APSeg: Auto-Prompt Network for Cross-Domain Few-Shot Semantic Segmentation

Weizhao He[1†], Yang Zhang[1†*], Wei Zhuo[3*], Linlin Shen[1,2,3], Jiaqi Yang[3,4], Songhe Deng[1], Liang Sun[1]

[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2]Shenzhen Institute of Artificial Intelligence and Robotics for Society
[3]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[4]School of Computer Science, University of Nottingham, China

{heweizhao, dengsonghe, sunliang}2022@email.szu.edu.cn {yangzhang, weizhuo, llshen}@szu.edu.cn jiaqi.yang2@nottingham.edu.cn

Abstract

Few-shot semantic segmentation (FSS) endeavors to segment unseen classes with only a few labeled samples. Current FSS methods are commonly built on the assumption that their training and application scenarios share similar domains, and their performances degrade significantly while applied to a distinct domain. To this end, we propose to leverage the cutting-edge foundation model, the Segment Anything Model (SAM), for generalization enhancement. The SAM however performs unsatisfactorily on domains that are distinct from its training data, which primarily comprise natural scene images, and it does not support automatic segmentation of specific semantics due to its interactive prompting mechanism. In our work, we introduce APSeg, a novel auto-prompt network for cross-domain few-shot semantic segmentation (CD-FSS), which is designed to be auto-prompted for guiding cross-domain segmentation. Specifically, we propose a Dual Prototype Anchor Transformation (DPAT) module that fuses pseudo query prototypes extracted based on cycle-consistency with support prototypes, allowing features to be transformed into a more stable domain-agnostic space. Additionally, a Meta Prompt (MPG) module is introduced to automatically generate prompt embeddings, eliminating the need for manual visual prompts. We build an efficient model which can be applied directly to target domains without fine-tuning. Extensive experiments on four cross-domain datasets show that our model outperforms the state-of-the-art CD-FSS method by 5.24% and 3.10% in average accuracy on 1-shot and 5-shot settings, respectively.

[†]Equal Contribution: Weizhao He and Yang Zhang
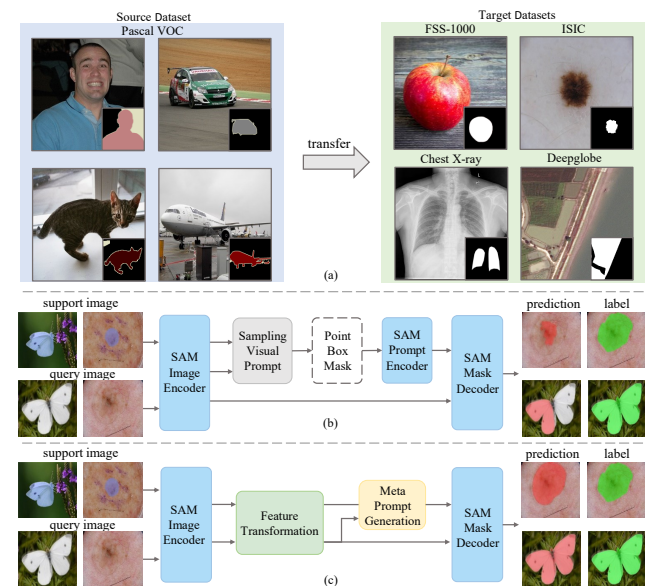[*]Corresponding Author: Yang Zhang and Wei Zhuo

1. Introduction

Figure 1. (a) In CD-FSS tasks, training (source) and testing (target) datasets come from different domains, and categories in the testing dataset are unseen during the training phase. (b) The framework of PerSAM [53], an existing one-shot segmentation method based on SAM. (C) The framework of our proposed APSeg.

Current deep neural networks [7, 8, 25, 45] depend heavily on extensive annotated data to attain satisfactory performance. Data annotation is a time-consuming task that requires significant human resources, especially for dense pixel-wise annotation for segmentation tasks [8, 20, 23]. The few-shot semantic segmentation (FSS) [33] is therefore introduced to close this gap, aiming to produce pixel-level predictions for a novel category with only a few of labeled samples. Although existing FSS methods [20, 26, 30, 38]

have achieved significant progress, they are commonly built on the assumption that their training and test images are from the same domain. When they are applied to a different domain, their performance decreases dramatically [21, 37, 42, 43]. The cross-domain generalizability is thus significant and necessary. FSS models typically require a large amount of base data for training. However, images for some tasks, such as cancer diagnosis and remote-sensing analysis, are scarce and challenging to obtain, making training a powerful model directly on their own data nearly impossible. In our work, we aim to transfer knowledge from easily accessible natural domains to data-scarce domains, as shown in Fig. 1(a). We believe that the transfer ability of general models trained on large-scale natural scene datasets will benefit domains that own only a handful of data.

To accomplish the cross-domain few-shot semantic segmentation (CD-FSS) task, PATNet [21] utilizes support prototypes for computing a transformation matrix, facilitating the conversion of domain-specific features into domain-agnostic ones. In addition, PATNet has a further fine-tuning process using the target domain data during the testing phase. Built on [21], RestNet [16] introduces a unified attention module to enhance query and support features prior to transformation. Residual connections are integrated to fuse features before and after the transformation, preserving important information from the original space. Furthermore, RestNet achieves better segmentation results by predicting twice. These works however are inefficient in application due to an additional finetuning process or double predictions. In addition, their backbone pretrained on ImageNet [32] limits their performance.

To this end, we attempt to take advantage of recent achievements on the foundation model, the Segment Anything Model (SAM) [19] to assist CD-FSS. With more than one billion masks under its training, SAM exhibits strong feature extraction and generalization abilities. However, recent studies [6, 9, 27, 39] have reported that applying SAM directly to new domains often yields subpar performance on zero-shot segmentation tasks, particularly when the data distribution significantly differs from the natural domain data used in SAM training. In addition, its interactive framework necessitates manual visual prompts, such as points or boxes, for precise segmentation, which restricts its capability for full automation. Furthermore, some recent works [6, 22, 48] demonstrate SAM is sensitive to manual visual prompts. Even slight deviations in provided prompts can remarkably affect segmentation accuracy. As shown in Fig. 1(b), PerSAM presents unsatisfactory performance due to its inability to extract high-quality visual prompts.

In our work, we propose an auto-prompt network for cross-domain few-shot semantic segmentation (APSeg), which builds a novel end-to-end framework that efficiently adapts SAM to CD-FSS tasks for accurate segmentation.

As shown in Fig. 1(c), the core of our framework is feature transformation and meta prompt generation. In particular, for feature transformation, we propose a Dual Prototype Anchor Transformation (DPAT) module to extract pseudo query prototypes based on cycle-consistency between support and query features. By fusing the pseudo query prototypes with the support prototypes, the computation of the transformation matrix can incorporate information from both support and query samples, which facilitates transforming input features into a more resilient domain-agnostic feature space. Combined with DPAT, the potential of SAM in cross-domain scenarios can be unleashed. In addition, for automatic prompt-embedding generation, we introduce a Meta Prompt Generator (MPG) module via a meta-learning procedure. Rather than relying on manual visual prompts such as point and box prompts, MPG leverages support features to guide the generation of meta prompt embeddings associated with target objects to substitute the output of the SAM's prompt encoder. With MPG, our method is robust and general for automatic segmentation. Our main contributions are summarized as follows:

- We propose a novel model that integrates a dual prototype anchor transformation (DPAT) module and a meta prompt generator (MPG) module for efficiently adapting SAM to CD-FSS tasks.
- The DPAT module is proposed for cross-domain feature transformation, which integrates support prototypes and pseudo query prototypes and transforms input features into a stable domain-agnostic space.
- The MPG module is introduced to generate prompt embeddings through meta-learning to establish a fully automatic framework for segmentation.
- Extensive experiments on four cross-domain datasets demonstrate that our model outperforms the state-of-the-art CD-FSS method by 5.24% and 3.10% in average accuracy on 1-shot and 5-shot settings, respectively. Especially, we attain 17.49% (1-shot) and 14.30% (5-shot) improvements on the Chest X-ray dataset.

In our method, the parameters of SAM's image encoder and mask decoder are frozen, and only a few parameters are trainable. Notably, our trained model can directly achieve promising results when applied to target domains without fine-tuning or multiple rounds of inference.

2. Related Work

2.1. Few-shot Segmentation

The goal of few-shot segmentation (FSS) is to segment new classes with a few annotated examples. Current FSS methods are commonly based on meta-learning, which can be largely grouped into two types: prototype-based methods [4, 41, 47, 54] and matching-based methods [20, 30, 44, 49]. Motivated by PrototypicalNet [36] for few-shot learning,
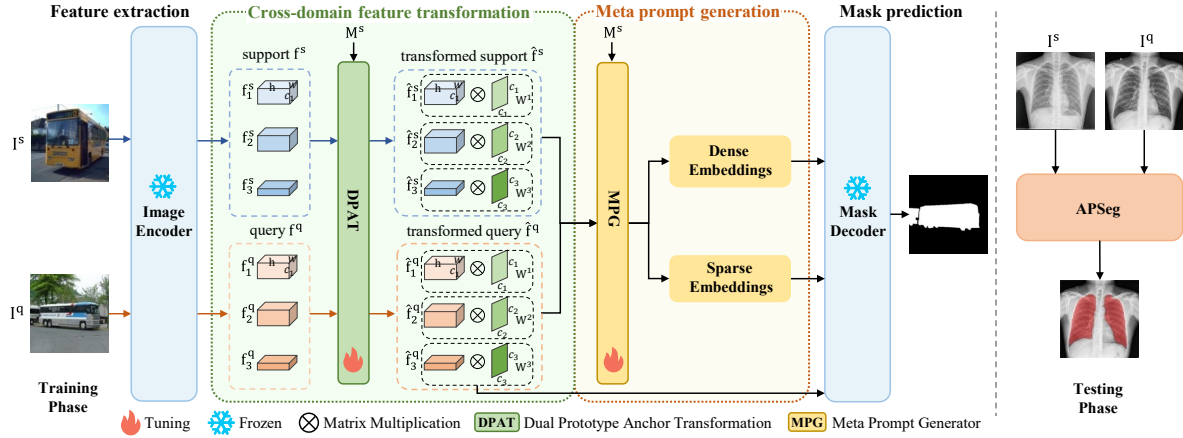
Figure 2. The overall architecture of our proposed APSeg in a 1-way 1-shot example. After obtaining the multi-layer features of support and query images, DPAT is employed to transform the domain-specific features into domain-agnostic ones by producing linear transformation matrices. Then the transformed features are passed into MPG to generate prompt embeddings. At last, the mask decoder takes the prompt embeddings and the transformed high-level query feature as input to make a prediction for the query image. In the testing phase, the trained model is directly applied to complete meta-testing in the target domain.

the prevalent FSS models utilize prototypes for specific-class representation. Recent works [23, 50] point out that a single prototype has a limitation to cover all regions of an object, especially for pixel-wise dense segmentation tasks. To alleviate this problem, methods of [23, 50] use expectation-maximization and cluster algorithms to generate multiple prototypes for different parts of the objects. Compared with prototype-based methods, matching-based ones [28, 38, 52] are designed to extract dense correspondences between query images and support annotations, harnessing pixel-level features to augment the support context with more intricate details. These methods however only focus on segmenting new categories from the same domain and fail to generalize unseen domains.

2.2. Cross-domain Few-shot Segmentation

In contrast to the traditional FSS setting, CD-FSS necessitates that models refrain from accessing target data during the training process. Furthermore, the data distribution and labeling space in the test phase differ from those in the training phase. This is a more realistic setting. PATNet [21] proposes a feature transformation module, which aims to convert domain-specific features into domain-agnostic features. In addition, the target domain data are desired to be utilized to fine-tune the model during the testing phase. Notably, PATNet outperforms current FSS methods on its established benchmark. RestNet [16] utilizes a lightweight attention module to enhance pre-transformation features and merge post-transformation features through residual connections to maintain the key information in the original domain. Meanwhile, a mask prediction strategy is introduced to mitigate the issue of overfitting to support samples in FSS and facilitates the model in a gradual acquisi-

tion of cross-domain segmentation capabilities. However, existing methods still utilize classical classification models [15, 35] as feature extractors with limited feature extraction capabilities compared to existing visual foundation models [19, 29, 31]. The performance of cross-domain segmentation is limited. Moreover, either additional training or multiple inferences is required when predicting masks, which makes the inference process complex and inefficient.

2.3. Segment Anything Model

The Segment Anything Model (SAM) [19], pretrained on massive amounts of labeled data, first introduced a foundation model for image segmentation. SAM relies on explicit points and bounding boxes at precise locations for accurate segmentation [10, 39]. Therefore, extensive manual guidance or a specialist detector is often required, leading to a complex multi-stage pipeline [40]. SAM cannot achieve automatic segmentation for specific semantics. To address this, PerSAM [53] proposes automatic sampling of visual prompts and some other methods suggest directly generating prompt embeddings [6, 48]. Inspired by these works, we propose an automatic prompting method in a meta-learning manner to adapt SAM to CD-FSS.

3. Method

3.1. Problem Definition

For CD-FSS, a source domain (X_s, Y_s) and a target domain (X_t, Y_t) exist, where the input distribution of the source domain and target domain are different and their label space has no intersection as well, i.e. $X_s \neq X_t$ and $Y_s \cap Y_t = \emptyset$. Here X denotes input distribution and Y denotes the label space. In our work, we train and test our
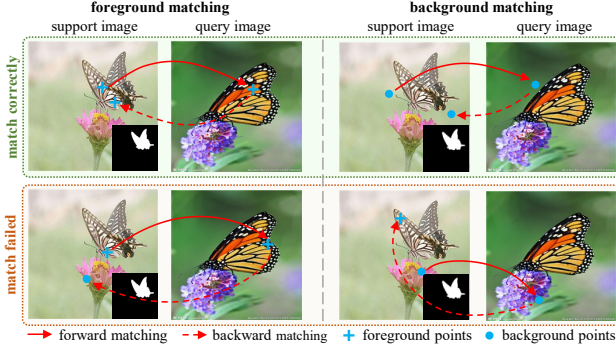
Figure 3. A visual example of a support-query pair to perform cycle-consistent selection (CCS).
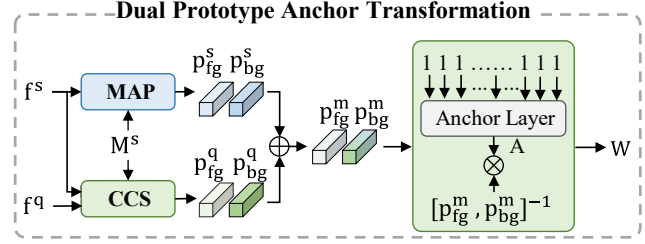

Figure 4. The specific implementation of the Dual Prototype Anchor Transformation module. Pseudo foreground and background prototypes of query are extracted through cycle-consistent selection (CCS). The extracted pseudo prototypes are then fused with support prototypes to calculate the transformation matrix $\mathbf{W}$ with an anchor layer.

model in a meta-learning episodic manner following [21], and our model is only trained on the source domain and has no access to the target data. Each episode data consists of a support set S and a query set Q with a specific category. The support set $S = (\mathrm{I}_i^s, \mathrm{M}_i^s)_{i=1}^K$ contains K image-mask pairs, where I_i^s denotes the i-th support image and M_i^s denotes the corresponding binary mask. Similarly, the query set is defined as $Q = (\mathrm{I}_i^q, \mathrm{M}_i^q)_{i=1}^K$. To train our model, support sets and images from query sets are used as model inputs to predict the query masks. To assess the trained model's performance, we test it on a support set and a query set from the target domain.

3.2. Method Overview

Our target is to train a general model on natural domains with rich annotations and transfer the knowledge to target domains with limited labeled data. As illustrated in Fig. 2, our proposed APSeg consists of two key modules: the Dual Prototype Anchor Transformation (DPAT) module and the Meta Prompt Generator (MPG). Specifically, given the support image I^s and query image I^q, the SAM image encoder is used to extract multi-level features from different layers. DPAT module is then employed to map support and query features to a new domain-agnostic space, facilitating the rapid adaptation of subsequent modules for previously unseen domains. Next, we introduce the MPG module whose task is to generate sparse and dense prompt embeddings through the meta-learning procedure for the SAM's mask decoder by utilizing the transformed features. At last, the generated prompt embeddings and the transformed high-level query features are passed into the mask decoder for target mask prediction.

3.3. Dual Prototype Anchor Transformation

To map support and query features into a new domain-agnostic space, we introduce a Dual Prototype Anchor Transformation (DPAT) module, as shown in Fig. 4. Previous method [21] solely relies on a set of support proto-

types and anchor layers to compute a transformation matrix. However, support prototypes cannot well represent complete information of a category due to intra-class variance. Therefore, we propose to enhance the support prototype set with query prototypes for better feature transformation. Specifically, our proposed DPAT consists of two procedures: dual prototype enhancement and cross-domain feature transformation. Inspired by [52], we propose cycle-consistent selection (CCS) to extract both the query foreground and background prototypes in the absence of query masks, which is used to enhance the support prototypes. Based on these enhanced prototypes which represent a category and its surroundings, an effective transformation matrix can be computed with a learnable domain-agnostic module. The transformation matrix is then applied to query features for cross-domain feature transformation.

Dual Prototype Enhancement Representative prototypes are significant for our cross-domain transformation. To this end, we build a cycling-examine procedure that reasons on both the query foregrounds and backgrounds to augment support prototypes. The process is shown in Fig. 3. We conduct forward matching to attain query features that have the highest similarity with the support foregrounds. We then use these identified forward-matched query features to re-locate corresponding support features reversely. If the located support features through reverse matching fall within support foregrounds, the identified query features will be averaged and used to derive the foreground prototypes. The background prototypes are obtained with the same process. Finally, support prototypes and query prototypes are fused by addition.

Specifically, given a support image I^s and a query image I^q, we initially employ a shared-weight image encoder to extract their multi-layer feature maps $\{f_l^s\}_{l=1}^3$ and $\{f_l^q\}_{l=1}^3$ respectively, where $f_l^s, f_l^q \in \mathbb{R}^{c_l \times h \times w}$. c_l, h, w are the channel dimension, height, and width of the feature map. DPAT takes support features $\{f_l^s\}_{l=1}^3$, query features $\{f_l^q\}_{l=1}^3$ and

support mask M^s as input. To simplify the notations, f^s and f^q are used to represent any one of $\{f_l^s\}_{l=1}^3$ and $\{f_l^q\}_{l=1}^3$. We then employ masked average pooling (MAP) on the support feature to obtain the foreground and background prototype, denoted as $\mathbf{p}_{fg}^s$ and $\mathbf{p}_{bg}^s$. By concatenating $\mathbf{p}_{fg}^s$ and $\mathbf{p}_{bg}^s$, we get the support prototype matrix $\mathbf{P}^s = \left[\mathbf{p}_{fg}^s, \mathbf{p}_{bg}^s\right]$. Considering that the query mask cannot be accessed during training, we extract pseudo query prototypes through CCS. First, the support foreground feature is obtained by multiplying the support mask M^s with the support feature f^s. Ne, the similarity between the support foreground feature and the query feature is calculated. For each element of the support foreground feature, we search for the element in the query feature map with the highest similarity score, and acquire the matched position set $i^{s \to q}$ from support to query as follows,

$$i^{s \to q} = \underset{i \in \{0,1,\ldots,h \times w-1\}}{\arg\max} \left(\text{sim}(f^s \odot M^s, f_i^q)\right) \qquad (1)$$

where $\text{sim}(\cdot, \cdot)$ is a cosine function. Based on $i^{s \to q}$, the matched query feature $f_{i^{s \to q}}^q$ can be extracted. Similarly, the matched positions $j^{s \leftarrow q}$ from query to support can obtained as below,

$$f_{i^{s \to q}}^q = \{f^q[i] : i \in i^{s \to q}\} \qquad (2)$$

$$j^{s \leftarrow q} = \underset{j \in \{0,1,\ldots,h \times w-1\}}{\arg\max} \left(\text{sim}(f_{i^{s \to q}}^q, f_j^s)\right) \qquad (3)$$

where $j^{s \leftarrow q}$ is the positions on support that have the most similar features with the corresponding reference query features $f_{i^{s \to q}}^q$. If the position in the matched position set $j^{s \leftarrow q}$ does not fall in the support mask M^s, we filter out the position from $i^{s \to q}$ and obtain the final matched position set $i'^{s \to q}$. According to the set $i'^{s \to q}$, the corresponding features are extracted from f^q and averaged to obtain the pseudo foreground query prototype $\mathbf{p}_{fg}^q$. The pseudo background query prototype $\mathbf{p}_{bg}^q$ can be obtained through a similar process, resulting in the query prototype matrix $\mathbf{P}^q = \left[\mathbf{p}_{fg}^q, \mathbf{p}_{bg}^q\right]$. Finally, a mixed prototype matrix can be obtained by $\mathbf{P}^m = \mathbf{P}^s + \mathbf{P}^q$.

Cross-domain Feature Transformation Features of the same class yield similar results when they are transformed in the same way. Support features and query features are transformed into a domain-agnostic space using the same transformation matrix $\mathbf{W}$ to avoid the detrimental impact caused by domain shift. Given the weight matrix of an anchor layer $\mathbf{A}$, the definition of a transformation matrix is as follows:

$$\mathbf{W}\mathbf{P}^m = \mathbf{A} \qquad (4)$$

where $\mathbf{P}^m = \left[\frac{\mathbf{p}_{fg}^m}{\|\mathbf{p}_{fg}^m\|}, \frac{\mathbf{p}_{bg}^m}{\|\mathbf{p}_{bg}^m\|}\right]$, $\mathbf{A} = \left[\frac{\mathbf{a}_{fg}}{\|\mathbf{a}_{fg}\|}, \frac{\mathbf{a}_{bg}}{\|\mathbf{a}_{bg}\|}\right]$ and $\mathbf{a}$ is the anchor vector which is independent of the input. Differ-
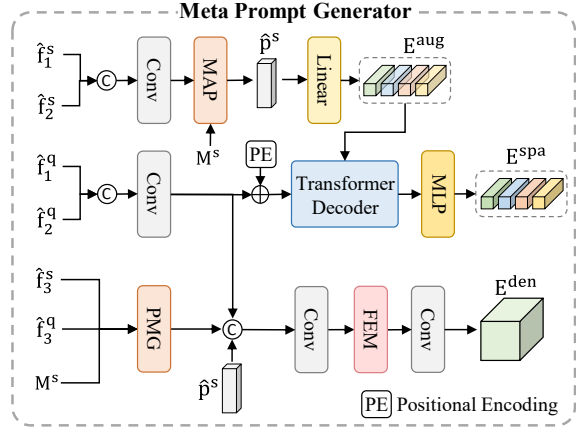

Figure 5. The specific implementation of the Dual Prototype Anchor Transformation module. PE indicates a positional encoding.

ent from previous work [21], we leverage the mixed prototype matrix that incorporates both support and query feature information during the computation process. Since the prototype $\mathbf{P}^m$ is a non-square matrix, the generalized inverse [1] of $\mathbf{P}^m$ is calculated with $\mathbf{P}^{m+} = \{\mathbf{P}^{m\mathbf{T}}\mathbf{P}^m\}^{-1}\mathbf{P}^{m\mathbf{T}}$. Therefore, the transformation matrix is calculated as $\mathbf{W} = \mathbf{A}\mathbf{P}^{m+}$. In our work, We have two different anchor layers for mid-level features $\{f_1^s, f_2^s, f_1^q, f_2^q\}$ and high-level features $\{f_3^s, f_3^q\}$. Finally, we can efficiently map support and query features to a stable, domain-agnostic space by multiplying them with $\mathbf{W}$.

Objects and things of even the same class can differ in shape and appearance. Due to limited samples of supports, it is intrinsically challenging to represent all the variance within objects and things of a class. Through the double-check procedure for both foreground and background regions, our proposed DPAT can effectively mitigate the challenges aroused by intra-class variances and generate a more stable transformation matrix for cross-domain feature transformation.

3.4. Meta Prompt Generator

To construct an end-to-end fully automated SAM-based segmentation framework for CD-FSS, we utilize the features of support-query pairs to directly generate prompt embeddings. In particular, a meta prompt generator (MPG) module is designed to obtain both sparse and dense embeddings simultaneously. Different from [4] that uses single support embeddings without alignments, our new pipeline extends to leverage multiple support embeddings and integrates feature alignment. Our design comprehensively takes account of intra-class variance and the multiple prompts coming along with the support embeddings enhance the segmentation. For clarity, we refer to the process of generating sparse embeddings and dense embeddings as sparse path and dense path respectively. In this way, our method

eliminates the need for external manual visual prompts, such as points or boxes.

Sparse path. In this process, the query features and several support embeddings augmented from the support prototypes are utilized to generate sparse embeddings through a transformer decoder [5], which then replace the original sparse embeddings in SAM. First, we concatenate the transformed mid-level query features $\hat{f}_1^s \in \mathbb{R}^{c_1 \times h \times w}$ and $\hat{f}_2^s \in \mathbb{R}^{c_2 \times h \times w}$ along the channel dimension, and then perform dimension reduction through a convolution layer to obtain $\hat{f}^s \in \mathbb{R}^{c_r \times h \times w}$. In the same way, we can also get $\hat{f}^q \in \mathbb{R}^{c_r \times h \times w}$,

$$\hat{f}^s = \mathcal{F}_{conv}(\hat{f}_1^s \oplus \hat{f}_2^s)$$
$$\hat{f}^q = \mathcal{F}_{conv}(\hat{f}_1^q \oplus \hat{f}_2^q) \tag{5}$$

where $\mathcal{F}_{conv}$ means performing a 1×1 convolution followed by a ReLU activation function and $\oplus$ denotes the concatenation operation in channel dimension.

Next, we take $\hat{f}^s$ and M^s as input and apply MAP to obtain the foreground class prototypes $\hat{\mathbf{p}}^s \in \mathbb{R}^{c_r}$. Then, a linear layer maps the $\hat{\mathbf{p}}^s$ to multiple augmented support embeddings $\mathbf{E}^{aug} \in \mathbb{R}^{k \times c_r}$,

$$\mathbf{E}^{aug} = \mathcal{F}_{linear}(\hat{\mathbf{P}}^s) \tag{6}$$

where k denotes the number of the augmented support embeddings. Here we generate several embeddings instead of only a single embedding for simulating multiple point prompts.

To supplement positional information, learnable position encodings are applied to $\mathbf{E}^{aug}$ and fixed sine-cosine positional encodings are applied to $\hat{f}^q$. We then input them into the transformer decoder and its output is further processed by a two-layer MLP to increase the channel dimensions, which yields $\hat{\mathbf{E}}^{aug} \in \mathbb{R}^{k \times c_o}$.

$$\hat{\mathbf{E}}^{aug} = \mathcal{F}_{mlp}(\mathcal{F}_{trans}(E^{aug}, \hat{f}^q)) \tag{7}$$

To align the generated sparse embeddings with those produced by SAM's prompt encoder, the sine function is employed to generate the final sparse embeddings $\mathbf{E}^{spa} \in \mathbb{R}^{k \times c_o}$ following [6].

$$\mathbf{E}^{spa} = \hat{\mathbf{E}}^{aug} + \mathcal{F}_{sine}(\hat{\mathbf{E}}^{aug}) \tag{8}$$

Dense path. In this process, the dense embeddings are modulated by query features and support prototypes. $\hat{f}_3^s$, $\hat{f}_3^q$ and M^s are first passed to prior mask generation (PMG) module [38] to generate a prior mask $M^{pr} \in \mathbb{R}^{1 \times h \times w}$. After concatenating $\hat{\mathbf{p}}^s$, $\hat{f}^q$ and M^{pr}, we perform a 1×1 convolution for dimension reduction to obtain $\hat{f}^{pr} \in \mathbb{R}^{c_r \times h \times w}$. The output is then passed into the feature enhancement (FEM) module [38] to get $\hat{f}^{fem}$. $\hat{f}^{fem}$ is further processed by a 1×1 convolution layer to increase the channel dimensions to obtain $\mathbf{E}^{den} \in \mathbb{R}^{c_o \times h \times w}$.

3.5. Training Loss

In the training of APSeg, we employ a Dice loss function, computed between the predicted mask $\hat{M}$ and the corresponding ground truth query mask M^q. The loss function, denoted as $\mathcal{L}$, is expressed as:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \text{DICE}\left(\mathcal{I}(\hat{M}), M^q\right) \tag{9}$$

Here, n represents the total number of training episodes in each batch, and DICE signifies the Dice loss function. The function $\mathcal{I}$ serves as an interpolation function, ensuring that $\hat{M}$ shares the same spatial size as M^q.

4. Experiment

Datasets. Following the previous approach [21], we use PASCAL VOC 2012 [13] with SBD [14] augmentation as training dataset and then evaluate the trained model on Chest X-ray [3, 17], ISIC [11], FSS-1000 [24] and Deepglobe [12] respectively.

Metric and Evaluation. We use the mean intersection-over-union (mIoU) as the evaluation metric, which is the same as the previous method. We take the mean-IoU of 5 runs [28] with different random seeds for each test. For all datasets except FSS-1000, each run has 1200 tasks. Every run of FSS-1000 has 2400 tasks.

Implementation Details. In our experiments, we employ the base version of the SAM and keep it frozen during training. To be consistent with the original input to SAM, we set spatial sizes of both support and query images to 1024×1024. For the SAM image encoder, we utilize feature maps derived from the mid-level features output by the 5th and 8th transformer blocks, in addition to the high-level features obtained from the final output of the image encoder. Concerning the DPAT module, we create two anchor layers dedicated to mid-level and high-level features, each configured with channel numbers set to 768 and 256, respectively. For the MPG module, the number of feature channels after dimension reduction is set to 64. Additionally, the number of sparse embeddings and dense embeddings channels output by MPG is set to 256. We implement the model in PyTorch and utilize the Adam [18] optimizer with a learning rate of $1e-3$.

4.1. Comparison with State-of-the-Arts

We compare our method against existing CNN-based and SAM-based approaches for CD-FSS. As shown in Tab. 1, the results demonstrate the superiority of the proposed method in this challenging task. Specifically, our approach exhibits improvements of 5.40% and 3.10% compared to the PATNet [21], under 1-shot and 5-shot settings, respectively. Moreover, APSeg surpasses the SAM-based
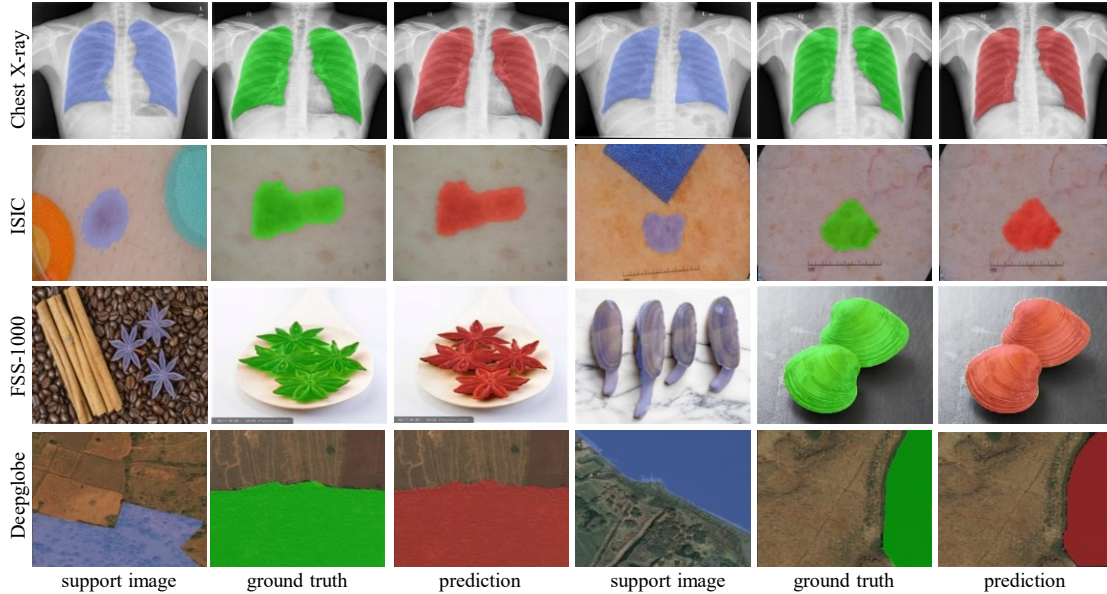
Figure 6. Qualitative results of APSeg in 1-way 1-shot segmentation on CD-FSS. Note that the model is trained on PASCAL VOC. Support labels are overlaid in blue. The prediction and ground truth of query images are in green and red respectively. Best view in color.

Methods	Backbone	Chest X-ray		ISIC		FSS-1000		Deepglobe		Average	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Few-shot Semantic Segmentation Methods											
AMP [34]	VGG-16	51.23	53.04	28.42	30.41	57.18	59.24	37.61	40.61	43.61	45.83
PGNet [50]	Res-50	33.95	27.96	21.86	21.25	62.42	62.74	10.73	12.36	32.24	31.08
PANet [41]	Res-50	57.75	69.31	25.29	33.99	69.15	71.68	36.55	**45.43**	47.19	55.10
CaNet [51]	Res-50	28.35	28.62	25.16	28.22	70.67	72.03	22.32	23.07	36.63	37.99
RPMMs [46]	Res-50	30.11	30.82	18.02	20.04	65.12	67.06	12.99	13.47	31.56	32.85
PFENet [38]	Res-50	27.22	27.57	23.50	23.83	70.87	70.52	16.88	18.01	34.62	34.98
RePRI [2]	Res-50	65.08	65.48	23.27	26.23	70.96	74.23	25.03	27.41	46.09	48.34
HSNet [28]	Res-50	51.88	54.36	31.20	35.10	77.53	80.99	29.65	35.08	47.57	51.38
PerSAM [53]	ViT-base	29.95	30.05	23.27	25.33	60.92	66.53	36.08	40.65	37.56	40.64
Cross-domain Few-shot Semantic Segmentation Methods											
PATNet [21]	Res-50	66.61	70.20	41.16	53.58	78.59	81.23	**37.89**	42.97	56.06	61.99
PATNet‡ [21]	ViT-base	76.43	-	44.25	-	72.03	-	22.37	-	53.77	-
RestNet [16]	Res-50	70.43	73.69	42.25	51.10	**81.53**	**84.89**	-	-	-	-
APSeg(ours)	ViT-base	**84.10**	**84.50**	**45.43**	**53.98**	79.71	81.90	35.94	39.98	**61.30**	**65.09**

Table 1. Comparison with previous FSS and CD-FSS methods under 1-way 1-shot and 1-way 5-shot settings. All the methods are trained on the source dataset and tested on the CD-FSS benchmark. The method marked with ‡ is implemented by ourselves.

method PerSAM [53] by 23.74% and 24.45%, affirming the effectiveness of our approach. Notably, our method showcases significantly superior performance compared to the current method when confronted with large domain gaps between the testing and training datasets. Specifically, compared with PATNet, our method achieves improvements of 17.49% and 14.30% on the chest X-ray dataset under the 1-shot and 5-shot settings, respectively. Similarly, on the ISIC dataset, improvements of 4.27% are observed for the 1-shot setting. Furthermore, our automatic prompting approach significantly surpasses the manual prompting strategy proposed by [53] in cross-domain performance. For a fair comparison, we also implement PATNet based on the SAM's image encoder. The model is trained with the same

input image size and test-time fine-tuning is not employed. The results demonstrate that APSeg still maintains a significant superiority. Qualitative results, illustrated in Fig. 6, validate that our proposed method attains substantial improvements in generalization performance in the presence of large domain gap while maintaining considerable accuracy with minor domain shift. More visualization results are provided in the supplementary materials.

4.2. Ablation study

Components analysis. We assess the effectiveness of our proposed DPAT module and MPG module by using the 1-shot mIoU averaged on 4 datasets. To establish a baseline model, we first remove the DPAT module and the sparse

path in the MPG module and then replace the final layer of the dense path with a 1×1 convolution layer for predicting segmentation masks. Tab. 2 illustrates the impact of each component on model performance. Overall, the incorporation of the two components suggested in this paper enhances the baseline by 18.44%. In the second row, MPG leverages the segmentation capabilities of SAM by autonomously generating semantic-aware prompt embeddings, eliminating the need for manual prompts and improving the baseline by 2.80%. Upon combining MPG and DPAT, DPAT unleashes the segmentation capabilities of SAM in cross-domain scenarios. It achieves this by transforming input features into a more stable domain-agnostic feature space, resulting in a significant performance improvement of 15.46% compared to the second row. More analysis and discussion about DPAT are shown in supplementary materials.

Method	Chest X-ray	ISIC	FSS-1000	Deepglobe	Average
Baseline	28.80	44.55	77.90	20.19	42.86
Baseline + MPG	32.62	36.67	79.77	34.31	45.84
Baseline + MPG + DPAT	84.10	45.43	79.71	35.94	61.30

Table 2. Ablation study on key components of APSeg on CD-FSS. Results are averaged over four datasets for 1-shot.

Sparse	Dense	1-shot mIoU
✓		41.52
	✓	44.09
✓	✓	45.43

Table 3. Ablation study on the choice of different types of prompt embeddings in MPG on ISIC.

Meta Prompt Generator. Tab. 3 shows the impact of the main components in the MPG, namely sparse embeddings and dense embeddings. We show the results for three combination scenarios: using only sparse embeddings, only dense embeddings, and both. Our observation reveals that better performance can be attained when combining sparse and dense embeddings. This emphasizes the significance of leveraging both types of embeddings to harness the segmentation capabilities of SAM in cross-domain scenarios.

The number of feature channels. After cross-domain feature transformation, we fuse the transformed mid-level features and perform dimension reduction, which can reduce the number of learnable parameters and avoid the trained model overfitting the training dataset. Observation in Tab. 4 shows that reducing the number of channels to 64 allows our method to achieve better CD-FSS performance with only a small number of additional learnable parameters.

The number of sparse embeddings. Our proposed MPG introduces an automatic generation mechanism for prompt embeddings, replacing SAM's original method of obtaining them through manual prompts fed into the prompt encoder. Employing more manual visual prompts has been

dim	1-shot mIoU	# Learnable Params
64	61.30	0.7 M
128	60.03	2.3 M
256	58.65	8.4 M

Table 4. Ablation study on different output feature channels. Results are averaged over four datasets for 1-shot.

shown to enhance SAM's interactive segmentation performance. Therefore, Tab. 5 shows the association between the number of sparse embeddings generated by MPG and the performance of cross-domain few-shot segmentation with APSeg. Notably, generating 4 sparse embeddings results in a 0.55% improvement in 1-shot scenarios. However, as indicated in the third row, continuing to generate more prompt embeddings may lead to a decline in performance.

num	1-shot mIoU
1	60.75
4	61.30
8	60.65

Table 5. Ablation study on the different number of sparse embeddings. Results are averaged over four datasets for 1-shot.

5. Conclusion

In this paper, we introduce APSeg, an auto-prompt method for guiding SAM to complete CD-FSS tasks. To achieve fully automatic segmentation based on SAM and release the segmentation capability of SAM in cross-domain scenarios, we propose the Meta Prompt Generator (MPG) module and Dual Prototype Anchor Transformation (DPAT) module to achieve this goal. By fusing the extracted pseudo query prototypes with support prototypes, DPAT enables domain-specific input features to be more stably converted into domain-agnostic features, significantly improving cross-domain generalization capabilities. In addition, MPG generates semantic-aware prompt embeddings with meta-learning, promoting the construction of a fully automatic CD-FSS framework based on SAM. Combining DPAT and MPG, extensive experimental results show that our APSeg achieves a new state-of-the-art in CD-FSS.

6. Acknowledgements

References

[1] Adi Ben-Israel and Thomas NE Greville. *Generalized inverses: theory and applications*. Springer Science & Business Media, 2003. 5

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

[3] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013. 6

[4] Leilei Cao, Yibo Guo, Ye Yuan, and Qiangguo Jin. Prototype as query for few shot semantic segmentation. *arXiv preprint arXiv:2211.14764*, 2022. 2, 5

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 6

[6] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 3, 6

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[9] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3367–3375, 2023. 2

[10] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023. 3

[11] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 6

[12] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018. 6

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

[14] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] Xinyang Huang, Chuang Zhu, and Wenkai Chen. Restnet: Boosting cross-domain few-shot segmentation with residual transformation network. *arXiv preprint arXiv:2308.13469*, 2023. 2, 3, 7

[17] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013. 6

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[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*, pages 8057–8067, 2022. 1, 2

[21] Shuo Lei, Xuchao Zhang, Jianfeng He, Fanglan Chen, Bowen Du, and Chang-Tien Lu. Cross-domain few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 73–90. Springer, 2022. 2, 3, 4, 5, 6, 7

[22] Chengyin Li, Prashant Khanduri, Yao Qiang, Rafi Ibn Sultan, Indrin Chetty, and Dongxiao Zhu. Auto-prompting sam for mobile friendly 3d medical image segmentation. *arXiv preprint arXiv:2308.14936*, 2023. 2

[23] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8334–8343, 2021. 1, 3

[24] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2869–2878, 2020. 6

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

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

[27] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 2

[28] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6941–6952, 2021. 3, 6, 7

[29] 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. 3

[30] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 1, 2

[31] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

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

[34] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5249–5258, 2019. 7

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

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

[37] Antonio Tavera, Fabio Cermelli, Carlo Masone, and Barbara Caputo. Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1626–1635, 2022. 2

[38] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):1050–1065, 2020. 1, 3, 6, 7

[39] Tassilo Wald, Saikat Roy, Gregor Koehler, Nico Disch, Maximilian Rouven Rokuss, Julius Holzschuh, David Zimmerer, and Klaus Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. In *Medical Imaging with Deep Learning, short paper track*, 2023. 2, 3

[40] An Wang, Mobarakol Islam, Mengya Xu, Yang Zhang, and Hongliang Ren. Sam meets robotic surgery: An empirical study on generalization, robustness and adaptation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–244. Springer, 2023. 3

[41] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 2, 7

[42] Wenjian Wang, Lijuan Duan, Yuxi Wang, Qing En, Junsong Fan, and Zhaoxiang Zhang. Remember the difference: Cross-domain few-shot semantic segmentation via meta-memory transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7065–7074, 2022. 2

[43] Yixin Wang, Zhe Xu, Jiang Tian, Jie Luo, Zhongchao Shi, Yang Zhang, Jianping Fan, and Zhiqiang He. Cross-domain few-shot learning for rare-disease skin lesion segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1086–1090. IEEE, 2022. 2

[44] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 517–526, 2021. 2

[45] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[46] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 763–778. Springer, 2020. 7

[47] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8721–8730, 2021. 2

[48] Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgicalsam: Efficient class promptable surgical instrument segmentation. *arXiv preprint arXiv:2308.08746*, 2023. 2, 3

[49] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8312–8321, 2021. 2

[50] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9587–9595, 2019. 3, 7

[51] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5217–5226, 2019. 7

[52] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems*, 34:21984–21996, 2021. 3, 4

[53] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 1, 3, 7

[54] 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