Learning Orthogonal Prototypes for Generalized Few-shot Semantic Segmentation

Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie*, Yongdong Zhang, Ting Yao
University of Science and Technology of China HiDream.ai Inc.
{lsa1997.mail, htxie, zhyd73}@ustc.edu.cn, {yihengzhang.chn, zhaofanqiu, tingyao.ustc}@gmail.com

Abstract

Generalized few-shot semantic segmentation (GFSS) distinguishes pixels of base and novel classes from the background simultaneously, conditioning on sufficient data of base classes and a few examples from novel class. A typical GFSS approach has two training phases: base class learning and novel class updating. Nevertheless, such a standalone updating process often compromises the well-learnt features and results in performance drop on base classes. In this paper, we propose a new idea of leveraging Projection onto Orthogonal Prototypes (POP), which updates features to identify novel classes without compromising base classes. POP builds a set of orthogonal prototypes, each of which represents a semantic class, and makes the prediction for each class separately based on the features projected onto its prototype. Technically, POP first learns prototypes on base data, and then extends the prototype set to novel classes. The orthogonal constraint of POP encourages the orthogonality between the learnt prototypes and thus mitigates the influence on base class features when generalizing to novel prototypes. Moreover, we capitalize on the residual of feature projection as the background representation to dynamically fit semantic shifting (i.e., background no longer includes the pixels of novel classes in updating phase). Extensive experiments on two benchmarks demonstrate that our POP achieves superior performances on novel classes without sacrificing much accuracy on base classes. Notably, POP outperforms the state-of-the-art fine-tuning by 3.93% overall mIoU on PASCAL-5^i in 5-shot scenario.

1. Introduction

Semantic segmentation is to assign semantic labels to every pixel of an image. With the recent development of CNNs [10, 13] and vision transformers [6, 18, 23, 24, 35, 41, 42], the state-of-the-art networks have successfully pushed the limits of semantic segmentation [1, 3, 25, 49] with remarkable performance improvements. Such achievements heavily rely on the requirements of large quantities of pixel-level annotations and it is also difficult to directly apply the models to the classes unseen in the training set. A straightforward way to alleviate this issue is to leverage Few-shot Semantic Segmentation (FSS) [30, 36, 44], which utilizes the limited support annotations from unseen/novel classes to adapt the models. Nevertheless, FSS performs on the assumption that the support images and the query image contain the same novel classes, and solely emphasizes the segmentation of one novel class in the query image at a time. A more practical scenario namely Generalized Few-shot Semantic Segmentation (GFSS) [33] is recently presented to simultaneously identify the pixels of base and novel classes in a query image.

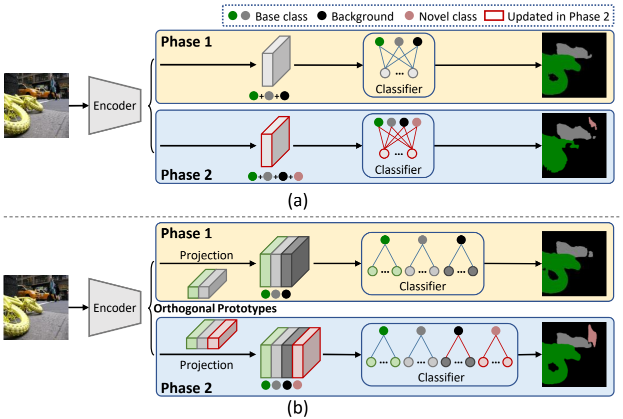

Figure 1. Comparisons between fine-tuning [26] and Projection onto Orthogonal Prototypes (POP). In novel class updating phase, fine-tuning (a) updates the network to predict base and novel classes, and inevitably compromises the well-learnt representations for base classes. Instead, POP (b) only updates prototypes for novel classes and executes predictions for each class separately on the features projected onto different orthogonal prototypes.

A typical GFSS solution has proceeded along two training phases: base class learning and novel class updating. In the first phase, models are trained on abundant base classes' annotations to classify the pixels of base categories, and then updated with the limited labeled novel examples in the second phase to additionally recognize pixels of novel

[*] Corresponding author.

classes. For instance, Myers Dean *et al*. [26] sample some base data plus novel examples as the supervision to fine-tune the network, as depicted in Figure 1(a). Despite having good performances on novel classes, there still exists a clear performance degradation on base classes. We speculate that this may be the results of compromising the well-learnt features of base classes in fine-tuning. As such, a valid question then emerges as is it possible for GFSS to nicely generalize the model to novel classes without sacrificing much segmentation capability on base classes. In an effort to answer this question, we seek to represent an image via a group of uncorrelated feature components each of which characterizes a specific class. By doing so, it is readily applicable to learn and integrate new components for novel classes without affecting the ones learnt for base classes.

To materialize our idea, we propose a new Projection onto Orthogonal Prototypes (POP) framework for GFSS. POP learns a series of orthogonal prototypes and each prototype corresponds to one specific semantic class. As shown in Figure 1(b), POP employs an encoder to extract feature maps of a given query image and then projects them onto prototypes. The projection on each prototype is regarded as the discriminative representation with respect to the corresponding class and exploited to predict the probability map of pixels belonging to the class. More specifically, POP deliberately devises the learning of prototypes from three standpoints. The first one is to freeze the base prototypes when learning novel ones from support images in the updating phase. In this way, feature projections on base prototypes maintain their discriminability of base classes. Second, POP encourages the prototypes of base and novel classes to be orthogonal through a prototype orthogonality loss. Such a constraint decorrelates features projected onto different prototypes and mitigates the inter-class confusion caused by extending to novel classes. Finally, in view that background no longer contains the pixels of novel classes in updating phase, known as “semantic shifting”, POP measures the residual of feature projection as the background representation instead of learning a prototype for “background”. This way further improves the differentiation between novel classes and background.

The main contribution of this work is the proposal of a Projection on Orthogonal Prototypes (POP) framework for generalized few-shot semantic segmentation. The solution also leads to the elegant views of how to adapt the trained model to novel classes without sacrificing well-learnt features, and how to represent background pixels dynamically in the context of semantic shifting, which are problems not yet fully explored in the literature. We demonstrate that POP outperforms the state-of-the-art fine-tuning [26] on two benchmarks (PASCAL-5^i and COCO-20^i) with evident improvements on both base and novel classes.

2. Related Work

Semantic Segmentation. Extensive FCN-based methods have been proposed and achieved remarkable performances. For example, pyramid pooling [12, 48, 50] and parallel dilated convolutions [1, 2, 46, 47] are utilized to aggregate multi-scale context. The attention mechanism is widely exploited to model long-range dependencies [8, 14, 17, 21, 43, 45, 51]. Recently, vision transformer has also been proven effective for semantic segmentation [3, 32, 39]. Though these efforts lead to high-quality segmentation, it is still difficult to directly apply the learnt models to unseen classes.

Few-shot Semantic Segmentation. Few-shot semantic segmentation (FSS) [30] executes pixel-wise labeling on new classes given a limited number of support examples. It mainly focuses on the 1-way scenario that requires binary maps for query images to identify the pixels belonging to the class labeled in support images. PL [5] and PANet [36] remould prototype learning for FSS by calculating the cosine similarities between pixels and prototypes obtained from support images. ASR [20] learns several orthogonal prototypes on the base data to represent novel categories. In addition, assigning multiple prototypes to each class is also a promising solution to improve FSS models [16,22,38,40].

Generalized Few-shot Semantic Segmentation. Despite showing great potential, it is not trivial for FSS approaches to simultaneously identify the pixels of both base and novel classes in a query image. Generalized Few-shot Semantic Segmentation (GFSS) is further presented by Tian *et al*. [33] to address this problem. GFSS aims to segment both base and novel classes from query images through one inference process without paired support images. To achieve this goal, CAPL [33] mines contextual cues from both support and query samples to enrich the classifier for novel class segmentation. Fine-tuning [26] is another practical and effective way to adapt models for novel classes.

Orthogonality. Feature orthogonality is known to be helpful to improve discriminative ability. DSN [31] constrains the orthogonality on base data to facilitate few-shot classification on novel classes. C-FSCIL [11] and OPL [29] rely on re-training the network to keep the features of each class orthogonal. Nevertheless, these approaches are not applicable for GFSS since they either ignore the influence between base and novel classes or compromise the segmentation capability on base classes after re-training.

Our work focuses on adapting the semantic segmentation model to novel classes without compromising well-learnt features for better generalization. Different from existing GFSS approaches may sacrifice base class segmentation quality for novel class updating, our POP contributes by studying not only learning orthogonal prototypes to independently indicate pixels of base and novel classes, but also how the background could be nicely represented in the context of semantic shifting.

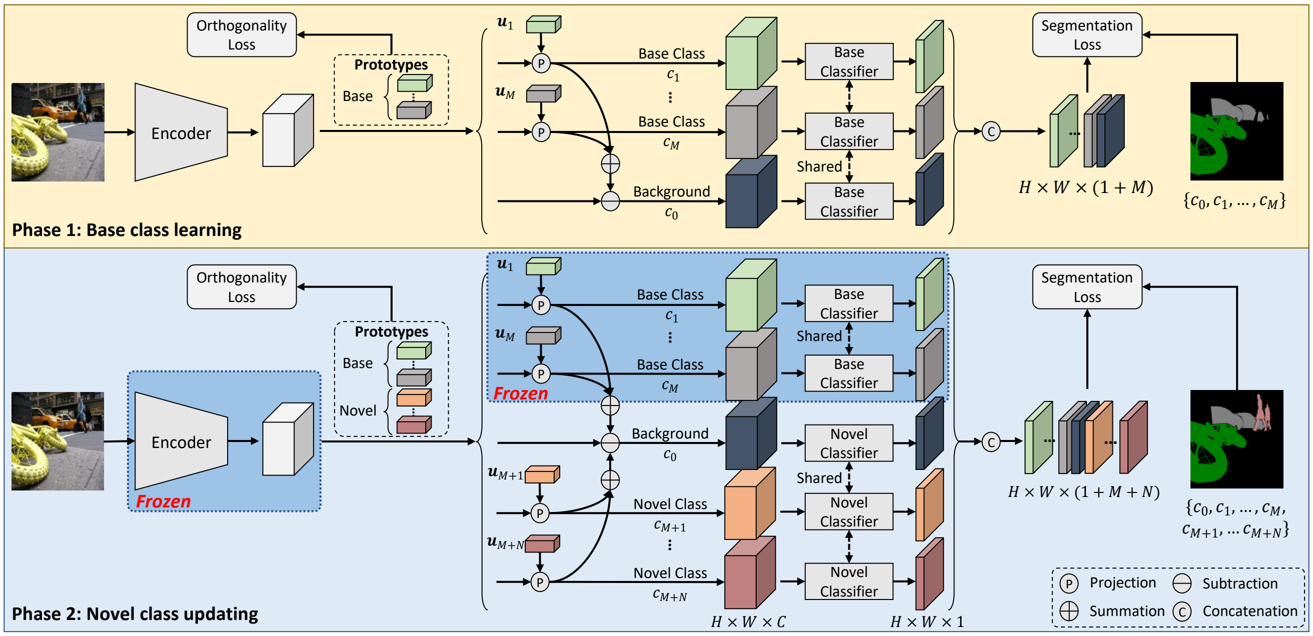

Figure 2. An overview of our Projection onto Orthogonal Prototypes (POP) framework for generalized few-shot semantic segmentation. The input image is first encoded into dense feature map via a CNN encoder, which is then projected onto a series of orthogonal prototypes to make prediction of each class independently through a shared classifier. In the base class learning phase, the entire network and prototypes for base classes are learnt on sufficient base data. In the second phase of novel class updating, we freeze all the modules learnt in the first phase and only optimize the newly minted prototypes for novel classes along with an additional classifier for novel classes and background class. As such, the proposed POP generalizes well to novel classes while maintaining the segmentation capability on base classes.

3. Our Method

In this paper, we devise Projection onto Orthogonal Prototypes (POP) framework for generalized few-shot semantic segmentation (GFSS), by decoupling the feature learning of base classes and novel classes. Figure 2 illustrates an overview of the whole architecture of POP. We begin this section by elaborating the problem formulation of GFSS with the two-phase training scheme. Then, a projection-based framework is introduced to learn class-wise orthogonal prototypes, and extract the residual representation as background descriptor. Finally, we present the overall objective combining orthogonal constraint along with the semantic segmentation loss.

3.1. Problem Formulation

In generalized few-shot semantic segmentation, we are given abundant annotated images for M base classes $C^b = \{c_1, \ldots, c_M\}$ and only K labeled images per class for N novel classes $C^n = \{c_{M+1}, \ldots, c_{M+N}\}$. The background category c_0 presents the pixels that do not belong to any target classes. The goal of this task is to simultaneously distinguish both base and novel classes from backgrounds, i.e., $M + N + 1$ classes in total. Inspired by recent success of two-phase training scheme in generalized few-shot semantic segmentation [26, 33], we formulate our POP framework in two-phase manner: base class learning and novel class updating. Accordingly, POP is first trained with sufficient data for $(M + 1)$-class segmentation $(C^b \cup \{c_0\})$ in the first phase, and then updates the segmentation head with the few-shot support data targeting for pixel-wise prediction of all $M + N + 1$ classes $(C^b \cup C^n \cup \{c_0\})$ in the second phase.

3.2. Projection onto Orthogonal Prototypes

GFSS necessitates the capability to segment both base and novel classes. In this case, the novel updating process should learn and integrate new components for novel classes without affecting the ones learnt for base classes. However, the existing approaches [26, 33] in this area still show obvious performance degradation on base classes in the second phase. They often suffer from a challenge of whether to drastically overhaul the network for the performance on novel classes or preserve the well-learnt representation for base classes. We attribute this issue to the highly-relevant representations to identify different classes, which interfere with each other when integrating novel classes.

To fulfill the objective of decoupling the feature learning of base classes and novel classes, we propose to orthogonally decompose the feature map into several uncorrelated components, each of which characterizes a specific class. Formally, we denote the dense feature map from the encoder network as $\boldsymbol{f} \in \mathbb{R}^{(H \times W) \times C}$, where H, W and C are the height, the width and the number of channels of feature map, respectively. Assuming the matrix rank of extracted features is r ($r \leq C$), the feature map $\boldsymbol{f}$ can be represented as the weighted summation of r orthogonal bases as:

$$\begin{aligned} \boldsymbol{f} &= \sum_{i=1}^{r} f^{(i)} \cdot \boldsymbol{u}_i, \\ \text{s.t. } & \boldsymbol{u}_i \cdot \boldsymbol{u}_j{}^T = 0 \ (i \neq j), \end{aligned} \tag{1}$$

where $\{\boldsymbol{u}_i \in \mathbb{R}^{1\times C}|_{i=1\ldots r}\}$ is a group of unit bases that are orthogonal to each other, and $f^{(i)} \in \mathbb{R}^{(H\times W)\times 1}$ is the weight map of the i-th component. This process is well known as orthogonal decomposition, and widely explored for Independent Component Analysis (ICA) [4]. In our work, we treat the basis $\boldsymbol{u}_i$ as the prototype of the i-th class and the corresponding component $f^{(i)} \cdot \boldsymbol{u}_i$ as the representation to identify the pixels of the i-th class. As such, calculating the representation of the i-th class is equivalent to projecting $\boldsymbol{f}$ to the i-th basis $\boldsymbol{u}_i$:

$$\boldsymbol{f}_{c_i} = f^{(i)} \cdot \boldsymbol{u}_i = (\boldsymbol{f} \cdot \boldsymbol{u}_i{}^T) \cdot \boldsymbol{u}_i, \tag{2}$$

where $\boldsymbol{f}_{c_i}$ is the representation the of the i-th class and is then taken as the input of a shared classifier. In view that the learnt prototypes are orthogonal to each other, the representations of different classes are decorrelated and easy to be distinguished. More importantly, when integrating a new target class, we only need to learn a new prototype that is orthogonal to the existing ones, thereby preserving the well-learnt representations of other classes.

Nevertheless, the limit of prototype orthogonality (i.e., $\boldsymbol{u}_i \cdot \boldsymbol{u}_j{}^T = 0$) is not differentiable and hard to be implemented by gradient propagation. Hence, we present a practical alternative to relax the hard limit to a soft orthogonal constraint. Specifically, a prototype orthogonality loss is devised by minimizing absolute value of the inner product between prototypes:

$$L_{orth} = \frac{1}{r'(r'-1)} \sum_{i\neq j} |\boldsymbol{u}_i \cdot \boldsymbol{u}_j{}^T|, \tag{3}$$

where $r' < r$ is the number of target classes. The orthogonality loss L_{orth} restricts the averaged correlation of different prototype pairs, and is minimized when the prototypes are orthogonal to each other.

3.3. Residual as Background Representation

The background class c_0 is a special category in segmentation tasks, and indicates the pixels that do not belong to any target classes. One common practice is to treat target classes and background class equally, which reaches reasonable performances on various segmentation tasks. However, such a simple strategy is not acclimatized to GFSS task due to the effect of the "semantic shifting" between two phases. More precisely, in the first phase, the pixels of both novel classes C^n and background class c_0 are labeled as background. When integrating novel classes in the second phase, the semantic meaning of background changes since it no longer includes the pixels of novel classes. We regard this gap as semantic shifting of background, which results in the confusion between novel classes and background class.

In order to bridge the gap between two phases, we propose to exploit the residual of feature projection as the background representation in POP, instead of constructing an independent prototype. The residual here is defined as the summation of remaining components after projecting $\boldsymbol{f}$ to the target classes. For instance, in the base class learning phase, the first M components in Eq.(1) are the projections for M base classes and the remaining $r - M$ components are background components:

$$\boldsymbol{f} = \underbrace{\sum_{i=1}^{M} f^{(i)} \cdot \boldsymbol{u}_i}_{base\ classes} + \underbrace{\sum_{i=M+1}^{r} f^{(i)} \cdot \boldsymbol{u}_i}_{background}. \tag{4}$$

Note that the background here actually includes the pixels of novel and background classes. In view of the difficulty on learning the remaining $r - M$ components, we reformulate the background representation as the residual of feature projection by subtracting the components of bases classes:

$$\boldsymbol{f}_{c_0} = \sum_{i=M+1}^{r} f^{(i)} \cdot \boldsymbol{u}_i = \boldsymbol{f} - \sum_{i=1}^{M} f^{(i)} \cdot \boldsymbol{u}_i, \tag{5}$$

where $\boldsymbol{f}_{c_0}$ is orthogonal to the feature $\boldsymbol{f}_{c_i}$ of each class c_i due to the orthogonality constraint between prototypes. Moreover, this reformulation avoids the estimation of matrix rank r, which is more practical.

For the novel class updating phase, the background components in Eq. (4) are separated into two parts, i.e., N components for N novel classes and $r - M - N$ components as the updated background representation. The decomposition of feature $\boldsymbol{f}$ in this phase is represented as

$$\boldsymbol{f} = \underbrace{\sum_{i=1}^{M} f^{(i)} \cdot \boldsymbol{u}_i}_{base\ classes} + \underbrace{\sum_{i=M+1}^{M+N} f^{(i)} \cdot \boldsymbol{u}_i}_{novel\ classes} + \underbrace{\sum_{i=M+N+1}^{r} f^{(i)} \cdot \boldsymbol{u}_i}_{background}. \tag{6}$$

Similar to Eq.(5), the updated background representation is measured by the residual of feature projection to base and novel classes:

$$\boldsymbol{f}_{c_0} = \boldsymbol{f} - \sum_{i=1}^{M} f^{(i)} \cdot \boldsymbol{u}_i - \sum_{i=M+1}^{M+N} f^{(i)} \cdot \boldsymbol{u}_i. \tag{7}$$

In this way, the background representation is dynamically adjusted with respect to the change of target classes. It remains orthogonal to the representations of both base classes and novel classes, such that the confusion between novel classes and background class is alleviated.

3.4. Optimization

Following the common practice [26, 33], the optimization of our POP framework contains two phases: base class learning and novel class updating.

Base class learning. In the first phase, the entire network is trained on sufficient base data to optimize the encoded representations and segment base classes. The overall training objective function of POP combines the semantic segmentation loss of base classes and the orthogonality loss between prototypes: $L = L_{seg} + \lambda L_{orth}$, where L_{seg} denotes the cross entropy loss for segmentation and λ is a scale factor of the orthogonality loss.

Novel class updating. The second phase involves the K-shot support data of novel classes and expands the set of target classes. We freeze all the modules learnt in the first phase and only optimize the newly constructed prototypes for novel classes along with an additional classifier for novel classes and background class. The loss function is the same as that in the first phase (i.e., $L = L_{seg} + \lambda L_{orth}$), and the orthogonality loss L_{orth} is utilized over the prototypes of both base classes and novel classes.

Data enrichment. The second phase demands the training images of both base and novel classes. As a result of highly unbalanced data between base classes and novel classes, simply merging the two training sets leads to poor performance on novel classes. Inspired by the fine-tuning approaches [26, 37], we establish a balanced training subset by randomly sampling K examples per base class. Such a subset in the second phase has been proven to be effective for partially retraining the knowledge of base classes while integrating novel classes [26, 37]. Besides, we further promote the subset by two data enrichment strategies:

1) Relabeling. Recall that there is semantic shifting problem between two training phases. In the training images of base classes, the pixels of novel classes are labeled as background. Such annotation misleads the network optimization in the second phase. Therefore, after each epoch, we utilize the latest POP model to relabel the background pixels in the image of base classes, and replace the background annotation with the class with the highest predicted probability across the novel classes and background class.

2) Resampling. The subset only selects a few samples from the images of base classes. It underuses the large amount of training data, and makes the network sensitive to the choice of few-shot samples. Hence, we propose to resample the training images of base classes after each epoch. This strategy, on one hand, well balances the number of training images for different classes, and on the other, makes full use of the entire base training set.

4. Experiments

We empirically verify the merit of our POP by conducting a thorough evaluation of generalized few-shot semantic segmentation on PASCAL-5^i [30] and COCO-20^i [27].

4.1. Datasets and Experimental Settings

Datasets. The PASCAL-5^i [30] is built on the PASCAL VOC 2012 [7] dataset with the extended annotations of SDS [9], and contains 12,031 images with high-quality pixel-level annotations of 20 classes. All the images are split into two sets of 10,582 and 1,449 for training and validation, respectively. Following the standard protocol in [30, 33], the 20 classes are evenly partitioned into 4 folds (5 classes per fold) for cross-validation. Furthermore, we evaluate the capability of our POP on a more challenging COCO-20^i [27] dataset. COCO-20^i is created upon COCO [19] , and includes 122,218 images being annotated with 80 object classes. In between, 82,081 and 40,137 images are exploited for training and validation. Similarly, experiments on COCO-20^i are also conducted with cross-validation on 4 folds (20 classes per fold) following [27,33].

Evaluation Settings. For both datasets, once we validate the model on one fold, the classes in this fold serve as "novel classes" and the classes in the other 3 folds plus background play the role of "base classes". In the base class learning phase, we select all images that have at least one pixel belonging to base classes from the original training set for model training. Note that the pixels of novel classes in the selected images are considered as background at this stage. In the novel class updating phase, we mimic the few-shot setting by randomly sampling K images containing pixels of novel classes from the original training set. All the images in the validation set are utilized for the evaluation on both base and novel classes. In the experiments, the Intersection over Union (IoU) per class and mean IoU (mIoU) over the base, novel, and all classes are adopted as the performance metrics. We average the mIoU over all the folds for cross-validation and regard the mean values as the final performance. For stability, we repeat each experiment five times with different random seeds and report the mean values. The same setting is also exploited in [26, 33].

Network Structure. In order to make the capacity of our model comparable to the baselines of generalized few-shot semantic segmentation, we follow [33] and exploit the PSPNet [50] originated from the ImageNet pre-trained ResNet-50 as the encoder. The spatial resolution of the feature map extracted through the encoder is $\frac{1}{8}$ of the input image and the probability maps produced by POP are upsampled via bilinear interpolation for pixel-level predictions. The number of prototypes depends on the datasets. For PASCAL-5^i [30], POP has 15 prototypes in the base class learning phase, and then 5 prototypes are additionally included for novel class updating. Similarly, POP maintains 60 and 80

Table 1. Performance comparisons on PASCAL-5^i under the generalized few-shot segmentation settings. We report mIoU (%) over base classes (Base), novel classes (Novel), and all classes (Base + Novel = Total). All models are based on ResNet-50. † indicates that the images containing novel classes are excluded by base class learning phase.

Methods	1-shot			5-shot			10-shot		
	Base	Novel	Total	Base	Novel	Total	Base	Novel	Total
CANet [44]	8.73	2.42	7.23	9.05	1.52	7.26	-	-	-
PFENet [34]	8.32	2.67	6.97	8.83	1.89	7.18	-	-	-
PANet [36]	31.88	11.25	26.97	32.95	15.25	28.74	-	-	-
BAM [15]	73.62	32.96	63.94	73.57	34.35	64.23	-	-	-
CAPL† [33]	65.48	18.85	54.38	66.14	22.41	55.72	69.09	27.17	59.11
FT [26]	66.84	18.82	55.41	72.03	46.40	65.93	73.02	52.55	68.14
FT-Triplet [26]	66.41	19.71	55.31	71.31	50.46	66.35	72.87	57.00	69.10
POP†	67.73	19.85	56.45	71.34	43.48	64.71	72.23	50.49	67.05
POP	**73.92**	**35.51**	**64.77**	**74.78**	**55.87**	**70.28**	**74.99**	**58.77**	**71.13**

Table 2. Performance comparisons on COCO-20^i under the generalized few-shot segmentation settings. Models are based on ResNet-50.

Methods	1-shot			5-shot			10-shot		
	Base	Novel	Total	Base	Novel	Total	Base	Novel	Total
CAPL† [33]	44.61	7.05	35.46	45.24	11.05	36.80	45.51	10.82	36.95
FT [26]	43.42	8.94	34.90	47.18	24.72	41.63	48.18	30.03	43.70
FT-Triplet [26]	43.64	9.23	35.14	46.61	28.84	41.36	46.61	34.49	43.27
POP†	47.09	7.27	37.26	48.61	20.07	41.56	49.13	26.48	43.54
POP	**54.71**	**15.31**	**44.98**	**54.90**	**29.97**	**48.75**	**55.01**	**35.05**	**50.08**

prototypes in two training phases on COCO-20^i, respectively. The dimension of the prototypes is set as 512.

Training Strategy. Our POP is implemented on PyTorch [28]. The mini-batch stochastic gradient descent with momentum 0.9 and weight decay 0.0001 is exploited to optimize the model. During the base class learning, the initial learning rate is set to 0.01 which is annealed down to zero following a "poly" policy whose power is fixed to 0.9. The batch size is set as 8 for both datasets and the training of each model takes 50 epochs. For novel class updating, we update the model with a fixed learning rate 0.01 for 500 epochs, and the batch size is set as 2 and 8 for PASCAL-5^i and COCO-20^i, respectively. For data augmentation, we employ the 473×473 patches randomly cropped from the random scaled images. Each patch is randomly flipped along horizontal direction. The trade-off parameter λ is set as 10 empirically.

Baseline Approaches. We compared the following approaches: (1) CAPL [33] capitalizes on the contextual co-occurrence cues to enrich the prototypes in the classifier for novel class segmentation. (2) FT [26] directly fine-tunes the model on some sampled data from base classes plus novel examples for novel class updating. Moreover, we also report the setting of FT which additionally leverages a triplet loss regularization to augment the learnt models. (3) Following [33], we form a prototype for each class by averaging features of pixels belonging to the corresponding class in all training images and utilize prototypes to segment the query image. By doing so, we can remould the three prototype-based FSS approaches of CANet [44], PFENet [34] and PANet [36] for GFSS. The reported results are directly drawn from [33]. (4) BAM [15] is an FSS model that segments all base classes and one novel class through one inference process. To adapt for GFSS, we execute multiple inferences each of which takes the query image and support images of one novel class as the inputs. Following [15], we fuse the model outputs as the final prediction.

4.2. Performance Comparison

Quantitative Analysis. We compare with several state-of-the-art techniques on PASCAL-5^i and COCO-20^i datasets. Table 1 summarizes the mIoU performances of capitalizing on different number of shots (K-shot) on PASCAL-5^i. Overall, the results across three K-shot settings consistently indicate that our POP successfully adapts the models for novel classes and achieves the better performances than other baselines including the remoulded FSS methods and GFSS approaches. In particular, the mIoU of POP on all classes, i.e., Total, reaches 64.77%, 70.28% and 71.13% on 1-shot, 5-shot and 10-shot scenarios, respectively, making the absolute gain over the best competitors by 0.83%, 3.93%, and 2.03%. CAPL, FT and POP generally exhibit higher mIoU against the reshaped FSS approaches of CANet, PFENet and PANet. This somewhat reveals the weakness of the episodic training/testing scheme of FSS, which once only distinguishes pixels be-

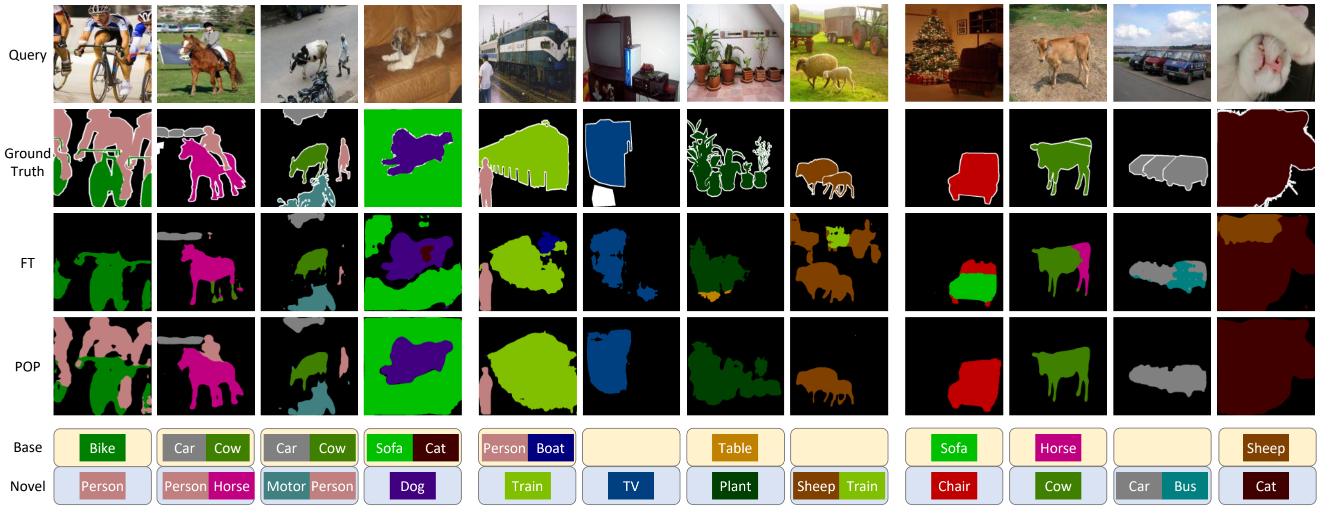

Figure 3. Twelve examples of semantic segmentation from PASCAL-5^i by POP and the state-of-the-art FT method under 5-shot setting.

longing to one specific class by abstracting the knowledge of such class from support images. As such, FSS models result in inferior performances for GFSS, where the models necessitate to identify novel classes and all possible base classes simultaneously. BAM in part alleviates the limitation of FSS scheme by additionally employing a base class segmentation branch, that is pre-trained on base data and frozen when FSS training to refine the predictions of novel classes. As such, adapting BAM to GFSS can segment the base classes well and yield good performances.

Compared to GFSS models of CAPL and FT in which the segmentation of base classes may be affected by novel class updating, POP improves the segmentation quality of base classes. The results basically indicate the advantage of freezing the well-learnt representations of base classes in model updating. On novel classes, the mIoU of POP is 35.51% in 1-shot scenario and further increased to 55.87% when knowing 5 labeled images (5-shot) for each category. As a good performer in 1-shot setting, BAM benefits from the frozen base representations and a well-tuned base-novel fusion strategy. Nevertheless, BAM still suffers from inter-class confusion and is inferior to POP, especially when having more available novel examples. The results in general verify the impact of learning orthogonal prototypes in POP. Please note that for a fair comparison, we follow CAPL [33] and also implement POP with the dataset filtration that excludes the images containing pixels of novel classes in base class learning phase. We denote such run as POP†. POP† constantly outperforms CAPL across different few-shot settings, demonstrating the effectiveness of representing an image via a group of uncorrelated features by POP.

Table 2 details the performance comparisons of 1-shot, 5-shot and 10-shot experiments on COCO-20^i dataset. Similar to the observations on PASCAL-5^i, POP surpasses the best-performing baseline FT-Triplet on 5-shot by 8.29%, 1.13% and 7.39% in terms of mIoU on base, novel and all classes, respectively. Furthermore, POP obtains better segmentation quality on base classes against other methods across different k-shot settings. The results again empirically validate our proposal.

Qualitative Analysis. Figure 3 showcases twelve semantic segmentation examples in PASCAL-5^i, by our POP and the state-of-the-art FT [26] method under 5-shot setting of GFSS. Clearly, POP attains much more promising segmentation results. Taking the query images in the first three columns on the left as the examples, POP nicely identifies the novel class “Person”, but FT fails to recognize this category. Moreover, the novel class “Dog” in the fourth example on the left is well segmented by both models and POP distinguishes the base class of “Sofa” more precisely.

4.3. Experimental Analysis

Evaluation on POP Designs. We first examine the performance contribution of different factors in POP. In between, PRO projects the features on the learnt prototypes to predict the class of each pixel and base prototypes are frozen when learning novel ones in the updating phase. ORT further encourages the orthogonality between prototypes of different classes to alleviate inter-class confusion. RES is our residual background representation that measures the residual of feature projection to categorize background pixels rather than learn a prototype for “background”.

Table 3 shows the mIoU improvement on the PASCAL-5^i dataset by considering one more factors in POP for GFSS. Note that the data enrichment strategies are not employed in the experiments here. We start from the basic run of FT [26] that capitalizes on the same features to classify the pixels of base and novel classes in the images. PRO successfully boosts up the total mIoU performance from 55.41% to 59.59% under 1-shot setting and introduces 5.53% mIoU improvement on base classes. This basically proves that PRO is a very effective and practical way for

Table 3. Ablation studies of POP on PASCAL-5^i. PRO represents segmentation via feature projection on learnt prototypes. ORT denotes the orthogonality constraint on prototypes. RES means that the background is represented by residual of feature projection.

PRO	ORT	RES	1-shot			5-shot		
			Base	Novel	Total	Base	Novel	Total
-	-	-	66.84	18.82	55.41	72.03	46.40	65.93
✓	-	-	72.37	18.71	59.59	73.39	46.99	67.10
✓	✓	-	72.82	23.65	61.11	73.76	50.69	68.27
✓	✓	✓	**73.11**	**32.29**	**63.39**	**74.32**	**53.46**	**69.36**

Table 4. Ablation studies of data enrichment strategies for novel class updating on PASCAL-5^i. RL denotes relabeling which provides pseudo labels for background pixels in images of base data. RS represents the resampling strategy that randomly selects different base data for each epoch of novel class updating.

RL	RS	1-shot			5-shot		
		Base	Novel	Total	Base	Novel	Total
-	-	73.11	32.29	63.39	74.32	53.46	69.36
✓	-	73.10	35.08	64.04	74.39	54.81	69.73
-	✓	72.86	27.74	62.12	74.31	52.12	69.03
✓	✓	**73.92**	**35.51**	**64.77**	**74.78**	**55.87**	**70.28**

GFSS to generalize the model to novel classes without sacrificing much accuracy on base classes. The performance gain of ORT is 0.45% and 4.94% in terms of mIoU on base and novel classes, respectively. The results verify the idea of learning orthogonal prototypes for feature projection. Measuring the residual of feature projection as background representation by RES further contributes a mIoU increase of 8.64% and 2.77% on novel classes in 1-shot and 5-shot settings. That empirically manifests the superiority of RES to dynamically fit semantic shifting in GFSS.

Evaluation on Data Enrichment. To better study the contribution of each design in data enrichment, we then conduct the comparisons on POP configured with different enrichment strategies. Table 4 lists the results of both 1-shot and 5-shot scenarios. In particular, relabeling the background pixels in base samples leads to over 1% mIoU improvements on novel classes. The results indicate that assigning pseudo labels to novel class pixels in base data could refine noisy background labels and ease the prototype learning for novel classes. The performance gain tends to be larger when updating the model on more base data via resampling. The performance however drops if solely using resampling. We speculate that this may be the result of involving more novel class pixels which may be incorrectly regarded as background due to semantic shifting.

Confusion Matrix Visualization. Figure 4 depicts the confusion matrices of FT and POP under 5-shot scenario on the second fold of PASCAL-5^i. We visualize the matrix after each training phase and highlight the difference in between. After the first phase of base class learning, both models achieve over 72.9% mIoU on base classes. Through the second phase of novel class updating, the mIoU on novel classes is 45.50% by FT but the performance decreases to 68.37% on base classes. In contrast, POP reaches 63.36% mIoU on novel class with only 0.13% mIoU drop on base classes. That again validates the good property of POP to identify novel classes without compromising base ones.

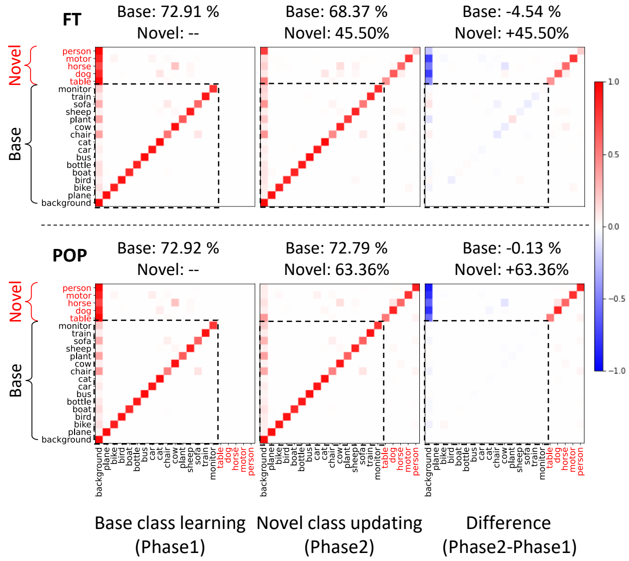

Figure 4. The difference of confusion matrices between two training phases. Compared to FT, POP obtains higher novel class performance after novel class updating without sacrificing much accuracy on base classes learnt in base class learning.

5. Conclusion

We have presented Projection onto Orthogonal Prototypes (POP) framework, which explores a principled way to generalize the model to novel classes without sacrificing much segmentation capability on base classes for GFSS. In particular, we study the problem by learning a group of prototypes and each represents a semantic class, thereby predicting each class separately with respect to the projections on its prototype. Technically, POP learns prototypes first on base classes and then updates the prototypes to support novel classes, with orthogonal constraint. Moreover, POP measures the residual of feature projection as the background representation to better mitigate semantic shifting. Experiments conducted on two datasets of PASCAL-5^i and COCO-20^i validate our proposal and analysis. Performance improvements are observed when comparing to both GFSS and remoulded FSS techniques.

Acknowledgements This work is partly supported by the National Key Research and Development Program of China (2022YFB3104700), the National Nature Science Foundation of China (62121002, 62022076, U1936210).

References

[1] 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 TPAMI*, 2017. 1, 2

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2

[3] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 1, 2

[4] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 1994. 4

[5] Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, 2018. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[7] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2009. 5

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2

[9] Bharath Hariharan, Pablo Arbeláez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[11] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *CVPR*, 2022. 2

[12] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *CVPR*, 2020. 2

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 1

[14] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 2

[15] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *CVPR*, 2022. 6

[16] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *CVPR*, 2021. 2

[17] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 2019. 2

[18] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual transformer networks for visual recognition. *IEEE TPAMI*, 2022. 1

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[20] Binghao Liu, Yao Ding, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Anti-aliasing semantic reconstruction for few-shot semantic segmentation. In *CVPR*, 2021. 2

[21] Sun-Ao Liu, Hongtao Xie, Hai Xu, Yongdong Zhang, and Qi Tian. Partial class activation attention for semantic segmentation. In *CVPR*, 2022. 2

[22] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *ECCV*, 2020. 2

[23] Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Jiebo Luo, and Tao Mei. Stand-alone inter-frame attention in video models. In *CVPR*, 2022. 1

[24] Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. Dynamic temporal filtering in video models. In *ECCV*, 2022. 1

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[26] Josh Myers-Dean, Yinan Zhao, Brian Price, Scott Cohen, and Danna Gurari. Generalized few-shot semantic segmentation: All you need is fine-tuning. *arXiv preprint arXiv:2112.10982*, 2021. 1, 2, 3, 5, 6, 7

[27] Khoi Duc Minh Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, 2019. 5

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[29] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *ICCV*, 2021. 2

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

[31] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *CVPR*, 2020. 2

[32] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 2

[33] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7

[34] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature en-

richment network for few-shot segmentation. *IEEE TPAMI*, 2020. 6

[35] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In *NeurIPS*, 2021. 1

[36] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, 2019. 1, 2, 6

[37] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020. 5

[38] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *ICCV*, 2021. 2

[39] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2

[40] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *ECCV*, 2020. 2

[41] Ting Yao, Yehao Li, Yingwei Pan, Yu Wang, Xiao-Ping Zhang, and Tao Mei. Dual vision transformer. *arXiv preprint arXiv:2207.04976*, 2022. 1

[42] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *ECCV*, 2022. 1

[43] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 2

[44] 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 *CVPR*, 2019. 1, 6

[45] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *ICCV*, 2019. 2

[46] Yiheng Zhang, Zhaofan Qiu, Jingen Liu, Ting Yao, Dong Liu, and Tao Mei. Customizable architecture search for semantic segmentation. In *CVPR*, 2019. 2

[47] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 2018. 2

[48] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Dong Liu, and Tao Mei. Transferring and regularizing prediction for semantic segmentation. In *CVPR*, 2020. 2

[49] Yiheng Zhang, Ting Yao, Zhaofan Qiu, and Tao Mei. Lightweight and progressively-scalable networks for semantic segmentation. *arXiv preprint arXiv:2207.13600*, 2022. 1

[50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 5

[51] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, 2019. 2