Generalized Few-shot Semantic Segmentation

Zhuotao Tian[1]   Xin Lai[1]   Li Jiang[2]   Shu Liu[3]   Michelle Shu[4]   Hengshuang Zhao[5,6]   Jiaya Jia[1,3]

[1]CUHK   [2]MPI Informatics   [3]SmartMore   [4]Cornell University   [5]HKU   [6]MIT

Abstract

Training semantic segmentation models requires a large amount of finely annotated data, making it hard to quickly adapt to novel classes not satisfying this condition. Few-Shot Segmentation (FS-Seg) tackles this problem with many constraints. In this paper, we introduce a new benchmark, called Generalized Few-Shot Semantic Segmentation (GFS-Seg), to analyze the generalization ability of simultaneously segmenting the novel categories with very few examples and the base categories with sufficient examples. It is the first study showing that previous representative state-of-the-art FS-Seg methods fall short in GFS-Seg and the performance discrepancy mainly comes from the constrained setting of FS-Seg. To make GFS-Seg tractable, we set up a GFS-Seg baseline that achieves decent performance without structural change on the original model. Then, since context is essential for semantic segmentation, we propose the Context-Aware Prototype Learning (CAPL) that significantly improves performance by 1) leveraging the co-occurrence prior knowledge from support samples, and 2) dynamically enriching contextual information to the classifier, conditioned on the content of each query image. Both two contributions are experimentally manifested for their substantial practical merit. Extensive experiments on Pascal-VOC and COCO also show that CAPL generalizes well to FS-Seg by achieving competitive performance. Code is available at `https://github.com/dvlab-research/GFS-Seg`.

1. Introduction

Development of deep learning has yielded significant performance gain to semantic segmentation tasks. Representative semantic segmentation methods [5, 64] benefit a wide range of applications for robotics, automatic driving, medical imaging, etc. However, once these frameworks are trained, without sufficient fully-labeled data, they are unable to deal with unseen classes in new applications. Even if the required data of novel classes are ready, fine-tuning costs additional time and resources.

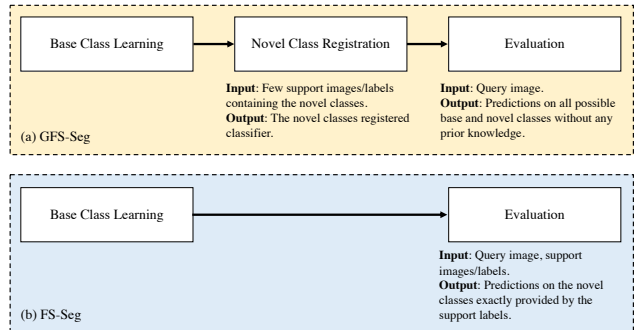In order to quickly adapt to novel classes with only lim-

Figure 1. Pipeline illustrations of GFS-Seg and FS-Seg. (a) GFS-Seg has an additional *novel class registration phase* that registers the novel information to the new classifier, therefore, in the last *evaluation phase*, GFS-Seg methods are able to make predictions on all possible base and novel classes as testing normal segmentation models without forwarding additional support samples that provide the prior knowledge of target classes. Contrarily, (b) FS-Seg models in the *evaluation phase* require support images/labels to provide the information of target classes exactly contained in each query image.

ited labeled data, few-shot segmentation (denoted as FS-Seg) [31] models are trained on well-labeled base classes and are tested on previously unseen novel classes. During training, FS-Seg divides data into the support and query sets. Samples of support set aim to provide FS-Seg models with target categorical information to identify target regions in query samples, with the purpose to mimic the situation where only a few labeled data of novel classes are available. After training, both support and query samples are sent to FS-Seg models to yield query predictions on previously unseen classes based on the support information.

Limitations of FS-Seg. However, FS-Seg requires support samples to contain classes that exist in query samples. It may be overly strong in many situations to have this prior knowledge because providing support samples in the same classes requires cumbersome manual selection. Besides, FS-Seg only evaluates the novel classes, while test samples in normal semantic segmentation may also contain the base classes. Experiments show that exemplar FS-Seg models cannot well tackle the practical situation of evaluation on both base and novel classes due to these constraints.

New benchmark and our solution. With these facts, we set up a new task, named Generalized Few-Shot Semantic Segmentation (GFS-Seg). As shown in Figure 1, a typical GFS-Seg method has three phases: 1) *Base class learning phase*, 2) *novel class registration phase* with few support samples containing novel classes, and 3) *evaluation phase* on both base and novel classes. The difference between GFS-Seg and FS-Seg is that, during evaluation, GFS-Seg does not require forwarding support samples that contain the same target classes in the test (query) samples to make predictions, because GFS-Seg should have obtained the information of base and novel classes during the *base class learning phase* and *novel class registration phase* respectively. GFS-Seg performs well on novel classes without sacrificing the accuracy of base classes when making predictions on them simultaneously without knowing what classes are contained in the query images in advance, achieving the essential step towards practical use of semantic segmentation in more challenging situations.

Inspired by [12, 26], we design a baseline for GFS-Seg with decent performance. Considering that the contextual relation is essential for semantic segmentation, we propose the Context-Aware Prototype Learning (CAPL) that provides significant performance gain to the baseline by updating the weights of base prototypes with adapted features. CAPL not only exploits essential co-occurrence information from support samples, but also adapts the model to various contexts of query images. The baseline method and the proposed CAPL can be applied to normal semantic segmentation models, *e.g.*, FCN [32], PSPNet [64] and DeepLab [5]. Also, CAPL demonstrates its effectiveness in the setting of FS-Seg by improving the baseline by a large margin, reaching state-of-the-art performance. Our overall contributions are as follows.

- We extend the classic Few-Shot Segmentation (FS-Seg) and propose a more practical setting – Generalized Few-Shot Semantic Segmentation (GFS-Seg).

- Based on our experimental results, we analyze existing performance gap between FS-Seg and GFS-Seg, and situations that recent popular FS-Seg models cannot well handle.

- We propose the Context-Aware Prototype Learning (CAPL) that yields significant performance gains to the baseline models in both settings of GFS-Seg and FS-Seg. It is applied to various normal semantic segmentation models without specific structural constraints.

2. Related Work

Semantic segmentation. Semantic segmentation is a fundamental while challenging task to accurately predict the label for each pixel. FCN [32] is the first framework designed for semantic segmentation by replacing the last fully-connected layer in a classification network with convolution layers. To get per-pixel predictions, encoder-decoder style approaches [1, 25, 29] are adopted to help refine the output in steps. The receptive field is vital for semantic segmentation – dilated convolution [5, 53] is introduced to enlarge the receptive field. Context information plays an important role for semantic segmentation and context modeling architectures are introduced, including global pooling [19] and pyramid pooling [5, 50, 63, 64]. Meanwhile, attention models [10, 16, 36, 42, 54, 59, 65] are also shown to be effective for capturing long-range relation inside scenes. Despite the success of these powerful segmentation frameworks, they cannot be easily adapted to unseen classes without fine-tuning on sufficient annotated data.

Few-shot learning. Few-shot learning aims at making prediction on novel classes with only a few labeled examples. Popular solutions include meta-learning based methods [4, 9, 30] and metric-learning ones [35, 37, 40, 56]. Data augmentation additionally helps models achieve better performance by combating overfitting. Therefore, synthesizing new training samples or features based on a few labeled data is also a feasible solution for tackling the few-shot problem [13, 44, 60]. Generalized few-shot learning was proposed in [13] where the query images can be from either base or novel categories.

Though combination of a supervised model (for base classes) and a prototype-based approach (for novel classes) was explored in low-shot visual recognition [12, 26], dense pixel labeling in semantic segmentation is different from the image-level classification where the latter does not contain contextual information for each target.

Few-shot segmentation. Few-Shot Segmentation (FS-Seg) places semantic segmentation in the few-shot scenario [2, 8, 11, 15, 18, 20, 21, 24, 27, 28, 33, 34, 38, 41, 48, 51, 52, 57, 61, 62, 66, 67], where dense pixel labeling is performed on new classes with only a few support samples. OSLSM [31] introduces this setting in segmentation and provides a solution by yielding weights of the final classifier for each query-support pair during evaluation. The idea of the prototype is used in PL [7] where predictions are based on the cosine similarity between pixels and the prototypes. Also, prototype alignment regularization is introduced in PANet [43]. Predictions can be also generated by convolutions. PFENet [39] uses the prior knowledge from a pre-trained backbone to help find the region of the interest, and the spatial inconsistency between the query and support samples is alleviated by Feature Enrichment Module (FEM).

Though FS-Seg models perform well on identifying novel classes given the corresponding support samples, as shown in Section 5, without the prior knowledge of the target classes contained in query images, even state-of-the-art

FS-Seg models cannot well tackle the practical setting involving both base and novel classes.

3. Generalized Few-shot Semantic Segmentation

Revisit the classic setting. In classic Few-Shot Segmentation (FS-Seg) [31], data is split into two sets for support S and query Q. An FS-Seg model needs to make predictions on Q based on the class information provided by S. It is trained on base classes C^b in the *base class learning phase* and is tested on previously unseen novel classes $C^n (C^b \cap C^n = \varnothing)$ in the *evaluation phase*. To deal with novel classes, the episodic paradigm was proposed in [40] to train and evaluate few-shot models. Each episode is formed by a support set S and a query set Q of the same class c. In a K-shot task, the support set S contains K samples $S = \{S_1, S_2, ..., S_K\}$ of class c. Each support sample S_i is a pair of $\{s_i, m_i\}$ where s_i and m_i are the support image and mask of class c.

For the query set, $Q = \{q, y\}$ where q is the input query image and y is the ground truth mask of class c. The input data batch used for model training is the query-support pair $\{q; S\} = \{q; s_1, m_1, ..., s_k, m_k\}$. The ground truth mask y of q is not accessible because it is to evaluate the prediction of the query image in each episode. More details for the episodic training paradigm in FS-Seg are included in [31].

In summary, there are two key criteria in the classic Few-Shot Segmentation task. (1) Samples of testing classes C^n are not seen by the model during training. (2) The model requires its support samples to contain target classes existing in query samples to make corresponding prediction.

Our generalized setting. Criterion (1) of classic few-shot segmentation evaluates model generalization ability to new classes with only a few provided samples. Criterion (2) makes this setting not practical in many cases since users may not know exactly how many and what classes are contained in each test image. So it is hard to feed the support samples that contain the same classes as the query sample to the model.

Besides, even if users have already known that there are N classes contained in the test image, FS-Seg models [3, 17, 23, 39, 55, 57, 58] may need to process NK additional manually selected support images/labels to make predictions over all possible classes for the test image. This is insufficient in real applications, where models are supposed to directly output predictions of all possible classes in the test image without processing NK additional support images/labels that contain prior knowledge of classes in the query image.

In GFS-Seg, base classes C^b have sufficient labeled training data and each novel class $C_i^n \in C^n$ only has limited K labeled samples (*e.g.*, $K = 1, 5, 10$). Similar to FS-Seg, models in GFS-Seg are first trained on base classes C^b

to learn good representations, which is the first *base class learning phase* in GFS-Seg. Then, when the first phase is accomplished, the model acquires information of N novel classes from the limited NK support samples and forms a new classifier in the *novel class registration phase*. In the last *evaluation phase*, GFS-Seg models are evaluated on images of the test set to predict labels from both base and novel classes $C^b \cup C^n$, rather than only evaluating novel classes C^n as FS-Seg.

Query images in GFS-Seg may contain either the novel classes, base classes, or both, and there is no prior knowledge of what classes are contained in the query images. Therefore, the major evaluation metric of GFS-Seg is the total mIoU that is averaged over all classes rather than the novel mIoU used in FS-Seg. The episodic training paradigm might not be a good choice for GFS-Seg, since the input data during *evaluation phase* is no longer the query-support pair $\{q; S\}$ used in FS-Seg. Instead, it is only the query image I_Q as testing common semantic segmentation models.

To better distinguish between FS-Seg and GFS-Seg, we illustrate a 2-way K-shot task of FS-Seg and a case of GFS-Seg with the same query image in Figure 2, where *Cow* and *Motorbike* are novel classes, and *Person* and *Car* are base classes. FS-Seg model (Figure 2(a)) is limited to predicting binary segmentation masks only for the classes that are included in the support set. *Person* on the right and *Car* at the top are missing in the prediction because the support set does not provide information for these classes, even if the model is trained on these base classes for sufficient steps.

In addition, if redundant novel classes that do not appear in the query image (*e.g.*, *Aeroplane*) are provided by support set of (a), they may adversely affect performance because FS-Seg has a prerequisite that the query images must contain the classes provided by support samples.

As shown in Section 5.1, FS-Seg models only learn to predict the foreground masks for the given novel classes. Thus, their performance is severely degraded in our generalized setting of GFS-Seg where all possible base and novel classes require predictions. Differently, GFS-Seg (Figure 2(b)) identifies base and novel classes simultaneously without the prior knowledge of classes contained in query image. Extra support classes (*e.g.*, *Aeroplane* at the top-left of Figure 2(b)) should not affect the model much.

4. Our Method

4.1. Prototype Learning

Prototype learning in FS-Seg. In Few-Shot Segmentation (FS-Seg) frameworks [7, 43] of a N-way K-shot FS-Seg task (with N novel classes, each with K support samples), all support samples s_j^i ($i \in \{1, 2, ..., N\}, j \in \{1, 2, ..., K\}$) are first processed by a feature extractor $\mathcal{F}$ and mask average pooling. They are then averaged over K
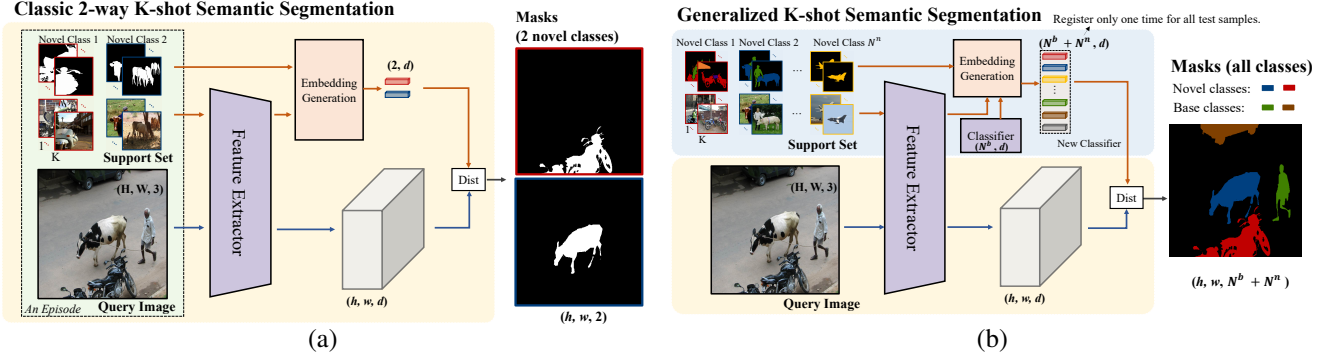
Figure 2. Illustrations of (a) classic Few-Shot Segmentation (FS-Seg) and (b) Generalized Few-Shot Semantic Segmentation (GFS-Seg). 'Dist' can be any methods measuring the distance/similarity between each feature and prototype and making predictions based on that distance/similarity. FS-Seg models only predict novel classes provided by the support set, while GFS-Seg models make predictions on base and novel classes simultaneously without being affected by redundant classes. Also, during evaluation, GFS-Seg models do not require the prior knowledge of what target classes exist in the query images by registering the novel classes to form a new classifier once for all test images (blue area in (b) representing the novel class registration phase).

shots to form N prototypes $\boldsymbol{p}^i$ ($i \in \{1, 2, ..., N\}$) as

$$\boldsymbol{p}^i = \frac{1}{K} * \sum_{j=1}^{K} \frac{\sum_{h,w}[\boldsymbol{m}_j^i \circ \mathcal{F}(\boldsymbol{s}_j^i)]_{h,w}}{\sum_{h,w}[\boldsymbol{m}_j^i]_{h,w}}, \quad i \in \{1, 2, ..., N\}, \quad (1)$$

where $\boldsymbol{m}_j^i \in \mathbb{R}^{h,w,1}$ is the class mask for class c^i on $\mathcal{F}(\boldsymbol{s}_j^i) \in \mathbb{R}^{h,w,d}$. $\boldsymbol{s}_j^i$ represents the j-th support image of class c^i, and $\circ$ is the Hadamard Product. After acquiring N prototypes, for query features $\mathcal{F}(\boldsymbol{q})$, the predictions are assigned by the class labels of the most similar prototypes.

Baseline for GFS-Seg. FS-Seg only requires identifying targets from novel classes. But our generalized setting in GFS-Seg requires predictions on both base and novel classes. It is hard and also inefficient for the FS-Seg model to form prototypes for base classes via Eq. (1) by forwarding all samples of base classes to the feature extractor, especially when the training set is large.

Common semantic segmentation frameworks can be decomposed into feature extractor and classifier parts. Feature extractor projects input image into d-dimensional latent space and the classifier of size $N^b \times d$ makes predictions on N^b base classes. The classifier of size $N^b \times d$ can be seen as N^b base prototypes ($\boldsymbol{P}^b \in \mathbb{R}^{N^b,d}$). Since forwarding all base samples to form the base prototypes is impractical, inspired by low-shot learning [12, 26], we learn the classifier via back-propagation during training on base classes.

Specifically, our baseline for GFS-Seg is trained on base classes as normal segmentation frameworks. After training, prototypes $\boldsymbol{P}^n \in \mathbb{R}^{N^n,d}$ for N^n novel classes are formed in Eq. (1) with $N^n \times K$ support samples. $\boldsymbol{P}^n$ and $\boldsymbol{P}^b$ are concatenated to form $\boldsymbol{P}^{all} \in \mathbb{R}^{N^b+N^n,d}$, the new classifier, to simultaneously predict base and novel classes.

Since the dot product used by classifier of common semantic segmentation frameworks produces different norm scales of $\mathcal{F}(\boldsymbol{s}_j^i)$ that negatively impact the average operation in Eq. (1), following [43], we adopt cosine similarity

as the distance metric ϕ to yield output $\boldsymbol{O}$ for pixels in query sample $\boldsymbol{q} \in \mathbb{R}^{h,w,3}$ as

$$\boldsymbol{O}_{x,y} = \arg\max_i \frac{\exp(\alpha\phi(\mathcal{F}(\boldsymbol{q}_{x,y}), \boldsymbol{p}^i))}{\sum_{\boldsymbol{p}^i \in \boldsymbol{P}^{all}} \exp(\alpha\phi(\mathcal{F}(\boldsymbol{q}_{x,y}), \boldsymbol{p}^i))}, \quad (2)$$

where $x \in \{1, ..., h\}, y \in \{1, ..., w\}, i \in \{1, ..., N^b+N^n\}$, and α is set to 10 in all experiments.

4.2. Context-Aware Prototype Learning (CAPL)

Motivation. Prototype Learning (PL) is applicable to few-shot classification and FS-Seg. But it is inferior to GFS-Seg in terms of performance. In the setting of FS-Seg, target labels of query samples are only from novel classes. Thus there is no essential co-occurrence interaction between novel and base classes that can be utilized for further improvement. However, in GFS-Seg, there is no such limitation on the classes contained in each test image – predictions on all possible base and novel classes are required.

Contextual cue always plays an important role in semantic segmentation (*e.g.*, PSPNet [64] and Deeplab [6]), especially in the proposed setting of GFS-Seg. For example, *Dog* and *People* are base classes. The learned base prototypes only capture the contextual relation between *Dog* and *People* during training. If *Sofa* is a novel class and some instances of *Sofa* in support samples appear with *Dog* (*e.g.*, a dog on sofa), merely mask-pooling in each support sample of *Sofa* to form the novel prototype may result in the base prototype of *Dog* losing the contextual co-occurrence information with *Sofa*. It hence yields inferior results.

To remedy it, for GFS-Seg, reasonable utilization of contextual information is the key to better performance, and our proposed method enables the model to meta-learn the behavior that the contextual information can be enriched by a simple adaptive fusion strategy without significantly altering model structure. The contextual cues can be mined from
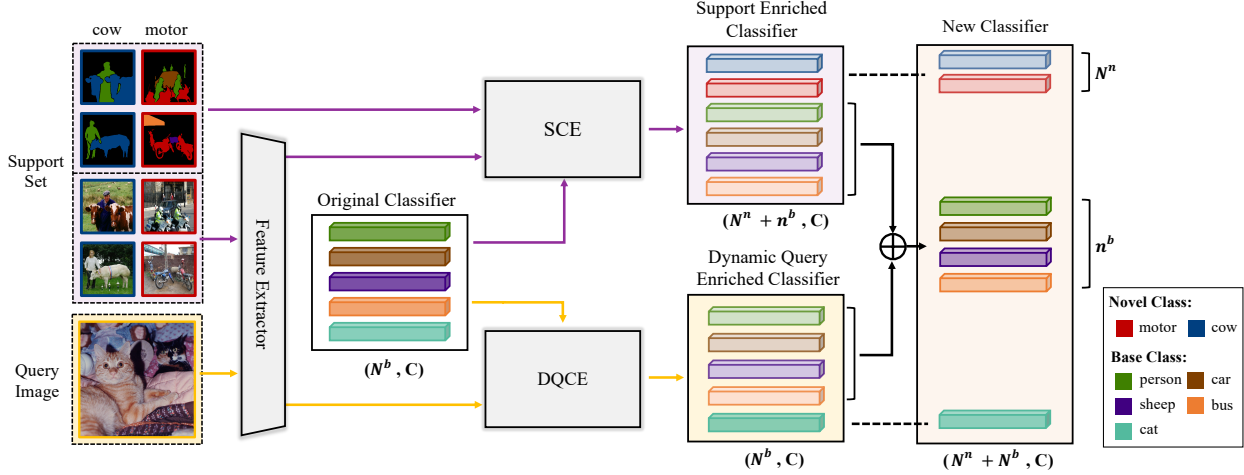
Figure 3. Visual illustration of CAPL. The weights of N^n novel classes (*e.g. motor* and *cow*) are directly set by the averaged novel features. Also, the weights of n^b base classes (*e.g. person*, *car*, *sheep* and *bus*) that appear in support samples are enriched by SCE with the original weights. DQCE dynamically enriches the weights of N^b base classes in the classifier with temporary contextual hints extracted from query samples. The new classifier takes the merits from both SCE and DQCE.

both support and query samples. Therefore, we propose the Context-Aware Prototype Learning (CAPL) to tackle GFS-Seg by effectively enriching the classifier with contextual information. To facilitate understanding, illustration of CAPL is shown in Figure 3.

Support Contextual Enrichment (SCE). SCE mines support context in the *novel class registration phase*. Let N^b and N^n denote the numbers of base classes $\boldsymbol{C}^b$ and novel classes $\boldsymbol{C}^n$ respectively. We use $n^b(n^b \leqslant N^b)$ to denote the number of base classes $c^{b,i} \in \boldsymbol{C}^b(i \in \{1,...,n^b\})$ contained in $N^n \times K$ support samples for N^n unseen novel classes $c^{n,u} \in \boldsymbol{C}^n(u \in \{1,...,N^n\})$. Before evaluation, the novel prototypes are formed as Eq. (1). The updated prototype $\boldsymbol{p}^{b,i}$ for base class $c^{b,i}$ is the weighted sum of the weights of original classifier $\boldsymbol{p}_{cls}^{b,i}$ and the new base prototypes $\boldsymbol{p}_{sup}^{b,i}$ generated from support features as

$$\boldsymbol{p}_{sup}^{b,i} = \frac{\sum_{u=1}^{N^n} \sum_{j=1}^{K} \sum_{h,w}[\boldsymbol{m}_j^{i,u} \circ \mathcal{F}(\boldsymbol{s}_j^u)]_{h,w}}{\sum_{u=1}^{N^n} \sum_{j=1}^{K} \sum_{h,w}[\boldsymbol{m}_j^{i,u}]_{h,w}} \quad (3)$$

where $\quad i \in \{1,...,n^b\}, u \in \{1,...,N^n\}$.

Now $\boldsymbol{p}^{b,i}$ is expressed as

$$\boldsymbol{p}^{b,i} = \boldsymbol{\gamma}_{sup}^i * \boldsymbol{p}_{cls}^{b,i} + (1 - \boldsymbol{\gamma}_{sup}^i) * \boldsymbol{p}_{sup}^{b,i}, \quad i \in \{1,...,n^b\}, \quad (4)$$

where $\boldsymbol{m}_j^{i,u}$ represents the binary mask for base class $c^{b,i}$ on $\mathcal{F}(\boldsymbol{s}_j^u)$, and $\boldsymbol{s}_j^u$ is the j-th support sample of novel class $c^{n,u}$. The adaptive weight $\boldsymbol{\gamma}_{sup}^i$ balances the impacts of the old and new base prototypes.

We note that $\boldsymbol{\gamma}_{sup}^i$ is data-dependent and its value is conditioned on each pair of old and new prototypes. For base class $c^{b,i}$, $\boldsymbol{\gamma}_{sup}^i$ is calculated as $\boldsymbol{\gamma}_{sup}^i = \mathcal{G}_{sup}(\boldsymbol{p}_{cls}^{b,i}, \boldsymbol{p}_{sup}^{b,i})$ where $\mathcal{G}_{sup}$ serves as a correlation estimator that produces the weighing factor $\boldsymbol{\gamma}_{sup}^i$.

Dynamic Query Contextual Enrichment (DQCE). The aforementioned SCE only takes place in the *novel class registration phase* and it exploits the hints from the support samples to offer the co-occurrence prior knowledge. However, during the *evaluation phase*, the new classifier is shared by all query images. Hence the introduced prior might be biased towards the content of the limited support samples, causing inferior generalization ability to different query images. To alleviate this issue, we propose the Dynamic Query Contextual Enrichment (DQCE) that adapts the new classifier to different context by dynamically incorporating essential semantic information mined from individual query samples.

Specifically, we first let the original classifier to yield a temporary prediction $\boldsymbol{y}_{qry} \in \mathbb{R}^{h_q w_q \times N^b}$ on the query feature $\mathcal{F}(\boldsymbol{q}) \in \mathbb{R}^{h_q w_q \times d}$, where h_q, w_q, N^b and d are height, width, base class number and dimension number respectively. Then, the categorical representatives $\boldsymbol{p}_{qry}^b \in \mathbb{R}^{N^b \times d}$ of query sample $\boldsymbol{Q}$ are yielded as

$$\boldsymbol{p}_{qry}^b = Softmax(\boldsymbol{y}_{qry}^t) \times \mathcal{F}(\boldsymbol{q}), \quad (5)$$

where the Softmax operation is performed on the second dimension of the transposed $\boldsymbol{y}_{cls}$ ($\boldsymbol{y}_{cls}^t \in \mathbb{R}^{N^b \times h_q w_q}$), so as to weigh the elements on the query feature map $\mathcal{F}(\boldsymbol{q})$.

Since the query label is not available, the dynamic prototypes $\boldsymbol{p}_{qry}^b$ yielded via query predictions are not as sound as $\boldsymbol{p}_{sup}^b$ and might introduce unnecessary noise. We still need reliability measure $\boldsymbol{\gamma}_{qry}^i$ for the estimated $\boldsymbol{p}_{qry}^b$. Similar to Eq. (4), a weighing factor $\boldsymbol{\gamma}_{qry}^i = \mathcal{G}_{qry}(\boldsymbol{p}_{cls}^{b,i}, \boldsymbol{p}_{qry}^{b,i})$ is required to accomplish the dynamic context enrichment for each base class. We obtain the dynamically enriched prototypes as

$$\boldsymbol{p}_{dyn}^{b,i} = \boldsymbol{\gamma}_{qry}^i * \boldsymbol{p}_{cls}^{b,i} + (1 - \boldsymbol{\gamma}_{qry}^i) * \boldsymbol{p}_{qry}^{b,i}, \quad i \in \{1,...,N^b\}. \quad (6)$$

It is worth noting that, different from Eq. (4) where only n^b classes contained in the support set are enriched, Eq. (6) considers all N^b base classes since there is no prior knowledge of what classes are contained in the current query image. Besides, we empirically find that cosine similarity is better for $\mathcal{G}_{qry}$ and MLP is more suitable for $\mathcal{G}_{sup}$. Ablations on other alternatives are in the supplementary file. Now,

$$\boldsymbol{p}_{capl}^{b,i} = \boldsymbol{p}^{b,i} + \boldsymbol{p}_{dyn}^{b,i}, \quad i \in \{1,...,N^b\}. \quad (7)$$

$\boldsymbol{P}_{capl}^b$ is then concatenated with the prototypes $\boldsymbol{P}^n$ of novel classes to form $\boldsymbol{P}_{capl}^{all} \in \mathbb{R}^{N^b+N^n,d}$. It predicts all possible base and novel classes during *evaluation*.

Training. Directly applying the weighted sum in Eq. (4) and Eqs. (6)-(7) is difficult because the averaged support/query features and classifier's weights are in different feature spaces. To make it tractable, we accordingly modify the training scheme to let the feature extractor $\mathcal{F}$ learn to yield features compatible with $\boldsymbol{P}_{cls}^b$.

For prior knowledge from support samples, we let B denote the training batch size and N^b denote the number of all base classes. Our work randomly selects $\lfloor \frac{B}{2} \rfloor$ training samples as the 'Fake Support' (FS) samples and the rest as 'Fake Query' ones. Let N_f^b denote the number of base classes contained in FS samples. We randomly select $\lfloor \frac{N_f^b}{2} \rfloor$ as the 'Fake Novel' (FN) classes $\boldsymbol{C}^{FN}$ and set the rest $N_f^b - \lfloor \frac{N_f^b}{2} \rfloor$ as the 'Fake Context' (FC) classes $\boldsymbol{C}^{FC}$. The selected $\boldsymbol{C}^{FN}$ and $\boldsymbol{C}^{FC}$ classes are both base classes. But they mimic the behaviors of real novel and base classes contained in the support set respectively. Formation of the updated prototypes $\boldsymbol{p}^{b,i}$ ($i \in \{1,...,N_f^b,...,N^b\}$) during training is

$$\boldsymbol{p}^{b,i} = \begin{cases} \boldsymbol{\gamma}_{sup}^i * \boldsymbol{p}_{cls}^{b,i} + (1-\boldsymbol{\gamma}_{sup}^i) * \boldsymbol{p}_{sup}^{b,i} & c^i \in \boldsymbol{C}^{FC} \\ \boldsymbol{p}_{sup}^{b,i} & c^i \in \boldsymbol{C}^{FN} \quad (8) \\ \boldsymbol{p}_{cls}^{b,i} & \text{Otherwise} \end{cases}$$

$$\boldsymbol{p}_{sup}^{b,i} = \frac{\sum_{j=1}^{\lfloor \frac{B}{2} \rfloor} \sum_{h,w} [\boldsymbol{m}_j^i \circ \mathcal{F}(\boldsymbol{s}_j^i)]_{h,w}}{\sum_{j=1}^{\lfloor \frac{B}{2} \rfloor} \sum_{h,w} [\boldsymbol{m}_j^i]_{h,w}}. \quad c^i \in \boldsymbol{C}^{FC} \cup \boldsymbol{C}^{FN} \quad (9)$$

Specifically, FC classes $\boldsymbol{C}^{FC}$ act as the base classes contained in support samples of novel classes during testing. The features of FC classes update $\boldsymbol{P}_{cls}^b$ with $\boldsymbol{\gamma}_{sup}$ as shown in Eq. (8). Also, features of FN classes $\boldsymbol{C}^{FN}$ yield FN prototypes via Eq. (9) that directly replace the learned weights of the corresponding base classes in $\boldsymbol{P}_{cls}^b$. Weights of the rest classes for this training batch are kept as the original ones in the classifier.

To make the classifier adapt to different contexts of individual query samples, the dynamic prototypes $\boldsymbol{p}_{dyn}^{b,i}(i \in \{1,...,N^b\})$ are generated by following Eqs. (5) and (6). Finally, the context-aware classifier $\boldsymbol{P}_{capl}^b$ is formed as Eq. (7), and the overall training objective is to minimize the standard cross-entropy loss calculated on the predictions made by $\boldsymbol{P}_{capl}^b$.

5. Experiments

This section presents the experimental results. Implementation details are provided in the supplementary file.

5.1. Comparison with FS-Seg Models in GFS-Seg

To show that models in the setting of FS-Seg are unable to perform well when predicting both base and novel classes, we evaluate four recently proposed exemplar FS-Seg frameworks in the setting of GFS-Seg. They are CANet [58], PFENet [39] SCL [55] and PANet [43]).

The major difference of these frameworks is on the 'Dist' shown in Figure 2(a) related to the way to process query and support feature and make predictions. SCL, PFENet, and CANet are relation-based models and PANet is cosine-based. Specifically, CANet uses convolutions as a relation module [37] to process the concatenated query and support prototypes. It yields prediction on the query image. PANet generates results by measuring the cosine similarity between every query feature and prototypes obtained from support samples.

Though FS-Seg only requires evaluation of novel classes, if the prototypes, i.e., averaged features, of base classes are available, the FS-Seg models should also be able to identify regions of them in query images. It is noted that some recent frameworks, *e.g.*, RePRI [3] and HSNet [23], are not practical to be applied to the setting of GFS-Seg because their inference process requires independent reasoning with the full support feature maps, instead of the prototypical feature vectors adopted by above methods where the prior mask cannot be used in PFENet. Thus, under an n-way k-shot setting, for each query image, these methods need to perform entire feature reasoning for additional nk times on the full support feature maps, causing low efficiency or even Out of Memory issues.

We use the publicly available codes and follow the default training configuration. We modify the inference code to feed all prototypes (base and novel) for each query image. The base prototypes are formed by averaging features belonging to base classes in all training samples. As shown in Table 1, CANet, SCL, PFENet, and PANet perform less satisfyingly than those implemented with CAPL. 'PANet + CAPL' differs from PANet only in the training and classifier formation strategies.

We note that the results of novel mIoU in Table 1 are worse than the GFS-Seg setting. The numbers are lower than those reported in the original papers of FS-Seg models. This discrepancy is caused by different settings. In GFS-Seg, models are required to identify all classes in a given testing image, including both base and novel classes. Note, in FS-Seg, models only need to find pixels belonging to one specific novel class with support samples that provide the prior knowledge of what the target class is. Therefore, it is much harder to identify the novel classes under the in-

Methods	1-shot			5-shot		
	Base	Novel	Total	Base	Novel	Total
CANet [58]	8.73	2.42	7.23	9.05	1.52	7.26
PFENet [39]	8.32	2.67	6.97	8.83	1.89	7.18
SCL [58]	8.88	2.44	7.35	9.11	1.83	7.38
PANet [43]	31.88	11.25	26.97	32.95	15.25	28.74
PANet + CAPL	63.06	14.96	51.60	63.81	19.66	53.30
DeepLab-V3 + CAPL	65.71	15.05	53.77	67.01	23.26	56.59
PSPNet + CAPL	65.48	18.85	54.38	66.14	22.41	55.72

Table 1. Comparisons with FS-Seg models in GFS-Seg where the base and novel classes are simultaneously identified. All models are based on ResNet-50. **Base**: mIoU results of all base classes. **Novel**: mIoU results of all novel classes. **Total**: mIoU results of all (base + novel) classes.

Methods	MLP	Cos	1-shot			5-shot		
			Base	Novel	Total	Base	Novel	Total
			Pascal-5^i					
Baseline	N/A	N/A	60.47	14.55	49.54	61.88	16.68	51.12
DQCE	✓	-	63.25	15.42	51.82	64.12	20.37	53.70
DQCE	-	✓	64.16	15.39	52.55	65.26	21.32	54.80
DQCE-Sw	-	✓	58.28	15.96	48.20	60.68	20.66	51.15
SCE	✓	-	62.17	17.88	51.63	63.62	20.50	53.35
SCE	-	✓	59.87	16.80	49.62	61.60	19.92	51.68
SCE-Sw	-	✓	60.92	16.39	50.32	62.75	21.18	52.83
CAPL	N/A	N/A	65.48	18.85	54.38	66.14	22.41	55.72
			COCO-20^i					
Baseline	N/A	N/A	36.68	5.84	29.06	36.91	7.26	29.59
CAPL	N/A	N/A	44.61	7.05	35.46	45.24	11.05	36.80

Table 2. Comparison of contextual enrichment strategies. 'MLP' and 'Cos' mean two-layers MLPs and cosine similarity to yield weighing factors γ_{qry} and γ_{sup} respectively. SCE mines co-occurrence cues from support data (Eqs. (3)-(4), and 'DQCE' extracts temporary query context (Eqs. (5)-(6)). 'CAPL' combines 'SCE (MLP)' and 'DQCE (Cos)' in Eq. (7).

terference of the base classes in GFS-Seg.

In particular, FS-Seg models fall short mainly because the episodic training/testing schemes of FS-Seg only focus on making models discriminative between background and foreground, where the decision boundary for each episode only lies between one target class and the background in each query sample. Also, FS-Seg requires query images to contain the classes provided by support samples, while GFS-Seg distinguishes between not only multiple novel classes but also all possible base classes simultaneously, even without the prior knowledge of classes in query samples. Moreover, to maintain the high generalization ability on unseen novel classes in FS-Seg, both CANet, SCL and PFENet fix their backbones during training, causing limited adaptation in the complex scenario of GFS-Seg that requires multi-class labeling.

5.2. Ablation Study

In this section, we investigate effectiveness of components of CAPL with PSPNet on Pascal-5^i in GFS-Seg where the base and novel classes are required to be simultaneously identified. The baselines for following ablation study are

Methods	Train	Test	1-shot			5-shot		
			Base	Novel	Total	Base	Novel	Total
Baseline	N/A	N/A	60.47	14.55	49.54	61.88	16.68	51.12
Baseline+	N/A	N/A	60.60	16.53	50.10	62.28	19.39	52.07
CAPL-Tr	✓	-	59.73	17.40	49.65	61.34	21.72	51.91
CAPL-Te	-	✓	60.89	7.00	48.06	61.13	10.90	49.17
CAPL	✓	✓	**65.48**	**18.85**	**54.38**	**66.14**	**22.41**	**55.72**

Table 3. Comparison of training & testing strategies of CAPL. 'Baseline+' only replaces fake novel prototypes during training. 'CAPL-Tr' adopts CAPL's training strategy while keeping the novel class registration and evaluation phases same as the baseline. 'CAPL-Te' performs CAPL during novel class registration and evaluation – its training scheme is not altered.

based on PSPNet [64] with ResNet-50 [14].

Design options for components of CAPL. The effects of the support contextual enrichment method (SCE) as well as its dynamic counterpart (DQCE) are investigated in Table 2. SCE enriches the classifier with the essential co-occurrence relations between novel and base classes (Eqs. (3)-(4)), while DQCE further adapts the enrichment process to content of individual query images (Eqs. (5)-(7)). Thus, $\mathcal{G}_{sup}$ and $\mathcal{G}_{qry}$ are used by SCE and DQCE to yield γ_{sup} and γ_{qry} respectively.

It can be observed in Table 2 that both SCE and DQCE are most conducive to the baseline when $\mathcal{G}_{qry}$='Cos' and $\mathcal{G}_{sup}$='MLP'. Cosine similarity works better for DQCE because it serves as the reliability estimator to weigh the prototypes extracted from the query features since they may introduce irrelevant information. MLP is more likely to give high confidence to the original classifier to avoid the risk brought by Eq. (5), yielding suboptimal results. Also, switching the weighing factors for $p_{cls}^{b,i}$ and $p_{qry}^{b,i}$ (DQCE-Sw) undoubtedly worsens the performance because the new prototype is dominated by untrustworthy $p_{qry}^{b,i}$.

Nonetheless, cosine similarity is less effective than MLP for SCE, and 'Cos' achieves comparable results to the baseline. This could be caused by the fact that $p_{sup}^{b,i}$ is produced by ground-truth masks, hence $p_{sup}^{b,i}$ is a reliable categorical representation, leading to a high similarity with $p_{cls}^{b,i}$ that directly suppresses context brought by $p_{sup}^{b,i}$ in Eq. (4). If we switch the weighing factors for $p_{cls}^{b,i}$ and $p_{sup}^{b,i}$ in Eq. (4) ('SCE-Sw' in Table 2), the results rely more on the content of $p_{sup}^{b,i}$, losing the generalization power of $p_{cls}^{b,i}$.

Differently, we observe that values of γ_{sup} yielded by MLP are within a certain range (0.4-0.7) without corrupting to 0 or 1. They are generally proportional to the discrepancy between $p_{cls}^{b,i}$ and $p_{sup}^{b,i}$ of base class i. Thus, SCE (MLP) outweighs SCE-Sw (Cos) by producing moderate data-conditioned values to adequately balance the novel contextual and original information.

Training and contextual enrichment strategies. Since the training of CAPL picks 'Fake Novel' and 'Fake Context' samples to make the averaged features compatible with

Methods	Venue	Backbone	Pascal-5^i		COCO-20^i	
			1-Shot	5-Shot	1-Shot	5-Shot
PANet [43]	ICCV-19	Res-50	48.1	55.7	20.9	29.7
PFENet [39]	TPAMI-20	Res-50	60.8	61.9	32.1	37.5
ASGNet [17]	CVPR-21	Res-50	59.3	63.9	34.5	42.5
SCL [55]	CVPR-21	Res-50	61.8	62.9	-	-
SAGNN [46]	CVPR-21	Res-50	62.1	62.8	-	-
RePri [3]	CVPR-21	Res-50	59.1	66.8	34.0	42.1
CWT [22]	ICCV-21	Res-50	56.4	63.7	32.9	41.3
MMNet [45]	ICCV-21	Res-50	61.8	63.4	37.5	38.2
CMN [47]	ICCV-21	Res-50	62.8	63.7	39.3	43.1
Mining [49]	ICCV-21	Res-50	62.1	66.1	33.9	40.6
HSNet [23]	ICCV-21	Res-50	**64.0**	**69.5**	39.2	46.9
CAPL (PANet)		Res-50	60.6	66.1	38.0	47.3
CAPL (PFENet)		Res-50	62.2	67.1	**39.8**	**48.3**
PFENet [39]	TPAMI-20	Res-101	60.1	61.4	32.4	37.4
SAGNN [46]	CVPR-21	Res-101	-	-	37.2	42.7
ASGNet [17]	CVPR-21	Res-101	59.3	64.4	-	-
CWT [22]	ICCV-21	Res-101	58.0	64.7	32.4	42.0
Mining [49]	ICCV-21	Res-101	62.6	68.8	36.4	44.4
HSNet [23]	ICCV-21	Res-101	**66.2**	**70.4**	41.2	49.5
CAPL (PFENet)		Res-101	63.6	68.9	**42.8**	**50.4**

Table 4. Class mIoU results on Pascal-5^i and COCO-20^i in FS-Seg where only the novel classes are required to be identified.

learned weights for building the updated classifier, we modify the training scheme of the baseline (denoted as 'Baseline+') accordingly for fair comparison. Because the baseline only replaces the novel prototypes during the evaluation, only 'Fake Novel' is sampled and used to replace the base prototypes during training Baseline+. Therefore, Baseline+ is analogous to the methods in few-shot classification [12, 26]. Besides, 'CAPL-Tr' denotes that CAPL is only performed during the training phase, keeping the novel class registration and evaluation same as the baseline. 'CAPL-Te' represents that CAPL is only performed during *novel class registration* and *evaluation phases*. There is no change of baseline training. Since γ_{sup} is yielded by a trainable two-layers MLP, we set γ_{sup} to the mean converged values of CAPL for CAPL-Te.

The results in Table 3 show that the training and contextual enrichment strategies of CAPL complement each other – both are indispensable. CAPL-Tr brings minor improvement to the baseline with only training alignment implemented. CAPL-Te proves that, without the proposed training strategy, the contextual enrichment method of CAPL produces sub-optimal results due to misalignment between the original classifier and the features.

5.3. Apply CAPL to FS-Seg

FS-Seg is an extreme case of GFS-Seg. To validate the proposed CAPL in the setting of FS-Seg, in Table 4, we incorporate CAPL to PANet and PFENet. CAPL achieves significant improvement to the baselines. Specifically, CAPL only alters the prototype construction process of PANet by making it dynamically adapt to different query and support pairs, following the method introduced in Section 4.

We also incorporate CAPL to PFENet whose decoder processes concatenation of the prior mask and middle-level

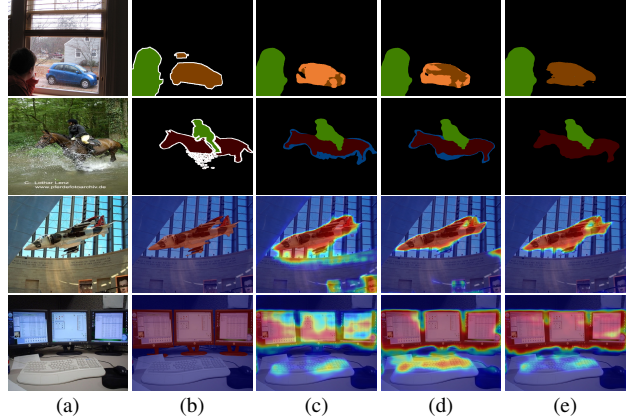(a)   (b)   (c)   (d)   (e)

Figure 4. Visualizations of GFS-Seg (top) and FS-Seg (bottom). (a): Input; (b): GT; (c): Baseline; (d): SCE; (e): SCE+DQCE.

features to make predictions. However, the prior mask is enhanced by CAPL. Since there is no significant structural change on both PANet and PFENet, the improved models have nearly the same inference speed as the original ones. More implementation details regarding this section are included in the supplementary material.

5.4. Visual Examples

Visual comparison is presented in Fig 4 where SCE and DQCE refine the baseline predictions. More examples are shown in the supplementary file.

6. Concluding Remarks

Summary. We have presented the new benchmark of Generalized Few-Shot Semantic Segmentation (GFS-Seg) with a novel solution – Context-Aware Prototype Learning (CAPL). Different from the classic Few-Shot Segmentation (FS-Seg), GFS-Seg aims at identifying both base and novel classes that FS-Seg models fall short. Our proposed CAPL achieves significant performance improvement by dynamically enriching the context information with adapted features. CAPL has no structural constraints on the base model and thus it can be easily applied to normal semantic segmentation frameworks. It also generalizes well to FS-Seg.

Limitation & Discussion with HSNet. CAPL dynamically leverages the contextual hints in both GFS-Seg and FS-Seg, while it does not introduce new designs for dense spatial reasoning between query and support features in FS-Seg. Thus, though new SOTA performance has been achieved on the challenging COCO-20^i (20 novel classes) where the semantic cues are better exploited, CAPL does not outperform another advanced method that adopts the hyper-correlations between query and support features, *i.e.*, HSNet [23], on Pascal-5^i (5 novel classes) in FS-Seg.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 2017. 2

[2] Ayan Kumar Bhunia, Ankan Kumar Bhunia, Shuvozit Ghose, Abhirup Das, Partha Pratim Roy, and Umapada Pal. A deep one-shot network for query-based logo retrieval. *PR*, 2019. 2

[3] Malik Boudiaf, Hoel Kervadec, Imtiaz Masud Ziko, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *CVPR*, 2021. 3, 6, 8

[4] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *CVPR*, 2018. 2

[5] 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. *TPAMI*, 2018. 1, 2

[6] 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. 4

[7] Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, 2018. 2, 3

[8] Zhibo Fan, Jin-Gang Yu, Zhihao Liang, Jiarong Ou, Changxin Gao, Gui-Song Xia, and Yuanqing Li. FGN: fully guided network for few-shot instance segmentation. In *CVPR*, 2020. 2

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 2

[10] 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

[11] Siddhartha Gairola, Mayur Hemani, Ayush Chopra, and Balaji Krishnamurthy. Simpropnet: Improved similarity propagation for few-shot image segmentation. In *IJCAI*, 2020. 2

[12] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018. 2, 4, 8

[13] Bharath Hariharan and Ross B. Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, 2017. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[15] Tao Hu, Pengwan Yang, Chiliang Zhang, Gang Yu, Yadong Mu, and Cees G. M. Snoek. Attention-based multi-context guiding for few-shot semantic segmentation. In *AAAI*, 2019. 2

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

[17] 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. 3, 8

[18] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. FSS-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, 2020. 2

[19] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *arXiv*, 2015. 2

[20] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Cr-net: Cross-reference networks for few-shot segmentation. In *CVPR*, 2020. 2

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

[22] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*, 2021. 8

[23] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *ICCV*, 2021. 3, 6, 8

[24] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *ICCV*, 2019. 2

[25] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015. 2

[26] Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *CVPR*, 2018. 2, 4, 8

[27] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha A. Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. In *ICLR Workshop*, 2018. 2

[28] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A. Efros, and Sergey Levine. Few-shot segmentation propagation with guided networks. *arXiv*, 2018. 2

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[30] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019. 2

[31] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017. 1, 2, 3

[32] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2017. 2

[33] Mennatullah Siam and Boris N. Oreshkin. Adaptive masked weight imprinting for few-shot segmentation. In *ICCV*, 2019. 2

[34] Mennatullah Siam, Boris N. Oreshkin, and Martin Jägersand. AMP: adaptive masked proxies for few-shot segmentation. In *ICCV*, 2019. 2

[35] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 2

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to com-

pare: Relation network for few-shot learning. In *CVPR*, 2018. 2, 6

[38] Pinzhuo Tian, Zhangkai Wu, Lei Qi, Lei Wang, Yinghuan Shi, and Yang Gao. Differentiable meta-learning model for few-shot semantic segmentation. In *AAAI*, 2020. 2

[39] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *TPAMI*, 2020. 2, 3, 6, 7, 8

[40] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, 2016. 2, 3

[41] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *ECCV*, 2020. 2

[42] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 2

[43] Kaixin Wang, JunHao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*, 2019. 2, 3, 4, 6, 7, 8

[44] Yu-Xiong Wang, Ross B. Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018. 2

[45] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *ICCV*, 2021. 8

[46] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *CVPR*, 2021. 8

[47] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *ICCV*, 2021. 8

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

[49] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. In *ICCV*, 2021. 8

[50] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 2

[51] Yuwei Yang, Fanman Meng, Hongliang Li, King N. Ngan, and Qingbo Wu. A new few-shot segmentation network based on class representation. In *VCIP*, 2019. 2

[52] Yuwei Yang, Fanman Meng, Hongliang Li, Qingbo Wu, Xiaolong Xu, and Shuai Chen. A new local transformation module for few-shot segmentation. In *MMM*, 2020. 2

[53] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2016. 2

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

[55] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. In *CVPR*, 2021. 3, 6, 8

[56] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[57] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *ICCV*, 2019. 2, 3

[58] 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. 3, 6, 7

[59] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 2

[60] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *CVPR*, 2019. 2

[61] Xiaolin Zhang, Yunchao Wei, Zhao Li, Chenggang Yan, and Yi Yang. Rich embedding features for one-shot semantic segmentation. *TNNLS*, 2021. 2

[62] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *arXiv*, 2018. 2

[63] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018. 2

[64] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2, 4, 7

[65] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 2

[66] Yinan Zhao, Brian L. Price, Scott Cohen, and Danna Gurari. Objectness-aware one-shot semantic segmentation. *arXiv*, 2020. 2

[67] Kai Zhu, Wei Zhai, and Yang Cao. Self-supervised tuning for few-shot segmentation. In *IJCAI*, 2020. 2