching capability of pre-trained ViT-B/16 and ResNet50 based on ImageNet. $< p, \mathbf{H} >$ denotes the cosine similarity between class-wise prototypes p and local feature $\mathbf{H}$ of Image. Prototypes p represent the average of embeddings belonging to the object-of-interest for each image, for example, the "people" and "bicycle" object of the Image in (1). It illuminates that the pre-trained transformer model has certain semantic grouping capabilities which may be helpful for few-shot semantic segmentation.

Subsequent studies [2, 17, 34] introduced different pre-training tasks to improve the perception ability further and widely applied on various downstream tasks. Unlike pre-trained ResNet models [15] which are based on conventional convolutional neural networks, vision transformers employ self-attention mechanisms to model global dependencies in images, affording them the capacity to hierarchically group semantically related visual concepts [2]. This means that they can learn to recognize and understand complex visual patterns by associating meaningful concepts within the image and enabling the potential for dense semantic prediction. As shown in Fig. 1, we visualize that without further learning, the pre-trained transformer already

*Corresponding author

has the capability of group patches with similar semantic meanings together, which could be a valuable property for segmentation tasks.

Few-shot semantic segmentation task aims to achieve precise segmentation of novel classes with a limited number of examples. A common paradigm involves utilizing a dedicated training dataset to acquire this capability. A pivotal obstacle in this learning process lies in ensuring that the few-shot segmentation capability extends effectively to classes that were not encountered in the training set. Utilizing pre-trained models, specifically pre-trained vision transformers, presents a promising avenue for mitigating this challenge. This is due to the fact that these pre-trained models are typically trained on large-scale datasets, encompassing a diverse array of visual concepts, even though their training is not specifically tailored for segmentation tasks. Consequently, one might anticipate that by employing a segmentation training dataset consisting of some base classes, it becomes possible for the few-shot segmentation capability to generalize effectively to classes that were not included in the training set.

Nonetheless, fine-tuning a pre-trained transformer directly with a decoder for segmentation often results in significant overfitting to the base classes. This occurs because the parameters introduced in the decoder, initially designed to produce optimal segmentation results for base classes, tend to be solely optimized for base class segmentation performance. This can lead to valuable knowledge about unseen classes learned from the pre-trained transformer being overlooked during the training process. To mitigate this issue, our approach draws inspiration from the recently introduced concept of a relationship descriptor (RD), originally developed for extending the CLIP model for zero-shot semantic segmentation. We adapt and refine the design of the relationship descriptor construction to tailor it specifically for few-shot semantic segmentation. Furthermore, we extend the construction of the RD to incorporate information from multiple layers. Building upon these adaptations, we propose a few-shot semantic segmentation framework that is not only easy to implement and train but also exhibits effective scalability with the number of support examples and categories.

We carried out extensive experiments on two well-known datasets, namely PASCAL-5^i and COCO-20^i, to evaluate the effectiveness of our proposed method. The results demonstrate that our method successfully tackles the issue of overfitting to base classes, and exhibits superior performance compared to state-of-the-art methods in various settings: generalized (GFSS) and binary (FSS) few-shot semantic segmentation.

2. Related Works

Few-Shot Learning (FSL) is a task that aims to enable models to quickly adapt to new classes with limited training examples and has been applied to various tasks including natural language understanding [1], image classification [5, 48], semantic segmentation [21, 47], object detection [53, 58] and so on. FSL can be broadly categorized into two paradigms: episodic meta-learning and representation learning. Meta-learning focuses on designing better meta-learners [10, 33, 36], metric distances [19, 40, 42], and architectures [11, 30, 37] through episodic training tasks. While representation-based methods [3, 43] simplify the procedure by facilitating more effective knowledge transfer from a base-class dataset to novel classes with limited support images.

Binary Few-shot Semantic Segmentation (FSS) was initially proposed in the work [47, 57] which designed a two-branch architecture to provide a classifier from support to query. Subsequent class-wise prototype methods [44, 47, 57, 60], influenced by [40], notably improved performance. Recent advancements include democratized graph attention mechanisms [46, 56], multi-scale correlation modules [52], and superpixel-guided clustering [21]. The latest developments by [20] and [18] focus on refining semantic masks and enhancing visual correspondence in few-shot segmentation.

Generalized Few-shot Semantic Segmentation (GFSS) introduced in [45], segments images of both base and novel classes without prior inference knowledge. FineTune [31] method emphasized test-time fine-tuning with a triplet loss for enhanced novel class performance. CCA [27] introduced a novel-class contrastive loss to address base-novel relationships, while DIaM [13] applied the InfoMax principle and knowledge distillation for classifier optimization. Additionally, POP [26] proposed an orthogonal constraint and an enrichment strategy to balance base and novel class performance during fine-tuning.

Pre-trained Vision Transformer has been highly valued in the research community due to its scalability [8] and exceptional feature representation capabilities [32]. Various downstream tasks have benefited from building upon pre-trained vision transformers for better performances, such as object detection [22, 41] and image segmentation [4, 23, 61]. For the few-shot semantic segmentation task, multiple works [28, 38, 59] have achieved great performance gain through using a pre-trained vision transformer backbone. In this work, we further investigate the semantic grouping capability of the pre-trained vision transformer and introduce it to the few-shot segmentation task for better novel class generalization.

3. Background

3.1. Few-shot Semantic Segmentation

Few-shot Semantic Segmentation task aims to segment the object-of-interest in a query image using only a few user-provided examples, where each example consists of an image paired with its corresponding object mask. The ability to perform few-shot segmentation is typically learned from a training set with well-annotated examples from a set of "base" classes, denoted as $\mathcal{C}^B$. The challenge of few-shot segmentation is how to ensure the model learned from "base" classes can be generalized to the novel classes during the test time. In the literature, there are two primary settings for few-shot segmentation:

Binary Few-shot Semantic Segmentation (FSS): This conventional setting evaluates the model exclusively on novel classes ($\mathcal{C}^N$, where $\mathcal{C}^N \cap \mathcal{C}^B = \emptyset$) and often focuses on object-background segmentation. The object mask only contains binary values indicating the object's presence.

Generalized Few-shot Semantic Segmentation (GFSS): The generalized setting extends beyond the standard framework to better mirror real-world scenarios, necessitating the recognition of both base and novel classes ($\mathcal{C}^B \cup \mathcal{C}^N$) as well as background within the same image. Also, this setting considers multiple object classes in the support examples, which requires the FSS method to scale well with the number of object categories.

3.2. Relationship Descriptor (RD) in ZegCLIP [61]

Recently, researchers have shown interest in adapting the image-level zero-shot classification ability of CLIP [35] model to the per-pixel prediction task, e.g., zero-shot semantic segmentation [7, 51]. One work ZegCLIP [61] proposed a novel relationship descriptor module (RD) to sufficiently utilize the text-image matching capability learned during the CLIP pretraining stage and effectively alleviate the base class overfitting problem. Specifically, the RD is the element-wise product between the image encoder embedding from the [CLS] token and the text embedding from the text encoder. Essentially, such a mechanism encodes the dimension-wise contribution to the image-text matching score calculated from a pre-trained CLIP model. It's important to note that the RD calculation takes place before applying any transformations to the text and image embeddings using newly introduced decoder parameters. This ensures that the matching capability of these embeddings is preserved, even for classes that are not part of the base classes. Therefore, using RDs can potentially alleviate the issue of overfitting to the based classes.

Our work is motivated by its effectiveness of leveraging the inherent matching capability of CLIP to alleviate the overfiting-to-base-class issue, despite our method not trying to leverage any text-image matching capability and needs a further adaption and extension of the idea of RD for few-shot segmentation.

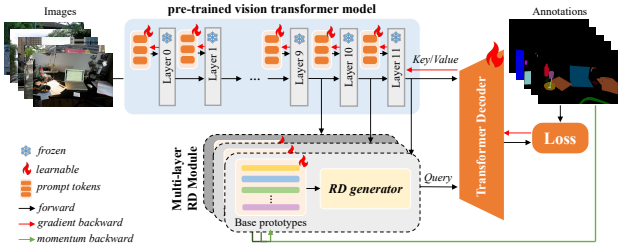

Figure 2. Overall training framework of our proposed few-shot semantic segmentation method.

3.3. Semantic Grouping Capability in Pre-trained Vision Transformers

Akin to the text-image matching capability of CLIP, the inherent capability we are interested in utilizing is the semantic grouping capability in a pre-trained vision transformer. This capability is well demonstrated in Fig. 1. We can see that even before any fine-tuning, the patch embedding extracted from a pew-trained visual transformer is similar to patches sharing the same semantic concept, i.e., the "person" riding bicycle or the "bicycle". In particular, we find this semantic grouping capability is quite prominent for vision transformers than CNNs. This motivates us to explore pre-trained vision transformers for few-shot segmentation tasks. The key research problem in this paper is thus how to extend the idea of RD to fully leverage this semantic grouping capability to improve generalization.

4. Our Method

4.1. Few-shot Segmentation via Prototype Embeddings and Relationship Descriptors

Without loss of generality, we can express a Few-shot Semantic Segmentation model in the form of $F(\mathbf{I}_q, \mathcal{S})$, where $\mathbf{I}_q$ represents the query image and $\mathcal{S}$ denotes the support set examples. To make the few-shot segmentation model scalable to the number of the classes and support set examples, recent work [45] advocates representing each support example as a prototype embedding vector, with one vector per object category. Specifically for one support image, we calculate the prototype embedding for class c as:

$$\mathbf{p}^c = \frac{1}{\sum_{i,j} \mathbf{M}^c[i,j]} \sum_{i,j} \mathbf{M}^c[i,j] \mathbf{H}_{i,j}, \tag{1}$$

where $\mathbf{H}_{i,j} \in \mathbb{R}^d$ is the patch embedding extracted from a pre-trained vision transformer and located at the (i,j) grid of the support image. $\mathbf{M}^c[i,j]$ is the value at the corresponding mask for object category c. It can be either "1" or "0", indicating foreground and background respectively.

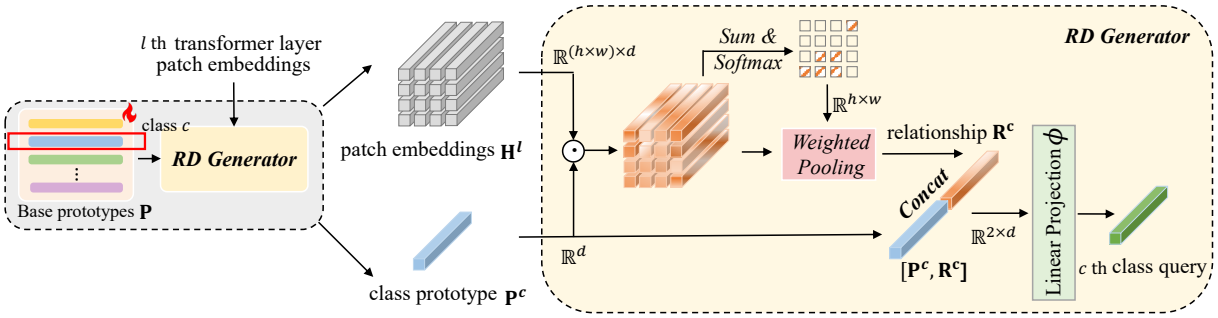

Figure 3. The detailed design of our Relationship Descriptor (RD) Generator module.

When there is more than one support example, we simply average the prototype across all examples. Once support examples are represented by $\mathbf{P} = [\mathbf{p}^1, \mathbf{p}^2, \cdots, \mathbf{p}^c]$, predicting semantic masks is performed by using a decoder $F(\mathbf{I}_q, \mathbf{P})$.

A straightforward idea to design $F(\mathbf{I}_q, \mathbf{P})$ is to use a matching function to compare $\mathbf{p}^c$ against patch embedding $\mathbf{H}_{i,j}$, producing how likely patch (i, j) is compatible with object category c and thus achieving semantic segmentation. However, this naive idea often leads to overfitting the base classes in the training set.

As mentioned in [61], the matching function inevitably introduces new parameters for achieving good matching quality[1]. However, those newly introduced parameters are optimized solely for the existing base classes, potentially disregarding valuable knowledge gained during the earlier transformer pretraining stage that could be beneficial for novel classes. In our experiments, we indeed observe this issue if we directly use prototypes to match patch embedding as shown in Tab. 4. Thus motivated by [61], we use an RD-augmented decoder to perform few-shot segmentation.

More specifically, the decoder consists of three transformer layers with a multi-head cross-attention model. The overall architecture is shown in Fig. 2. The transformer takes two types of inputs. **The first one** is the patch embedding from a pre-trained vision transformer, denoted as $\mathbf{H}$; and **the second one** is the concatenation of the prototypes $\mathbf{P}$ and the relationship descriptor $\mathbf{R}$ which is derived from the prototypes. The former will be used to produce the key and value vectors of the transformer and the latter will be used for producing the query vectors of the transformer. The final layer of attention, between the query for a given class and the patch embeddings, is then used to produce the segmentation mask for that particular class. The high-level calculation can be written as $\boldsymbol{Masks} = \mathrm{Trans}(\phi([\mathbf{P}, \mathbf{R}]), \varphi(\mathbf{H}))$, where $[\cdot]$ denotes concatenation and $\{\phi, \varphi\}$ represent the projection layers on class queries and patch features respectively. Please refer to [61] and Appendix for more details about the decoder design.

[1]Note that in ZegCLIP, the matching process is between text embedding and patch embedding, while in our case is matching between prototype and patch embedding.

4.2. Relationship Descriptors Design for Few-shot Segmentation

Unlike the ZegCLIP [61] model, which utilizes the inherent matching ability between image and text embeddings, the inherent matching capability we wish to leverage is on the matching between patch embeddings. This could further boil down to the matching between a class prototype and patch embeddings in a query image.

First, we define a relationship descriptor for each patch that can be done by performing an element-wise multiplication between a patch embedding and the class prototype. Considering the potentially large number of patch embeddings and the need to scale the method across multiple classes, it's more efficient to consolidate these relationship descriptors into a single descriptor per class. This consolidation can be effectively done by placing greater emphasis on descriptors whose corresponding patch embedding shows a higher matching score with the class prototype. The approach to achieve this involves a weighted summation method, which prioritizes descriptors based on their matching scores:

$$\mathbf{r}^c = \sum_{i,j} \alpha_{i,j} \mathbf{H}_{i,j} \odot \mathbf{p}^c, \quad \alpha_{i,j} = \frac{\exp(\mathbf{H}_{i,j}^\top \mathbf{p}^c / \tau)}{\sum_{m,n} \exp(\mathbf{H}_{m,n}^\top \mathbf{p}^c / \tau)}, \tag{2}$$

where τ is a temperature hyperparameter and is set to 0.1. $\mathbf{H}_{i,j} \odot \mathbf{p}^c \in \mathbb{R}^d$ with $\odot$ denoting the elementwise product. $\mathbf{H}_{i,j}^\top \mathbf{p}^c$ is the inner product (a scalar) between $\mathbf{H}_{i,j}$ and $\mathbf{p}^c$, indicating how $\mathbf{H}_{i,j}$ is related to class c. The final class-query procedure based on relationship descriptor generation is detailed presented in Fig. 3 and the collection of relationship descriptor can be formulated as $\mathbf{R} = [\mathbf{r}^1, \mathbf{r}^2, \cdots, \mathbf{r}^c]$.

4.3. Multi-layer Relationship Descriptors

Since semantic grouping capability exists in different layers of a vision transformer, it is possible to extend the relationship descriptor definitions to multiple layers. We use the l_ith layer patch embeddings of query and support images to perform the calculation in Eq. 2. Similar augmentation can be made to patch embeddings and prototypes. Specifically when multi-layer relationship descriptors are used,

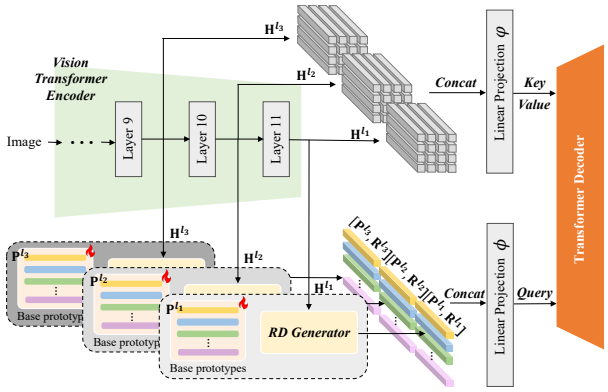

Figure 4. The framework of fusing multiple relationship descriptors from different transformer layers.

the enriched class query of the decoder transformers are $[\mathbf{P}^{l_1}, \mathbf{R}^{l_1}, \mathbf{P}^{l_2}, \mathbf{R}^{l_2}, \cdots]$ and the mask is calculated via:

$$\boldsymbol{Masks} = \text{Trans}\left(\phi([\mathbf{P}^{l_1}, \mathbf{R}^{l_1}, \mathbf{P}^{l_2}, \mathbf{R}^{l_2}, \cdots]), \varphi([\mathbf{H}^{l_1}, \mathbf{H}^{l_2}, \cdots])\right) \quad (3)$$

4.4. Training and Inference

Training algorithm: The aim of the training is to empower the network with the capability to conduct Few-Shot Semantic Segmentation using class prototypes derived from the support examples. In this work, we adopt a simple yet effective training procedure to learn the few-shot semantic segmentation capability from the training set. Specifically, our method maintains C prototypes for all objects that appear in the training set and updates the prototype dynamically using a momentum-style update equation:

$$\mathbf{p}^c \leftarrow (1-\eta) * \mathbf{p}_i^c + \eta * \mathbf{p}^c, \quad (4)$$

where $\mathbf{p}^c$ is the prototype embedding for the object in class c and $\mathbf{p}_i^c$ is the prototype embedding calculated from the i-th image. $\eta \in [0, 1)$ is the momentum coefficient to update parameters smoothly instead of learning linear classifiers as class-wise prototypes in the previous GFSS methods [45]. Following [12, 16], we initialize η with a relatively large value of 0.996 and increase η gradually with iteration up to 1.0.

At every training step, we utilize the up-to-date prototypes to compute RDs and generate queries for the decoder transformer. It's worth noting that because $\mathbf{p}^c$ is dynamically updated, it effectively introduces variations in the prototypes since $\mathbf{p}^c$ will differ at each iteration.

Training loss: Following previous works [55, 61], we apply the combination of the focal loss and dice loss to train the network, namely,

$$\mathcal{L}_{mask} = \lambda \cdot \mathcal{L}_{focal}(\mathbf{M}, \mathbf{M}^{gt}) + \beta \cdot \mathcal{L}_{dice}(\mathbf{M}, \mathbf{M}^{gt}), \quad (5)$$

where $\{\lambda, \beta\}$ is set to 20.0 and 1.0 to balance two loss items.

Our training method avoids the episodic training strategy of traditional few-shot learning and a pre-calculated prototype will be supplied at each training step, which is akin to the standard fully-supervised semantic segmentation at each training step. As shown in our experimental study, this method successfully adapts to both generalized and traditional binary few-shot segmentation settings.

Learnable parameters: Following ZegCLIP [61], we fix the parameters of the backbone vision transformer and introduce a few deep prompt tokens to make it adapt to the segmentation task. So the to-be-trained parameters are only those learnable deep prompts and the parameters in the transformer decoder.

Inference: In the context of Generalized Few-Shot Semantic Segmentation (GFSS), the class-wise prototypes are momentum-updated by using the training samples from the base classes. Meanwhile, we determine the prototypes for novel classes based on the examples in the support set. In the binary FSS setting that focuses on distinguishing foreground from background, the representation of the background can be achieved using prototypes from the designated background class, as well as prototypes from base classes. This approach is adopted because, during the inference stage in conventional FSS settings, base classes are treated as part of the background. More details are provided in the Appendix.

It's important to highlight that, in contrast to the techniques outlined in [13, 26, 27, 31], our approach has the capability to directly execute Generalized Few-Shot Semantic Segmentation using the support set examples, without requiring any additional fine-tuning on those examples. For a fair comparison, we also provide the results after a fast update on learnable parameters on the few annotated novel support set to further improve the performance.

5. Experiments

5.1. Datasets and Setup

To evaluate the effectiveness of our proposed method, we conducted extensive experiments on two public benchmark datasets for GFSS and FSS settings: **PASCAL-**5^i and **COCO-**20^i. The PASCAL-5^i dataset is created from PASCAL VOC 2012 [9] with additional annotations from SDS [14] and contains a total of 20 classes with an extra "background" category. COCO-20^i dataset is more difficult which is derived from MSCOCO [24] and includes 80 categories with "background".

For a fair comparison, we followed the evaluation protocol used in [13, 26] and divided the classes of each dataset into four non-overlapping folds. For each fold, we considered the selected classes as the novel part (5 classes for PASCAL-5^i and 20 classes for COCO-20^i), while the re-

Table 1. Comparison of our proposed method with previous works on the PASCAL-5^i dataset under the GFSS setting. mIoU(N) and mIoU(B) denote the mean mIoU of the novel(N) and base(B) classes among four folders, while F0-F3 shows the detailed mIoU results of each novel fold. Note that the rows with gray background represent the results obtained after test-time finetuning on novel classes.

Method	1-shot							5-shot						
	mIoU(N)	F0	F1	F2	F3	**mIoU(B)**	**hIoU**	**mIoU(N)**	F0	F1	F2	F3	**mIoU(B)**	**hIoU**
CANet[57]	2.4	-	-	-	-	8.7	3.8	1.5	-	-	-	-	9.1	2.6
PFENet [44]	2.7	-	-	-	-	8.3	4.0	1.9	-	-	-	-	8.8	3.1
SCL [54]	2.4	-	-	-	-	8.9	3.8	1.8	-	-	-	-	9.1	3.1
PANet [47]	11.3	-	-	-	-	31.9	16.6	15.3	-	-	-	-	33.0	20.9
CAPL+PANet [45]	15.0	-	-	-	-	63.1	24.2	19.7	-	-	-	-	63.8	30.1
CAPL+DeeplabV3 [45]	15.1	-	-	-	-	65.7	24.6	23.3	-	-	-	-	67.0	34.6
CAPL+PSPNet [45]	17.5	11.5	26.0	20.3	12.0	66.1	27.7	24.6	16.7	34.6	27.4	19.6	66.9	36.0
CAPL-ViT [45]	23.2	16.5	32.4	25.3	18.6	72.6	35.2	28.5	20.2	37.8	31.2	24.5	73.4	41.1
Ours-single	45.4	45.6	53.5	38.8	43.5	74.2	56.3	51.2	55.7	61.9	40.9	46.3	74.5	60.7
Ours-multiple	**51.0**	58.1	60.4	40.5	44.8	**76.3**	**61.1**	**55.3**	63.0	66.5	42.5	49.2	**76.5**	**64.2**
FineTune [31]	19.7	-	-	-	-	66.4	30.4	50.5	-	-	-	-	71.3	59.1
CCA [27]	22.6	18.0	34.1	22.8	15.5	68.4	34.0	32.1	27.6	46.0	30.1	24.7	70.5	44.1
DIaM [13]	35.1	29.4	46.7	27.1	37.3	70.9	47.0	55.3	53.7	63.6	54.0	50.2	70.9	62.1
POP [26]	35.5	-	-	-	-	73.9	48.0	55.9	-	-	-	-	75.0	64.1
Ours-single	48.9	49.3	57.7	41.3	47.1	76.4	59.6	52.1	55.8	59.3	42.9	50.5	76.6	62.0
Ours-multiple	**52.6**	58.6	60.4	42.5	48.9	**76.5**	**62.3**	**56.7**	64.0	65.9	44.3	52.4	**76.9**	**65.3**

Table 2. Qualitative results on COCO-20^i datasets under GFSS setting. Note that the rows with gray background represent the results obtained after test-time finetuning on novel classes.

Method	mIoU(N)	F0	F1	F2	F3	mIoU(B)	hIoU
1-shot							
CAPL [45]	7.6	5.3	9.2	6.9	9.1	44.4	13.0
CAPL-ViT [45]	9.7	6.5	10.8	11.2	10.2	46.1	16.0
Ours-single	17.2	13.4	18.5	22.0	14.7	46.5	25.1
Ours-multiple	**21.1**	17.1	21.3	25.2	20.9	49.5	**29.6**
FineTune [31]	9.2	-	-	-	-	43.6	15.2
CCA [27]	8.8	6.6	10.0	9.3	9.4	46.9	14.1
DIaM [13]	17.2	15.9	19.5	16.9	16.6	48.3	25.4
POP [26]	15.3	-	-	-	-	**54.7**	23.9
Ours-single	20.6	17.6	21.8	24.1	19.0	48.4	28.9
Ours-multiple	**23.2**	20.3	22.3	27.5	22.7	49.6	**31.5**
5-shot							
CAPL [45]	11.0	6.5	14.0	10.6	13.0	44.9	17.7
CAPL-ViT [45]	11.9	7.2	13.5	14.1	12.9	46.9	19.0
Ours-single	22.0	19.4	22.6	25.7	20.5	47.2	30.0
Ours-multiple	**27.0**	22.6	28.3	30.4	26.5	50.1	**35.1**
FineTune [31]	28.8	-	-	-	-	46.6	35.6
CCA [27]	12.7	9.2	15.3	12.1	14.1	47.1	20.0
DIaM [13]	28.7	24.9	33.9	27.2	29.0	48.4	36.0
POP [26]	30.0	-	-	-	-	**54.9**	38.8
Ours-single	25.1	25.3	25.4	26.6	23.1	49.6	33.3
Ours-multiple	**33.1**	32.1	36.9	34.2	29.3	50.2	**39.9**

maining classes (15 for PASCAL-5^i and 60 for COCO-20^i) were considered as base categories. In the training stage, we learn the segmentation ability on the base dataset, and then randomly selected either 1 or 5 samples as support sets for each novel class. In the testing stage, we used the original testing set which contains both base and novel classes for evaluation without any pair-wise input in the generalized few-shot segmentation task. We randomly selected different samples as the novel support set several times by controlling the seeds, as done in [45]. The averaged results of each fold are provided in this work.

5.2. Evaluation Metrics

Following previous works, we measure pixel-wise classification accuracy (pAcc) and the mean of class-wise intersection over union (mIoU) on both base and novel classes, denoted as $mIoU(B)$ and $mIoU(N)$ respectively. We also evaluate the harmonic mean of the IoU ($hIoU$) among base and novel classes which is formulated as:

$$hIoU = \frac{2 * mIoU(B) * mIoU(N)}{mIoU(B) + mIoU(N)}. \quad (6)$$

5.3. Implementation Details

Our proposed method is implemented based on the open-source toolbox MMSegmentation [6] with PyTorch 1.10.1. All experiments we provided are based on the pre-trained vision transformers with the ViT-B/16 architecture and conducted on 4 Tesla V100 GPUs. The batch size is set to 16 with 512x512 as the resolution of images. The total training iterations on base datasets are 10K for PASCAL-5^i and 40K for COCO-20^i. The number of prompt tokens is set to 10 and 50 for PASCAL-5^i and COCO-20^i according to [61]. During the testing stage, similar to prior works [13, 26], we perform quick adaptation by fine-tuning the learned parameters for 100 iterations on PASCAL-5^i and 400 iterations on COCO-20^i, respectively. The optimizer is set to AdamW with the default training schedule in the MMSeg toolbox. Similar to the architecture proposed in [55, 61], we employ three plain vision transformer layers serving as the lightweight decoder network.

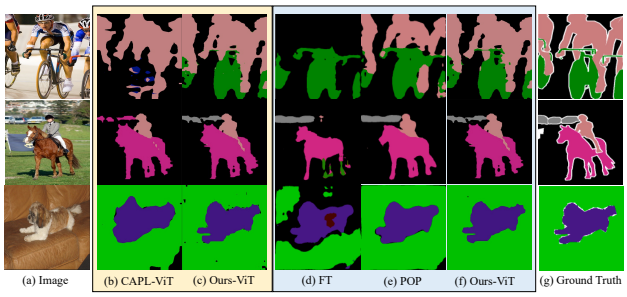

Figure 5. Visualization results on PASCAL-5^i compared with other methods on GFSS setting. Note that (b)-(c) are the methods without test-time tuning, while (d)-(f) are the methods with test-time tuning on novel support images.

5.4. Comparison with Existing GFSS Methods

We first compare our methods against existing ones in Tab. 1 and Tab. 2. Quantitative improvements achieved by our proposed are shown in Fig 5. We consider using single-layer RDs (our-single) and multi-layer RDs (our-multiple). In addition to the inference approaches mentioned previously, we also explore using the test-tuning strategy, that is, fine-tuning the model several iterations on the novel support examples, to achieve a fair comparison to methods [13, 26, 27, 31]. The results obtained are marked in gray background. As seen from Tab. 1 and Tab. 2, our method shows a clear advantage over existing methods when no test-tuning strategy is used. In particular, using multi-layer RDs leads to consistent improvement over single layer RDs, achieving state-of-the-art performance in GFSS. After test-tuning, our methods gains further improve and outperform the comparable methods.

Moreover, we reported the results by replacing the pre-trained ResNet backbone with ViT-B/16 in CAPL [45] and using the same decoder in our proposed method. The results (CAPL-ViT) demonstrate that even though the CAPL method with ViT achieved better segmentation performance on base classes, it still fails to generalize to novel categories. It validates that the competitive performance of our proposed method is not coming from the use of ViT, but the effectiveness of using relationship descriptor. Besides, the visualization in Fig. 6 indicates that enriched multiple RDs further improve the generalized recognition ability while achieving better segmentation boundaries.

5.5. Apply Our Pre-trained Model to FSS Setting

In accordance with previous studies, we evaluated the binary Few-Shot Semantic Segmentation (FSS) performance on novel pair-wise support-query sets. Although our model was not specifically designed for binary segmentation, it demonstrated remarkable results on PASCAL-5^i and COCO-20^i, as reported in Tab. 3 and qualitative performance is presented in Fig. 6.

Table 3. Comparison of our proposed method with the state-of-the-art FSS methods. Note that our method has not been trained on binary segmentation as well as test-time tuning on novel classes.

Method	Backbone	PASCAL-5^i		COCO-20^i	
		1-shot	5-shot	1-shot	5-shot
PANet [47]	RN-50	48.1	55.7	20.9	29.7
PFENet [44]	RN-50	60.1	61.4	32.4	37.4
BAM [20]	RN-50	67.8	70.9	46.2	51.2
MSANet [18]	RN-50	69.1	74.0	51.1	56.8
FECANet [25]	RN-50	69.3	74.9	50.9	58.3
ASGNet [21]	RN-101	59.3	63.9	34.5	42.5
SAGNN [49]	RN-101	62.1	62.8	37.2	42.7
HSNet [29]	RN-101	66.2	70.4	41.2	49.5
CAPL [45]	RN-101	63.6	68.9	42.8	50.4
DACM [50]	RN-101	69.1	73.3	43.0	49.2
CLIPSeg [28]	CLIP-ViT/B	52.3	-	33.2	-
CLIPSeg+ [28]	CLIP-ViT/B	59.3	-	33.2	-
PGMA-Net [39]	CLIP-ViT/B	74.1	74.6	-	-
PGMA-Net [39]	CLIP-RN50	74.1	75.2	54.3	57.1
PGMA-Net [39]	CLIP-RN101	77.6	78.6	59.4	**61.8**
FPTrans [59]	ViT-B/16	64.7	73.7	42.0	53.8
FPTrans [59]	DeiT-B/16	68.8	78.0	47.0	58.9
HSNet [38]	Swin-B	67.3	71.6	47.3	55.1
DCAMA [38]	Swin-B	69.3	74.9	50.9	58.3
Ours-single	ViT-B/16	77.7	78.0	57.1	59.2
Ours-multiple	ViT-B/16	**78.9**	**80.3**	**60.1**	61.2

5.6. Ablation study

A. How about the results without applying RD?

RD (relationship descriptor) was first proposed in ZegCLIP [61] to unleash the latent matching potential of the pre-trained transformer. By design, it only applies to the CLIP model that matches text and image modalities and it does not support the multiple layers design as [CLS] from other layers do not match across modalities in the pre-trained CLIP model. In fact, how to generalize the idea of RD to other domain or alternative construction of RD are not clear before this work. To experimentally verify this argument, we devised a baseline approach that exclusively employed class-wise prototypes without RDs as queries for the segment decoder. The results are shown in the first row of Table 4. Evidently, when RD is not utilized, a significant performance drop is observed. This outcome clearly shows the indispensable role played by RD in our method for mitigating overfitting to the training set.

B. Effect of our proposed design of RD

ZegCLIP [61] constructed relationship descriptors based on [cls] token, leveraging the inherent matching capability of [cls] tokens with CLIP text embedding. Our work is based on the vision transformer and we propose to leverage the inherent semantic grouping (matching) capability of patch tokens. Thus, we have a different design of RD, as shown in Eq. 2, than the one in ZegCLIP. Moreover, we propose a multi-layer strategy and fuse RD from different layers in Eq. 3. One alternative solution is to straightforwardly borrow the approach of RD construction from ZegCLIP by us-

Table 4. Using an alternative design for relationship descriptor (RD). [CLS] indicates a design that directly calculates RD via the [CLS] and the prototype embedding, as directly borrowed from ZegCLIP [61]. Our method achieves better performance especially on novel classes, highlighting the contribution of our customized RD.

RD	format	PASCAL -5^i						COCO -20^i					
		1-shot			5-shot			1-shot			5-shot		
		mIoU(N)	mIoU(B)	hIoU	mIoU(N)	mIoU(B)	hIoU	mIoU(N)	mIoU(B)	hIoU	mIoU(N)	mIoU(B)	hIoU
N/A	-	15.9	68.6	25.8	16.1	71.0	26.2	7.2	42.8	12.3	7.8	43.4	13.2
single	[cls]	41.9	73.8	53.5	48.8	74.2	58.9	13.7	45.4	21.0	16.8	46.5	24.7
	ours	45.4	74.2	56.3	51.2	74.5	60.7	17.2	46.5	25.1	22.0	47.2	30.0
		3.5(↑)	**(0.4↑)**	**(2.8↑)**	**(2.4↑)**	**(0.3↑)**	**(1.8↑)**	**(3.5↑)**	**(1.1↑)**	**(4.1↑)**	**(5.2↑)**	**(0.7↑)**	**(5.3↑)**
multiple	[cls]	48.6	74.6	58.9	53.9	75.3	62.8	17.5	49.9	25.9	20.6	50.2	29.2
	ours	51.0	76.3	61.1	55.3	76.5	64.2	21.1	49.5	29.6	27.0	50.1	35.1
		(2.4↑)	**(1.7↑)**	**(2.2↑)**	**(1.4↑)**	**(1.2↑)**	**(1.4↑)**	**(3.6↑)**	**(0.4↓)**	**(3.7↑)**	**(6.4↑)**	**0.1(↓)**	**(5.9↑)**

ing the element-wise product between the [CLS] embedding and prototype embedding. We investigate this variant (denoted as [cls]), and report the results in Table 4. Interestingly, we find this approach can also achieve quite good performance than the baseline without using RDs. The significantly lower performance compared to our RD design clearly demonstrates the unique contribution of our customized RD approach.

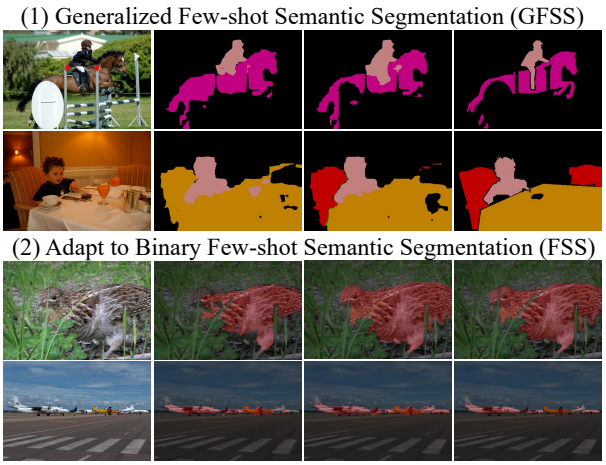

Figure 6. Quantitative results of (1) generalized and (2) binary few-shot semantic segmentation. (a) are the testing images, (b) and (c) are the results of our method with single and multiple relationship descriptors, (d) denotes ground truths. After enriched RDs, (c) shows better segmentation and recognition ability.

C. The choice of layers for multi-layer RDs

Our proposed multi-layer relationship descriptors (RDs) have exhibited a clear advantage over the use of single-layer RDs, as evident in Tab. 4 and the quantitative enhancements demonstrated in Fig. 6. To further investigate, we conducted comparisons among various choices of layers for constructing multi-layer RDs, with the experimental results presented in Fig. 7. As seen, the incorporation of the last layer is a main factor for achieving competitive performance. In addition to that, we identify the best combination from our investigation is utilizing the last three layers.

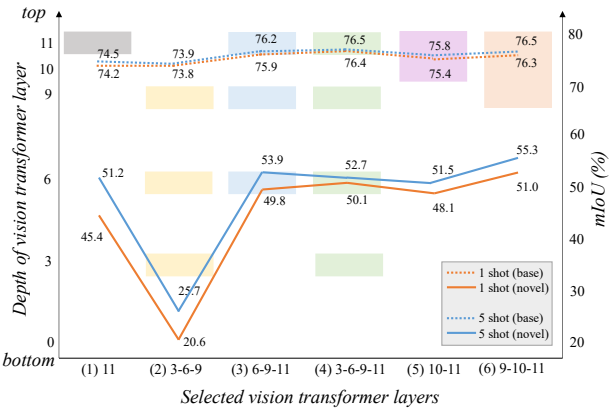

Figure 7. Quantitative results on PASCAL VOC 2012 of fusing relationship descriptors from different vision transformer layers.

6. Conclusion

This work presents a novel approach to few-shot semantic segmentation that leverages the hidden matching capability of a pre-trained vision transformer. The proposed method focuses on a more practical but challenging generalized few-shot segmentation setting and addresses the potential overfitting to base classes in the segmentation training set by using the relationship descriptor technique. In addition, our work can handle arbitrary input to conduct both generalized and traditional settings of the few-shot semantic segmentation tasks. We demonstrate the effectiveness and simplicity of our approach through extensive experimentation on various pre-trained vision transformers, achieving superior generalization performance and providing insights for choosing the appropriate transformer model. This work introduces a new paradigm for designing few-shot dense prediction models and is expected to contribute significantly to the research community.

Limitation Limited by the identification ability and resolution of ViT, similar objects may be misclassified and small objects may be ignored in complex scenes.

Acknowledgement This work was done in Adelaide Intelligence Research (AIR) Lab and Lingqiao Liu is supported by Centre of Augmented Reasoning.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1

[3] Weiyu Chen, Yencheng Liu, Zsolt Kira, Yuchiang Frank Wang, and Jiabin Huang. A closer look at few-shot classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 2

[4] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 2

[5] Philip Chikontwe, Soopil Kim, and Sang Hyun Park. Cad: Co-adapting discriminative features for improved few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14554–14563, 2022. 2

[6] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[7] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11583–11592, 2022. 3

[8] 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. 1, 2

[9] 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 (IJCV)*, 88:303–308, 2009. 5

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1126–1135. PMLR, 2017. 2

[11] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 2

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 33: 21271–21284, 2020. 5

[13] Sina Hajimiri, Malik Boudiaf, Ismail Ben Ayed, and Jose Dolz. A strong baseline for generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11269–11278, 2023. 2, 5, 6, 7

[14] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 991–998. IEEE, 2011. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 5

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 1

[18] Ehtesham Iqbal, Sirojbek Safarov, and Seongdeok Bang. Msanet: Multi-similarity and attention guidance for boosting few-shot segmentation. *arXiv preprint arXiv:2206.09667*, 2022. 2, 7

[19] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *Proceedings of the International Conference on Machine Learning Deep Learning Workshop (ICML workshop)*. Lille, 2015. 2

[20] 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 (CVPR)*, pages 8057–8067, 2022. 2, 7

[21] 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 (CVPR)*, pages 8334–8343, 2021. 2, 7

[22] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 2

[23] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 280–296. Springer, 2022. 2

[24] 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 *Proceedings of the IEEE conference on European Confer-*

ence on Computer Vision (ECCV), pages 740–755. Springer, 2014. 5

[25] Huafeng Liu, Pai Peng, Tao Chen, Qiong Wang, Yazhou Yao, and Xian-Sheng Hua. Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network. *IEEE Transactions on Multimedia (TMM)*, 2023. 7

[26] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11319–11328, 2023. 2, 5, 6, 7

[27] Weide Liu, Zhonghua Wu, Yang Zhao, Yuming Fang, Chuan-Sheng Foo, Jun Cheng, and Guosheng Lin. Harmonizing base and novel classes: A class-contrastive approach for generalized few-shot segmentation. *arXiv preprint arXiv:2303.13724*, 2023. 2, 5, 6, 7

[28] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7076–7086, 2022. 2, 7

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

[30] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2554–2563. PMLR, 2017. 2

[31] Josh Myers-Dean, Yinan Zhao, Brian Price, Scott Cohen, and Danna Gurari. Generalized few-shot semantic segmentation: All you need is fine-tuning. *arXiv preprint arXiv:2112.10982*, 2021. 2, 5, 6, 7

[32] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 34:23296–23308, 2021. 2

[33] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 3

[36] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2

[37] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016. 2

[38] 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 *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 151–168. Springer, 2022. 2, 7

[39] Chen Shuai, Meng Fanman, Zhang Runtong, Qiu Heqian, Li Hongliang, Wu Qingbo, and Xu Linfeng. Visual and textual prior guided mask assemble for few-shot segmentation and beyond. *arXiv preprint arXiv:2308.07539*, 2023. 7

[40] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 4077–4087, 2017. 2

[41] 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 (ICCV)*, pages 7262–7272, 2021. 2

[42] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1208, 2018. 2

[43] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 266–282. Springer, 2020. 2

[44] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44 (2):1050–1065, 2020. 2, 6, 7

[45] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11563–11572, 2022. 2, 3, 5, 6, 7

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

[47] 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 (ICCV)*, pages 9197–9206, 2019. 2, 6, 7

[48] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8012–8021, 2021. 2

[49] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmenta-

tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5475–5484, 2021. 7

[50] Zhitong Xiong, Haopeng Li, and Xiao Xiang Zhu. Doubly deformable aggregation of covariance matrices for few-shot segmentation. In *Proceedings of the IEEE conference on European Conference on Computer Vision (ECCV)*, pages 133–150. Springer, 2022. 7

[51] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*, 2021. 3

[52] Xianghui Yang, Bairun Wang, Kaige Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou. Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation. *arXiv preprint arXiv:2008.06226*, 2020. 2

[53] Ze Yang, Ya-Li Wang, Xian-Yu Chen, Jian-Zhuang Liu, and Yu Qiao. Context-transformer: Tackling object confusion for few-shot detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12653–12660, 2020. 2

[54] 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 (CVPR)*, pages 8312–8321, 2021. 6

[55] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, and Yifan Liu. Segvit: Semantic segmentation with plain vision transformers. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5, 6

[56] 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 (ICCV)*, pages 9587–9595, 2019. 2

[57] 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 (CVPR)*, pages 5217–5226, 2019. 2, 6

[58] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P Xing. Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2

[59] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *arXiv preprint arXiv:2210.06908*, 2022. 2, 7

[60] 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

[61] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11175–11185, 2023. 2, 3, 4, 5, 6, 7, 8