Visual Prompting for Generalized Few-shot Segmentation:
A Multi-scale Approach

Mir Rayat Imtiaz Hossain[1,2]   Mennatullah Siam[3,1]   Leonid Sigal[1,2,4]   James J. Little[1]

[1]University of British Columbia   [2]Vector Institute for AI   [3] Ontario Tech University

[4]Canada CIFAR AI Chair

Abstract

The emergence of attention-based transformer models has led to their extensive use in various tasks, due to their superior generalization and transfer properties. Recent research has demonstrated that such models, when prompted appropriately, are excellent for few-shot inference. However, such techniques are under-explored for dense prediction tasks like semantic segmentation. In this work, we examine the effectiveness of prompting a transformer-decoder with learned visual prompts for the generalized few-shot segmentation (GFSS) task. Our goal is to achieve strong performance not only on novel categories with limited examples, but also to retain performance on base categories. We propose an approach to learn visual prompts with limited examples. These learned visual prompts are used to prompt a multiscale transformer decoder to facilitate accurate dense predictions. Additionally, we introduce a unidirectional causal attention mechanism between the novel prompts, learned with limited examples, and the base prompts, learned with abundant data. This mechanism enriches the novel prompts without deteriorating the base class performance. Overall, this form of prompting helps us achieve state-of-the-art performance for GFSS on two different benchmark datasets: COCO-20^i and Pascal-5^i, without the need for test-time optimization (or transduction). Furthermore, test-time optimization leveraging unlabelled test data can be used to improve the prompts, which we refer to as transductive prompt tuning.

1. Introduction

The emergence of foundational models, trained on broad datasets, in natural language processing (e.g., GPT-3 [5]) and in vision-language (e.g., CLIP [31]), exhibits powerful generalization and strong performance on multiple downstream tasks. These models have been adapted through different prompting techniques [6, 43] for them to be used in the few-shot scenario.

However, despite successes, including in localization [6, 43, 50], prompting, which allows for few-shot demon-
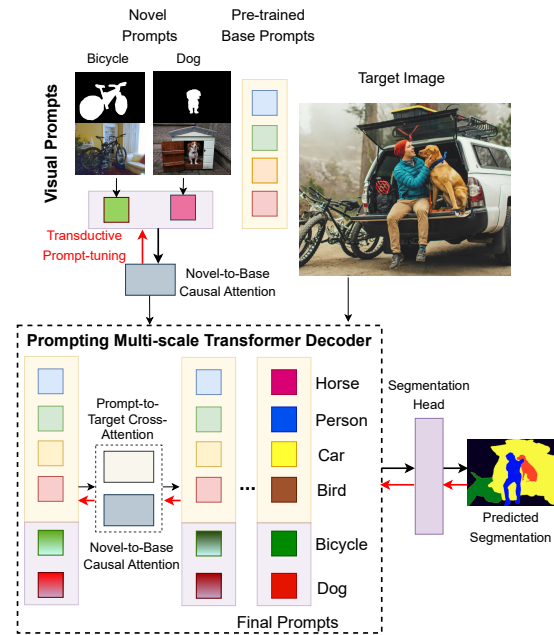

Figure 1. **Overview of the Proposed Approach.** Prompting multi-scale transformer decoders for generalized few-shot segmentation. Our approach is a simple approach that allows test-time transductive prompt tuning (see red arrows).

strations in dense prediction, and specifically in semantic segmentation tasks, is relatively under-explored.

Few-shot semantic segmentation [24, 40, 45, 46, 48, 49] aims to segment *novel* (unseen) classes with few labelled training examples. Most state-of-the-art methods rely on meta-learning, to leverage abundant training data as a form of "data augmentation" to construct many tasks that resemble test-time few-shot inference [26, 45]. In general, few-shot learning approaches can be classified as either *inductive* or *transductive*. Inductive approaches mainly rely on the training data, while transductive ones [4], in addition, exploit unlabelled test data in an unsupervised manner to improve the performance *e.g.*, by leveraging entropy priors over predictive class distributions. However, both cases largely focus on the performance of the *novel* categories; which isn't particularly realistic. A recently proposed *gen-*

eralized few-shot segmentation (GFSS) setting defines a more realistic scenario where the goal is to perform well on *all* classes, novel and base [36]. This is considerably more challenging.

Prompting has been proven to be effective for learning from few demonstrations, as seen in LLMs [5]. It was also explored in visual prompt tuning [16] to efficiently fine-tune vision transformers for new tasks. We posit that prompting within a transformer-based [38] architecture would similarly offer an effective and flexible mechanism for GFSS. This involves learning prompts that can be utilized to perform cross-attention with input image to make predictions. While learning prompts for base classes, with abundant data, is relatively straightforward, it becomes considerably more challenging for novel classes with few examples. Specifically, one must ensure that novel prompts learned from few samples are sufficiently distinct from the base prompts to avoid novel-base class misclassification.

To address the aforementioned challenges, we develop a simple yet highly effective visual prompting of transformer decoders for dense prediction at multiple scale that relies on novel-to-base causal[1] attention without meta-training. We look upon queries in DETR-style architectures as a form of visual prompts and devise a mechanism to initialize and learn novel prompts. Then the novel-to-base causal attention allows base prompts to affect novel prompt representations, but not the other way around. This, intuitively, allows novel prompts to be repulsed and/or attracted towards their base counterparts. This attention is shared across scales (layers of the transformer) which, as we show, leads to more robust learning and improved performance. The multi-scale refinement of prompts helps interacting and reasoning between image features at multiple scales which in turn helps in better dense prediction. Finally, we extended this architecture in a transductive setting, where both novel and base prompts can be fine-tuned at test-time with respect to unsupervised objective to further improve performance. The proposed architecture and methodology is shown in Fig. 1.

In summary, our contributions are:

- Designing a multi-scale visual prompting transformer decoder architecture for GFSS, featuring learnable prompts that allow the creation of new prompts for novel classes, initialized through masked average pooling of support images (with their masks).
- Within this architecture, proposing and learning a multi-scale (shared) novel-to-base cross-attention mechanism between the novel and base prompts.
- Proposing a novel transductive prompt tuning, which allows the visual prompts to be tuned on test (unlabeled) images, hence the term *transductive*.

---

[1]The term 'causal' [28] indicates its unidirectional nature.

2. Related work

Few-shot Learning. Few-shot learning, a longstanding concept in computer vision, serves as a paradigm to evaluate the data efficiency and adaptability of learning algorithms. It focuses on making predictions for *novel* classes with a few labeled examples, often by leveraging information from disjoint base classes for which there is abundant data. Few-shot learning has been extensively studied for image classification [12, 32–34, 39]. Most approaches can be categorized as either *transfer* or *meta*-learning-based. Transfer learning methods [11] typically pre-train the model on base classes with abundant data and then fine-tune this model for novel classes with few samples. Meta-learning [12, 32] approaches, on the other hand, simulate the inference stage during meta-training by sampling tasks of support and target sets. Meta-learning approaches can be sub-categorized into *gradient-based* and *metric-learning* based. Gradient-based approaches [12, 32] learn the commonalities among different tasks. Metric-based (or prototype-based) methods [34, 39] learn image embeddings and compare the distances between them. Despite its popularity and strong performance, meta-learning is found to be sensitive to the meta-training setup [7] and worse performance with domain shift [9] is observed.

Generalized few-shot learning, introduced in [14] as a more realistic variant, allows query images from either base or novel categories. Recently, the transductive setting has been proven to be effective for few-shot learning [3, 4, 13, 29, 52]. Transductive approaches exploit information of unlabelled test images during inference to improve classification [11, 22, 27, 30]. These approaches build on classic semi-supervised learning approaches, such as graph-based label propagation [23] or entropy minimization [11]. Recent variants of few-shot learning problems have also been established in other more granular domains, *e.g.*, detection [17, 41, 42, 44] and segmentation.

Few-shot segmentation. Less common than classification, few-shot segmentation is emerging as the dense counterpart to the aforementioned approaches. Few-shot segmentation approaches can be categorized to *prototype-based* and *dense comparison-based* methods. Prototype-based methods mostly relied on extracted prototypes from the support set to guide the segmentation of the target image, either as a single prototype [40, 49], part-aware [24] or a prototype mixture [45]. Dense comparison methods rely on dense comparisons between the few template images and the target image [46, 48] with extension to multiscale [26]. A recent approach proposed generalized few-shot segmentation (GFSS) task where the unlabelled target images are used to improve the segmentation of novel classes while maintaining the performance on base classes through knowledge distillation [13]. A concurrent work [21] on GFSS learns a set

of orthogonal prototypes for base and novel categories, yet their GFSS setup is easier as they sample in their inference balanced base class images along with the novel ones, unlike the setup we follow [13]. Recent work [50] proposed a prompting strategy within a meta-learning framework using few examples. This approach is designed for 1-way segmentation only and performs prompting at a single scale unlike our proposed approach, a non-meta learning scheme, which performs multi-scale prompting and is designed for k-way GFSS.

Prompting. An emerging direction in learning with limited labelled data is to utilize foundation models [2] that have strong generalization capabilities. These models, pre-trained on large-scale data, demonstrate strong generalization in few-shot settings [1]. Prompting is often used to adapt these foundation models to few-shot tasks, where it formulates the inference problem similarly to the training. Examples include auto-regressive modelling [1] where inference involves generating the next tokens after receiving the correct prompt. For few-shot learning, prompts include few-shot demonstrations followed by the query. Prompting can employ discrete prompts [1] or learning continuous prompts [20] through back-propagation. In computer vision, both vision-language prompting [1] and visual prompting [16] of transformer-based models have shown success in various tasks, particularly in few-shot scenarios. However, effectively prompting models designed for localization and segmentation remains an open question.

Concurrent Work. A recent method [6] proposed prompting of transformer decoders for few-shot object detection. Although conceptually related, their approach is designed for a different task (detection) compared to ours (semantic segmentation), trained within a meta-learning framework unlike ours. Further, [6] lacks novel-to-base attention and multi-scale prompt refinement which are at the core of our architectural design. A recent foundation model for semantic segmentation, SAM [18], exhibits strong performance when prompted by point, bounding box, or text. However, it is not specifically designed for few-shot inference, potentially requiring fine-tuning with the support set and other non-trivial changes. More importantly, the use of extensive data to train the model makes it unsuitable for few-shot tasks since it has likely encountered many target classes during training. In addition to aforementioned differences, including multi-scale prompting and novel-to-base attention, our method is the first to explore transductive prompt-tuning where the visual prompts can be further optimized with the unlabelled target images.

3. Problem Formulation

Classic Few-shot Segmentation. In the classic few-shot segmentation scenario, the categories are divided into two sets of disjoint sets: **base** and **novel**. The problem is framed as an n-**way** k-**shot** segmentation task, where n is the number of novel classes and k is the number of examples available for each class. A typical solution involves *meta-learning* with two-stage training. In the meta-training stage, a support set is created with n randomly selected base classes, and k examples from each class are used to train the model. In the meta-testing stage, the model is evaluated on tasks with k examples from n novel classes. An alternative approach is non-meta learning, where a model is initially trained on abundant data for base classes, then frozen and fine-tuned for novel classes using few examples. The focus is on segmenting novel classes in test images without considering base class segmentation.

Generalized Few-shot Segmentation (GFSS). *Generalized* few-shot segmentation, proposed by [36], extends the classic few-shot setting. In this new setup the model must preserve the segmentation performance on base classes in addition to generalizing to novel classes. While during fine-tuning the support set (typically) only contains k-shot examples for each novel category, the model is still expected to make predictions of both novel and base classes on the test images. We follow the recent more practical setup of generalized few-shot segmentation [13] within a non-meta learning framework, details provided in Section 5.1.

4. Method

We perform prompting using a set of learnable visual prompts, where, intuitively[2], each prompt corresponds to a class embedding. These learned visual prompts cross-attend with image features at multiple scales. Initially, we pre-train a base model with abundant examples of base classes, learning visual prompts for them. We then fix these base prompts and fine-tune new visual prompts for novel classes with a few examples. However, learning prompts from limited data is challenging because it leads to confusion between base and novel categories. To improve novel class generalization, we introduce a uni-directional causal attention mechanism between novel and base prompts. Intuitively, this allows better contextualization and repulsion / attraction between base and novel prompts. Section 5 demonstrates the empirical effectiveness of this mechanism.

Figure 2 shows our proposed architecture. It centers around learnable embeddings called *visual prompts*, which interact with image features at multiple scales. Our design is inspired by Mask2former [10] with a couple of key differences: (i) the number of visual prompts aligns with the number of classes, each representing a class as a learned embedding; (ii) we adopt per-pixel classification, instead of mask classification, thereby eliminating the query classification component of [10] and defining custom segmentation heads capable of making per-pixel predictions.

---

[2]This may not be precisely true in practice, but is a reasonable abstraction for intuitive understanding.
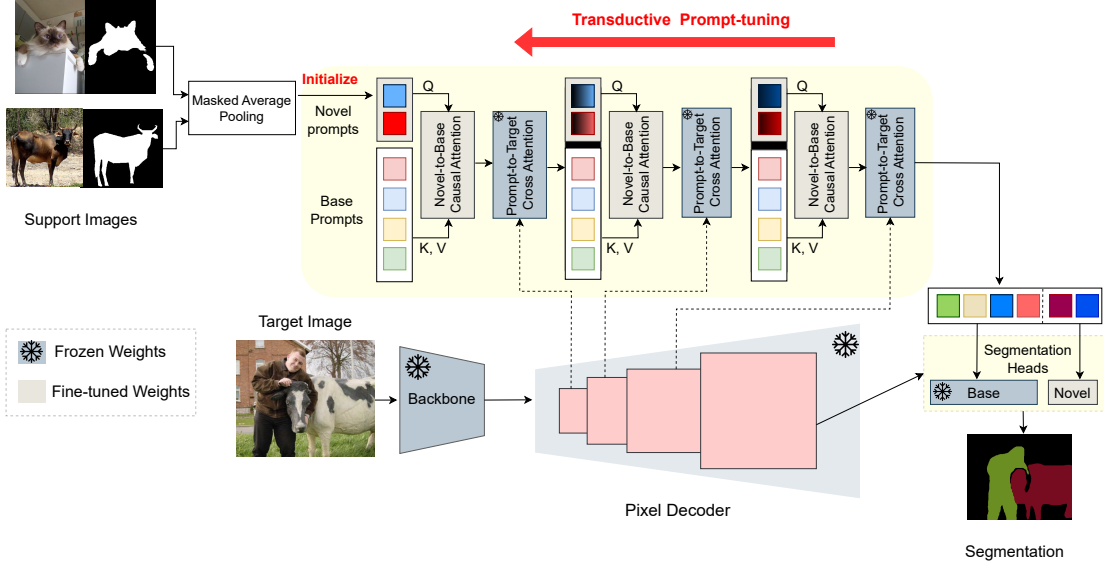
Figure 2. **Detailed architecture of our proposed visual prompting of multiscale transformer decoder.** Our design initializes the novel visual prompts using the support set. This is followed by consecutive novel-to-base causal attention, $\mathcal{CA}$, and prompt-to-target features cross attention, $\mathcal{C}$, across scales. Note that causal attention uses shared weights across the scales and decoder layers. Our design allows for transductive fine-tuning of the visual prompts leveraging the unlabelled test image.

4.1. Visual Prompting Multiscale Transformers

Base Visual Prompts. Given B base classes, we define B visual prompts each representing a class and initialize them randomly. These visual prompts first self-attend to learn the relationship between themselves and then are used to prompt the image features at multiple scales. The refinement of base prompts at each decoder layer is given by:

$$V_B^{(l)} = \mathcal{A}(V_B^{(l-1)}) + \mathcal{C}(V_B^{(l-1)}, F^{(l-1)}), \qquad (1)$$

where $\mathcal{A}$ and $\mathcal{C}$ refer to multi-headed self and cross attention respectively; $V_B^{(l)} \in \mathbb{R}^{B \times C}$ are the base visual prompts at layer l; $F^{(l)} \in \mathbb{R}^{H_l W_l \times C_l}$ is the flattened image feature at layer l; H_l and W_l denote image dimensions at that layer; C_l denotes the embedding dimension.

After refining the prompts through multiple levels of transformer attention, a segmentation head, denoted as $\mathcal{H}(\cdot)$, is defined. This segmentation head takes the final refined prompts, $V_B^{(L)}$, and the highest resolution image features, $F^{(L)}$, as input, producing per-pixel per-class predictions $O_{base} \in \mathbb{R}^{B \times H \times W}$ (see Section 4.3).

Novel Visual Prompts. After training on the base classes, we freeze all the layers of the model except the visual prompts representing the base classes. Given there are N novel classes, we add N new visual prompts. We further add a causal unidirectional attention module that allows the the novel prompts to be enhanced by the base prompts (see Section 4.2). The output is passed to a segmentation head to make predictions for the novel classes.

The novel visual prompts are first refined by a unidirectional novel-to-base causal attention module. The enhanced

novel prompts are then concatenated with the base prompts. Overall the concatenated prompts are refined by:

$$
\begin{aligned}
V_N^{(l)} &= \mathcal{CA}(V_B^{(l)}, V_N^{(l)}) \\
V_A^{(l)} &= \left[V_B^{(l)}, V_N^{(l)} \right] \\
V_A^{(l)} &= \mathcal{A}(V_A^{(l-1)}) + \mathcal{C}(V_A^{(l-1)}, F^{(l-1)}),
\end{aligned}
$$

where $V_N^{(l)} \in \mathbb{R}^{N \times C}$ are novel prompts at layer l; N is the total number of novel class and C is the embedding size; $V_A^{(l)} \in \mathbb{R}^{(B+N) \times C}$ are all visual prompts including base and novel; $[.,.]$ is the concatenation operator. The operation $\mathcal{CA}$ denotes the causal novel-to-base attention and is explained in Section 4.2.

Each of the N novel visual prompts (representing a novel class) is initialized using masked average global pooling of the k-shot support images corresponding to the class masked by the ground truth binary masks.

$$V_n^{(0)} = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{x,y} M_n^k(x,y) F_n^k(x,y)}{\sum_{x,y} M_n^k(x,y)}, \forall n \in N \quad (2)$$

where $V_n^{(0)} \in \mathbb{R}^{1 \times C}$ is the initial novel visual prompt for novel class n; F_n^k denotes the image features of k-th shot for novel class; M_n^k denotes the corresponding ground-truth binary mask; x,y are the spatial locations; N is the total number of novel classes and K is the total number of shots.

4.2. Novel-to-base Causal Attention

A key component of our proposed approach is the uni-directional novel-to-base prompt causal[3] attention. The base visual prompts are learned with abundant examples; this is not the case for novel prompts. Learning novel prompts directly or having novel prompts impact the base ones could therefore lead to degradation of performance. However, having novel prompts aware of their base counterparts could be quite helpful. We hypothesize that causal attention between the novel visual prompts and base prompts could help contextualize novel class embeddings, reducing confusion with the base counterparts. Motivated by this intuition, we propose a novel-to-base causal attention layer that is repeated at each scale (or layer of the transformer decoder); see Fig. 2.

The proposed causal attention module is a single layer cross-attention where the Key $K^{(l)}$, Value $V^{(l)}$ and $Q^{(l)}$ for layer l of the decoder are given by:

$$Q^{(l)} = V_N^{(l)} W^Q; \quad K^{(l)} = V_B^{(l)} W^K; \quad V^{(l)} = V_B^{(l)} W^V$$

where W^Q, W^K, W^V are weight matrices for Q, K and V respectively. The parameters W^Q, W^K, W^V are shared at all the layers of this cross-attention module to reduce the number of trainable parameters during fine-tuning on novel classes to prevent overfitting and poor generalization.

This uni-directional novel-to-base causal attention refines the novel visual prompts $V_N^{(l)}$ at each decoder layer l by:

$$V_N^{(l)} = \text{softmax}(Q^{(l)}(K^{(l)})^{\text{T}})V^{(l)}$$

This results in a set of contextualized novel visual prompts.

4.3. Segmentation Head

Base Segmentation Head. The base segmentation head, $\mathcal{H}_B$, during the base training in the proposed model is relatively simple and takes the form of a three layer MLP that projects a refined prompt into, effectively, a class prototype; computes the (dot product) similarity of each pixel feature with that prototype, and then applies softmax along the channel / class dimension to obtain per-pixel and per-class probability:

$$\begin{aligned} O_{base} &= \mathcal{H}_B(V_B^{(L)}, F^{(L)}) \\ &= \text{MLP}\left(V_B^{(L)}\right) \cdot \left[F^{(L)}\right]^{\text{T}}, \end{aligned} \quad (3)$$

where $O_{base} \in \mathbb{R}^{B \times H \times W}$ and the class assignment can then be made by an argmax along the channel dimension.
Novel Segmentation Head. While we could, in principle, use the same form of segmentation head during few-shot

inference as that defined for the base classes (see Eq. 3), simply fine-tuning it for novel classes, this proves ineffective in practice. Specifically, fine-tuning the MLP leads to significant overfitting given the few examples and lack of base class data during this step. So, instead, we resort to a simpler form of the novel segmentation head, $\mathcal{H}_N$, where we only learn a residual to the MLP learned during base training (which we keep fixed), mainly:

$$\begin{aligned} O_{novel} &= \mathcal{H}_N(V_N^{(L)}, F^{(L)}) \\ &= \left[\text{MLP}_\bullet(V_N^{(L)}) + W_N\right] \cdot \left[F^{(L)}\right]^{\text{T}}, \end{aligned}$$

where O_{novel} is output for novel classes and $\bullet$ designates that parameters of MLP are frozen and only $W_N \in \mathbb{R}^{N \times C}$ weight matrix is learned. This weight matrix is initialized in the same manner as novel prompts, in Eq. 2.

4.4. Transductive Prompt Tuning

Recent approaches [4, 13] show the effectiveness of a good transductive setting for few-shot dense prediction tasks. Transductive approaches perform test-time optimization for each unlabeled test image and utilize specific characteristics and information of the test instances to improve prediction.

To demonstrate the flexibility of our approach, in this section we outline the process of adapting our model for transductive inference, improving GFSS performance through transductive fine-tuning of the visual prompts. We build on the transductive losses proposed in DIaM [13] and we describe them here for completeness.

The losses are designed to maximize the mutual information between the learned features and the corresponding predictions on the test image, which is achieved through maximizing: $\text{H}(O) - \text{H}(O|I)$, where I, O are random variables associated with the pixels and predictions distributions respectively; $\text{H}(O)$ is the marginal and $\text{H}(O|I)$ is the conditional entropy. The conditional entropy $\text{H}(O|I)$ involved is given by sum of cross-entropy loss over the supervised support set and entropy of the predicted probability given the test image I. The marginal entropy is given by KL divergence loss relying on an estimated region proportion prior following [4]. Additional details are in [13].

Additionally, similar to [13], we use knowledge-distillation loss to preserve base class performance, given by:

$$\mathcal{L}_{\text{KD}} = \text{KL}(O_{base}^{new} \parallel O_{base}^{old}), \quad (4)$$

where O_{base}^{old} is the prediction probability of the base classes of the frozen model after base training; O_{base}^{new} is the prediction probability of the base classes on the model with our additional components to handle novel categories that is being fine-tuned. These include the multi-scale causal novel-to-base cross attention and the novel prompts.

The overall transductive objective function is given by:

$$\mathcal{L}_{\text{trans.}} = \alpha\text{H}(O|I) - \text{H}(O) + \gamma\mathcal{L}_{\text{KD}}, \quad (5)$$

where α and γ are balancing hyper-parameters.

We noticed that applying transduction from the first iteration underperforms, likely due to a poor initial estimate of the marginal distribution. To address this, we only apply per-pixel cross-entropy loss for a set number of iterations and then incorporate the remaining transductive losses from Eq. 5 when we achieve a more accurate estimate of the marginal distribution.

5. Experiments

5.1. Dataset and Implementation Details

Datasets. We evaluate our method on two few-shot segmentation benchmarks [36]: COCO-20^i and PASCAL-5^i. Following [13] we report average performance of our model on 10K test images for COCO-20^i and on all test images on PASCAL-5^i. The results are reported as an average across five different runs on four different splits.

Data Pre-processing. We adopt the same setup as DIAM [13]. During base training, images containing both base and novel categories are kept, but the pixels corresponding to novel classes are relabelled as background. This approach, as noted by [13], presents challenges due to potential ambiguity when the base model predicts novel categories as background. In fine-tuning, only novel classes are labeled in the support set, with base classes marked as background. To avoid unfair shot increases, images with multiple novel categories are excluded from the support set.

Evaluation Protocol. We evaluate performance using the standard mean intersection-over-union (mIoU) metric. Following [13], the mean score for generalized few-shot segmentation is computed as the average between *base* and *novel* mIoUs. This approach addresses bias in datasets like PASCAL-5^i and COCO-20^i, where base classes outnumber novel classes threefold.

Implementation Details. We use PSPNet [51] with Resnet-50 [15] as feature extractor. We utilize MSDeformAttn-based pixel decoder [53] from Mask2Former [10]. The dimension C for visual prompts is chosen to be 256. Contrary to [10], the base model is trained with per-pixel cross-entropy loss. We use AdamW optimizer [25] with base learning rate of 1×10^{-4}, a weight decay of 0.05 and a batch size of 16. Following [13, 19], we train base model for 20 epochs for COCO-20^i and 100 epochs for PASCAL-5^i.

During novel fine-tuning, we freeze the parameters of the feature extractor, pixel decoder, and base segmentation head, and only optimize the parameters of the visual prompts, novel segmentation head, and novel-to-base causal attention module. For the inductive case, the model is fine-tuned with per-pixel cross-entropy loss for 100 iterations. For the transductive setting, the model is fine-tuned for 40 iterations with only cross-entropy loss and then fine-tuned for 60 more iterations with full transductive losses mentioned in Eq. 5. The hyper-parameters α and γ are set to

be 100 and 25 respectively. We use AdamW optimizer with a learning rate of 5×10^{-3} and a weight decay of 0.05 for fine-tuning.

5.2. Results

Comparison with the State-of-The-Art. Table 1 compares our approach in both *inductive* and *transductive* setting against the state-of-the-art. As observed, our model outperforms existing approaches. Interestingly, our approach, in an inductive setting (where optimization is performed only using a supervised few-shot support set, and evaluation is conducted on all test images), outperforms transductive approaches like DIaM [13] for both 1-shot and 5-shot cases on the COCO-20^i dataset and for the 1-shot case on Pascal-5^i. This is notable despite transductive approaches requiring test-time optimization for each unlabeled test image separately, leading to significantly longer inference times.

For a fair comparison in the *inductive* setting, we fine-tuned DIaM without the transduction losses. We also fine-tuned POP [21] in our inductive setup. POP [21] randomly samples balanced base class image, making it easier for them to retain base class information and uses a form of transduction where their model relabels novel class pixels initially labelled as background. We observe that the performance of POP [21] degrades substantially, especially for the base classes in our inductive setting.

Our inductive approach has 5.78% higher mIoU for 1-shot and 11.16% higher mIoU for 5-shot on COCO-20^i than inductive DIaM. For PASCAL-5^i, the improvement from inductive DIaM is 7.71% for 1-shot and 13.29% for 5-shot. In fact, compared to the best inductive GFSS approach BAM [19], our inductive setting achieves an improvement of approximately 2.78% and 7.59% for 1-shot and 5-shot respectively on COCO-20^i and 5.24% and 12.32% for 1-shot and 5-shot respectively on PASCAL-5^i. Our method, in the inductive setting, even outperforms transductive DIaM [13] by 2.03% and 2.28% for 1-shot and 5-shot respectively on COCO-20^i and by 1.79% for 1-shot on PASCAL-5^i

On applying *transductive* fine-tuning to our method, overall mIoU increases considerably, outperforming transductive DIaM [13] and POP [21] on both 1-shot and 5-shot inference for both datasets. The improvement over DIaM for COCO-20^i is 3.3% for 1-shot and 3.93% for 5-shot. For PASCAL-5^i, our transductive setting outperforms DIaM by 5.11% and 3.15% for 1 and 5-shots.

5.3. Ablation

We performed ablation studies (refer to Table 2) to explain our design choices and contribution of each components.

Visual Prompts Initialization. As seen in **(1)** and **(2)** of the Table 2, initializing novel prompts with masked mean pooling from the support set images improves both base and novel class accuracy compared to random initialization.

| Method | Learning | COCO-20^i | | | | | |
| | | 1-shot | | | 5-shot | | |
		Base	Novel	Mean	Base	Novel	Mean
CANeT [36] (CVPR22)	Inductive	43.21	7.21	25.21	43.71	11.00	27.36
BAM [19] (CVPR22)	Inductive	49.84	14.16	32.00	49.85	16.63	33.24
DIaM (w/o trans.)* [13] (CVPR23)	Inductive	42.69	15.32	29.00	38.47	20.87	29.67
POP† [21] (CVPR23)	Inductive	30.38	9.63	20.00	24.53	16.19	20.36
Ours (w/o trans.)	Inductive	**51.55**	**18.00**	**34.78**	**51.59**	**30.06**	**40.83**
RePRI [4] (CVPR21)	Transductive	5.62	4.74	5.18	8.85	8.84	8.85
DIaM [13] (CVPR23)	Transductive	48.28	17.22	32.75	48.37	28.73	38.55
POP† [21] (CVPR23)	Transductive	**54.71**	15.31	35.01	**54.90**	29.97	42.44
Ours (w/ trans.)	Transductive	<u>53.80</u>	**18.30**	**36.05**	<u>53.81</u>	**31.14**	**42.48**

Note: The first method in the COCO table is labeled "CAPL [36]".

| Method | Learning | PASCAL-5^i | | | | | |
| | | 1-shot | | | 5-shot | | |
		Base	Novel	Mean	Base	Novel	Mean
CANeT [49] (CVPR19)	Inductive	8.73	2.42	5.58	9.05	1.52	5.29
PFENET [35] (TPAMI20)	Inductive	8.32	2.67	5.50	8.83	1.89	5.36
PANET [40] (ICCV19)	Inductive	31.88	11.25	21.57	32.95	15.25	24.1
SCL [47] (CVPR21)	Inductive	8.88	2.44	5.66	9.11	1.83	5.47
MiB [8] (CVPR20)	Inductive	63.80	8.86	36.33	68.60	28.93	48.77
CAPL [36] (CVPR22)	Inductive	64.80	17.46	41.13	65.43	24.43	44.93
BAM [19] (CVPR22)	Inductive	71.60	27.49	49.55	71.60	28.96	50.28
DIaM (w/o trans.)* [13] (CVPR23)	Inductive	66.79	27.36	47.08	64.05	34.56	49.31
POP† [21] (CVPR23)	Inductive	46.68	19.96	33.32	41.50	36.26	38.80
Ours (w/o trans.)	Inductive	**74.58**	**34.99**	**54.79**	**74.86**	**50.34**	**62.60**
RePRI [4] (CVPR21)	Transductive	20.76	10.50	15.63	34.06	20.98	27.52
DIaM [13] (CVPR23)	Transductive	70.89	35.11	53.00	70.85	55.31	63.08
POP† [21] (CVPR23)	Transductive	73.92	35.51	54.72	74.78	55.87	65.33
Ours (w/ trans.)	Transductive	**76.39**	**39.83**	**58.11**	**76.42**	**56.12**	**66.27**

Table 1. Comparison to SoTA methods on COCO-20^i and PASCAL-5^i. The mIoU results are reported as the average across 5 different runs. All methods use ResNet-50 backbone. † POP [21] uses an easier and less practical GFSS setup, e.g., balanced base class training in addition to support set during novel fine-tuning and uses a form of transduction; inductive setting is implemented by us following our setup. Results of the rest of the methods are obtained from [13] except the inductive setting of DIaM*, which we implemented ourselves. The **w/o trans.**, **w/ trans.** denote without and with transduction respectively.

| | Causal Attention | Prompt Initialization | Transduction | mIoU | | |
				Base	Novel	Mean
(1)	None	Random	No	50.18	11.22	30.70
(2)	None	Masked Pooling	No	50.43	11.31	30.87
(3)	First Layer Only	Masked Pooling	No	51.53	12.26	31.90
(4)	Separate Weights	Masked Pooling	No	51.06	18.05	34.56
(5)	Shared Weights	Random	No	50.75	17.41	34.08
(6)	Shared Weights	Masked Pooling	No	<u>51.55</u>	18.00	<u>34.78</u>
(7)	Shared Weights	Masked Pooling	Yes	**53.80**	**18.30**	**36.05**

Table 2. **Ablation Study on COCO-20^i.** The 1-shot performane, averaged over all folds.

This is more pronounced with novel-to-base causal attention **(5)**–**(6)**, amplifying the performance gain from masked pooling initialization.

Novel-to-base Causal Attention. As seen in **(4)**–**(7)** of Table 2, novel-to-base causal attention yields the most significant improvement in novel class accuracy compared to **(1)**–**(2)**. This aligns with our hypothesis that uni-directional causal attention between rich base prompts (learned with many examples) and novel prompts (learned with few examples) increases separation among novel classes and pushes the feature space of base prompts away from the novel prompts. It is further confirmed by Figure 3 and Fig-
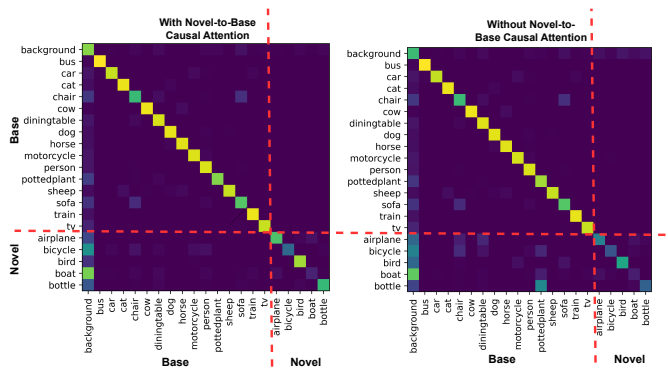
Figure 3. **Confusion Matrix for:** (left) model with Novel-to-Base Causal Attention; (right) model without Novel-to-Base Causal Attention. Note that Novel-to-Base Causal Attention reduces confusion between novel and base categories (bottom-left block).

ure 4. As observed in Figure 3, the novel-to-base causal attention leads to lower confusion between base and novel classes. Figure 4 shows a t-SNE [37] visualization of the features of base and novel visual prompts for 10 different
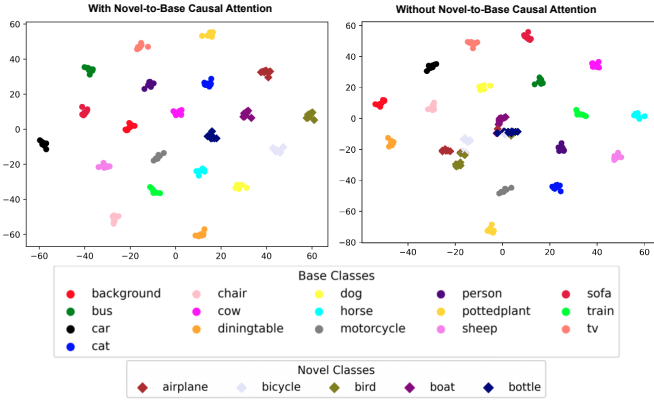
Figure 4. **TSNE visualization of the learned base and novel prompt features:** (left) with Novel-to-Base Causal Attention; (right) without Novel-to-Base Casual Attention.

Figure 5. **Qualitative Results** for 1-shot on Pascal-5^i. The left-omost two columns show image and ground truth mask; (Third) Baseline without causal attention; (Fourth); Ours in inductive setting; (Fifth) Ours in transductive setting; (Last) DIaM [13]. Last row illustrates a *failure*.

test images. As observed, the incorporation of novel-to-base causal attention leads to much clearer separation of visual prompt features, particularly for novel prompts.

Multiscale Causal Attention. We also ablated the placement of the causal attention module. We observe that applying it only at the first layer **(3)** slightly improves novel mIoU compared to the model without causal attention. Yet, it is notably inferior to applying at every decoder layer **(4)**–**(7)**. We could have opted for separate parameters for the causal attention module at each layer. However, as seen in **(4)**, having separate weights offers little improvement to novel mIoU despite significant increase in total number of parameters to optimize. In fact, it decreases the base accuracy, leading to overall worse performance.

Qualitative Results. Figure 5 compares the qualitative results of our model for 1-shot GFSS on Pascal-5^i in both inductive and transductive settings with the baseline lacking causal attention and DIaM [13]. Notably, for first row, both our inductive and transductive approaches predict more accurate segmentation of the *cow* compared to DIaM. In second row, the causal attention improves *person* segmentation followed by transduction. In the third row, DIaM completely misses the novel category *chair*. The causal attention allows our model to better differentiate the base class *dining table* from the novel class *chairs*. Transduction improves accuracy of the segmentation of *dining table*. In fourth row, we can observe that both DIaM and our model without causal attention misdetects novel class *cat* for *cow* at multiple locations. Causal attention allows our model to make a uniform and more accurate prediction of *cat*. In the fifth row, we observe that DIaM [13] completely misses detecting the base category *potted plant* and poorly segments *person*. The baseline without causal attention misdetects the *potted plant* with novel class *bus* and misdetects *dining table* with *chair*. Causal attention improves the accuracy which is further improved by transduction. Last row
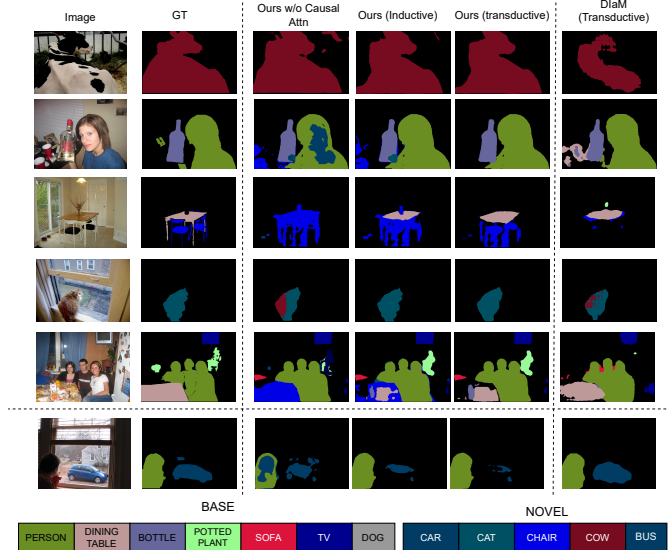
demonstrates a case where our model performs worse than DIaM [13]. Again, we observe that adding causal attention reduces misclassification of *person* with *car*, although overall the segmentation of *car* is less accurate than DIaM [13].

From the qualitative results, the effectiveness of novel-to-base causal attention at distinguishing base from novel classes becomes apparent. Our model in the inductive setting outperforms transductive DIaM in most cases, despite the inference time being faster.

6. Conclusion

In this work, we proposed a technique to learn visual prompts from few-shot examples of novel unseen cateogries to prompt a pre-trained segmentation model that not only generalizes well on the unseen novel class but strongly retains performance over the base classes on which the model was pre-trained. We achieve this by prompting the image features at multiple scales and proposing a uni-directional novel-to-base causal attention mechanism that enriches the novel visual prompts through interaction with the base visual prompts. Our proposed method can be applied in both inductive and transductive settings and achieves state-of-the-art performance on two benchmark few-shot segmentation datasets.

References

[1] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 23716–23736, 2022. 3

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. v. Arx, M. S. Bernstein, A. Bosselut J. Bohg, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, S. Ermon E. Durmus, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, D. Narayanan A. Narayan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 3

[3] Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:2445–2457, 2020. 2

[4] M. Boudiaf, H. Kervadec, Z. Masud, P. Piantanida, I. Ben Ayed, and J. Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13979–13988, 2021. 1, 2, 5, 7

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, G. Sastry P. Shyam, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 1877–1901, 2020. 1, 2

[6] A. Bulat, R. Guerrero, B. Martinez, and G. Tzimiropoulos. Fs-detr: Few-shot detection transformer with prompting and without re-training. In *International Conference on Computer Vision (ICCV)*, pages 11793–11802, 2023. 1, 3

[7] T. Cao, M. Law, and S. Fidler. A theoretical analysis of the number of shots in few-shot learning. *International Conference on Learning Representations (ICLR)*, 2020. 2

[8] F. Cermelli, M. Mancini, S.R. Bulo, E. Ricci, and B. Caputo. Modeling the background for incremental learning in semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9233–9242, 2020. 7

[9] W. Chen, Y. Liu, Z. Kira, Y. Frank Wang, and J. Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[10] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 3, 6

[11] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[12] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017. 2

[13] S. Hajimiri, M. Boudiaf, I. B. Ayed, and J. Dolz. A strong baseline for generalized few-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11269–11278, 2023. 2, 3, 5, 6, 7, 8

[14] B. Hariharan and R. B. Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[16] M. Jia, L. Tang, B. Chen, C. Cardie, S. Belongie, B. Hariharan, and S. Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pages 709–727, 2022. 2, 3

[17] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell. Few-shot object detection via feature reweighting. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, A. C. Berg S. Whitehead, W. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[19] C. Lang, G. Cheng, B. Tu, and J. Han. Learning what not to segment: A new perspective on few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8057–8067, 2022. 6, 7

[20] X. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[21] 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, 6, 7

[22] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. Hwang, and Y. Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[23] Y. Liu, J. Lee, M. Park, S. Kim, E. Yang, S. J. Hwang, and Y. Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[24] Y. Liu, X. Zhang, S. Zhang, and X. He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[25] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[26] J. Min, D. Kang, and M. Cho. Hypercorrelation squeeze for few-shot segmentation. In *International Conference on Computer Vision (ICCV)*, pages 6941–6952, 2021. 1, 2

[27] A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[28] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 2

[29] Guodong Qi, Huimin Yu, Zhaohui Lu, and Shuzhao Li. Transductive few-shot classification on the oblique manifold. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8412–8422, 2021. 2

[30] L. Qiao, Y. Shi, J. Li, Y. Wang, T. Huang, and Y. Tian. Transductive episodic-wise adaptive metric for few-shot learning. In *International Conference on Computer Vision (ICCV)*, pages 3603–3612, 2019. 2

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1

[32] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[33] M. Ren, E. Triantafillou, S. Ravi, J.Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2018.

[34] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[35] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(2):1050–1065, 2020. 7

[36] Z. Tian, X. Lai, L. Jiang, S. Liu, M. Shu, H. Zhao, and J. Jia. Generalized few-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11563–11572, 2022. 2, 3, 6, 7

[37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(11), 2008. 7

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 2, 5

[39] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3630–3638, 2016. 2

[40] K. Wang, J. Liew, Y. Zou, D. Zhou, and J. Feng. PANet: Few-shot image semantic segmentation with prototype alignment. In *International Conference on Computer Vision (ICCV)*, pages 9197–9206, 2019. 1, 2, 7

[41] X. Wang, T. E Huang, T. Darrell, J. E. Gonzalez, and F. Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning (ICML)*, 2020. 2

[42] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[43] X. Wu, F. Zhu, R. Zhao, and H. Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7031–7040, 2023. 1

[44] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin. Meta R-CNN: Towards general solver for instance-level low-shot learning. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[45] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 763–778, 2020. 1, 2

[46] X. Yang, B. Wang, K. Chen, X. Zhou, S. Yi, W. Ouyang, and L. Zhou. BriNet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation. In *British Machine Vision Conference (BMVC)*, 2020. 1, 2

[47] B. Zhang, J. Xiao, and T. Qin. Self-guided and cross-guided learning for few-shot segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8312–8321, 2021. 7

[48] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *International Conference on Computer Vision (ICCV)*, pages 9587–9595, 2019. 1, 2

[49] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen. CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5217–5226, 2019. 1, 2, 7

[50] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *Advances in Neural Information Processing Systems*, 35:6575–6588, 2022. 1, 3

[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 2881–2890, 2017. 6

[52] Hao Zhu and Piotr Koniusz. Transductive few-shot learning with prototype-based label propagation by iterative graph refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23996–24006, 2023. 2

[53] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations (ICLR)*, 2021. 6