Computer Science
&
Data Science

Capstone Report - Fall 2022

Towards Generalized Few-shot Segmentation: On Contrastive Learning and Background Information Modeling

Haoming Liu,
Chengyu Zhang,
Xiaochen Lu

supervised by
Li Guo

Preface

Deep learning models are rapidly expanding in size, thus requiring more data for training. This raises some concerns about the adequacy of the data since the annotated samples can be extremely limited on some downstream tasks. This work tackles a classic computer vision task in low-data regimes, few-shot segmentation (FSS), aiming to improve the baseline performance and generalization ability of FSS models. We believe this work is highly informative and insightful for those who are interested in data-efficient learning, as the plug-and-play techniques introduced in this work are applicable to most relevant tasks in the field of computer vision, or deep learning in general.

Acknowledgements

We thank Professor Li Guo for her consistent guidance throughout the project. We thank Professor Hongyi Wen for his suggestions on the project write-up. This work was supported through the NYU IT High Performance Computing resources, services, and staff expertise.

Abstract

Few-shot segmentation (FSS) aims to train a segmentation model that can quickly adapt to a novel task given only a few annotated samples. It tackles the lack of annotated data for rare classes and improves the generalization ability of deep CNN based segmentation models. Adopting the transfer learning paradigm, this work produces a strong FSS baseline by incorporating contrastive learning components to obtain more discriminative feature embeddings. On top of that, we further expand this work into generalized few-shot segmentation (GFSS) and introduce two background modeling techniques. Specifically, we reformulate the per-pixel cross-entropy loss in the context of GFSS and further introduce a distillation loss to alleviate the catastrophic forgetting of existing knowledge. The proposed methods demonstrate significant performance gains on PASCAL-5^i and COCO-20^i benchmarks and are portable to most data-efficient learning tasks. The code is publicly available at: `https://github.com/hmdliu/GFSS-Capstone`.

Keywords

Generalized Few-shot Segmentation; Incremental Learning; Contrastive Learning; Knowledge Distillation

Contents

1 Introduction

Semantic segmentation refers to the task of predicting the class label for each pixel in the image. It is a fundamental task in scene understanding and has various real-world applications, such as remote sensing and autonomous vehicles. Along with the boom of deep learning, many deep CNN based segmentation models have achieved great performance on a set of pre-defined categories with sufficient human-annotated labels (e.g., PASCAL VOC, MS-COCO) [1, 2]. However, such training paradigms are still far from satisfactory since the training samples for some categories may not be abundant in practice. Accordingly, the task of few-shot segmentation (FSS) is proposed, which aims to predict the region mask of an unseen class based on a few annotations. It accommodates the lack of annotations on some downstream tasks and improves the generalization ability of segmentation models in low-data regimes.

Figure 1 gives an overall illustration of few-shot segmentation: FSS models are first trained on the data-abundant base classes and then tested on the data-scarce novel classes. The two dominant paradigms for FSS are meta learning and transfer learning, with the latter being adopted by most recent works due to its simplicity and superior performance [3, 4]. More concretely, the transfer learning paradigm aims to learn generic feature embeddings through training on sufficient base images. On top of that, only the classifier needs to be fine-tuned while adapting to a novel class. In this work, we term these two stages as pre-training and fine-tuning, respectively.

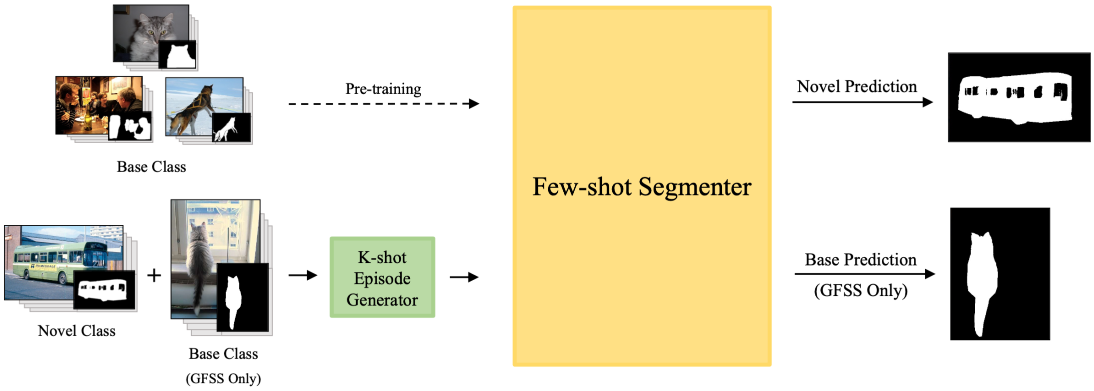

Figure 1: Few-shot Segmentation & Generalized Few-shot Segmentation Overview.

As most transfer learning based methods tend to make improvements in the fine-tuning stage, we wish to explore more possibilities in the pre-training stage. Contrastive learning is a common approach to producing more discriminative features, and it has demonstrated its power through its tremendous success in self-supervised learning. Inspired by a few recent works that incorporate contrastive learning into semantic segmentation [5, 6], we introduce a pixel-wise contrastive loss to the pre-training stage of FSS to improve the generalization ability of feature embeddings.

In addition, we observe that the conventional evaluation scheme for FSS is deficient in that the base class performance is generally disregarded, whereas the schemes for few-shot object detection usually perform evaluations on both novel and base classes [7, 8]. Due to the need of reserving base class knowledge after adapting to novel classes in some real-world scenarios, we extend our target task to generalized few-shot segmentation (GFSS) to ensure the comprehensiveness of the evaluation. However, this introduces the problem of background shift as pre-training stage and fine-tuning stage have different label spaces (i.e., base classes will be labeled as background in the ground truth in the fine-tuning stage of GFSS). To cope with this issue, we reformulate the per-pixel cross-entropy loss in the context of GFSS. Furthermore, we borrow ideas from the literature on incremental learning and devise a distillation loss to preserve the base class knowledge learned in the pre-training stage while adapting to a new task.

To sum up, the main contribution of this work can be summarized as follows:

- We propose to incorporate contrastive learning into the pre-training stage of FSS to gain more discriminative embeddings for the novel class adaptation.
- We expand our work into GFSS by reformulating the conventional per-pixel cross-entropy loss to tackle the background shift. Furthermore, we introduce a distillation loss to mitigate the catastrophic forgetting of the base class knowledge.
- We validate the effectiveness of the proposed methods through extensive experiments on the PASCAL-5^i and COCO-20^i benchmarks and discuss their other potential usage.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation is a dense prediction task that classifies each pixel of an image into a set of pre-defined semantic categories. Since the introduction of FCN [9], most semantic segmentation works have defaulted to fully convolutional architectures. PSPNet [10] incorporates spatial pyramid pooling to aggregate semantic information at different granularities. Deeplab [11] explores a dilated sampling strategy and proposes atrous spatial pyramid pooling (ASPP), which enlarges the receptive field and preserves the resolution of feature maps. Considering the promising performance PSPNet demonstrated in FSS [12, 13, 14], we adopt it as our backbone.

2.2 Few-shot Segmentation

Few-shot segmentation (FSS) aims to achieve effective knowledge transfer from the data-abundant base classes to the data-scarce novel classes. The works in this field can be divided into two paradigms: meta learning and transfer learning. **Meta learning** aims to extract task-level, class-agnostic meta knowledge for the model to quickly adapt to new tasks with very few annotated support examples. OSLSM [15] pioneered the field utilizing a conditioning branch to generate the classifier weights for query branch prediction. CANet [16] applies masked global average pooling on support images to generate prototypes and densely compare between the prototype and the novel query features. PFENet [13] achieves a competitive performance through a multi-scale aggregation module that refines the query feature with the support features and their generated priors. **Transfer learning** based approaches assume that the feature representations learned from the base classes are already generalizable, so instead of meta-training the feature extractor, they freeze the feature extractor pre-trained on base classes and focus on fine-tuning the classification layer for adaptation to new tasks. RePRI [14] introduces Shannon entropy and KL divergence to the loss functions and uses transductive inference. CWT [12] adds a transformer block to adaptively refine the classifier weights. Our work adopts a similar transfer learning based training scheme for FSS but focus more on learning stronger embeddings in the pre-training stage.

2.3 Contrastive Learning

Contrastive learning is a popular approach for representation learning, where the model learns in a discriminative manner by contrasting similar (positive) sample pairs against dissimilar (negative) sample pairs. In the context of computer vision, contrast is typically performed image-wise. The positive samples are generally multiple views of the same image obtained through applying strong perturbations [17], whereas the negative samples are usually sampled randomly [18]. Remarkably, contrastive learning is now a dominant branch of self-supervised learning, thanks to the great works such as SimCLR [19], MoCo [20] and DINO [21].

A few recent works in semantic segmentation have integrated contrastive learning into standard supervised training and achieved better performance. [5] proposes a pixel-wise contrastive learning paradigm across images and develops a region memory to enlarge the visual data space. [6] investigates and compares three variants of pixel-wise, label-based contrastive losses with sizable performance improvement. Inspired by these works, we formulate a simple but effective pixel-wise contrastive loss and tailor a sampling strategy for FSS pre-training.

2.4 Generalized Few-shot Segmentation

The goal of generalized few-shot segmentation (GFSS) is to train an FSS model that is able to give promising segmentation masks on both novel and base classes, which is more practical in real-world applications. [22] first introduces this task and proposes a prototype learning based method that yields significant performance gains over both the FSS and GFSS baselines. ABPNet [23] identifies the problem of background shift and meta-learns two modules for background modeling and prototype querying to address the problem. [24] explores a simple transfer learning baseline and formulates an augmented triplet loss for the fine-tuning stage. To the best of our knowledge, our work is the first transfer learning based approach that tackles the background shift problem.

3 Methodology

3.1 Problem Formulation

We follow the conventional FSS training scheme [15] and further expand it into GFSS. Formally, we are given two datasets $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ with category sets $\mathcal{C}_{base}$ and $\mathcal{C}_{novel}$ respectively, where $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$. Our objective is to learn a generalizable segmentation model through training on the base dataset $\mathcal{D}_{base}$ with sufficient annotated samples such that the model performs well on some unseen classes (drawn from $\mathcal{C}_{novel}$) with limited training samples (drawn from $\mathcal{D}_{novel}$). In general, FSS tasks are generated in an episodic manner. Specifically, we sample a series of episodes from $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ to simulate the few-shot scenario. Under a K-shot setting, each episode is composed of a small support set $\mathcal{S} = \{(x_k^s, m_k^s)\}_{k=1}^{K}$ and a query set $\mathcal{Q} = \{(x^q, m^q)\}$, where $x^{\cdot}$ and $m^{\cdot}$ represent a raw image and its corresponding binary mask for some class c. While evaluating the model, we randomly sample a class c from category set $\mathcal{C}_{test}$ and test the model on an episode of this class, where $\mathcal{C}_{test} = \mathcal{C}_{novel}$ for FSS and $\mathcal{C}_{test} = \mathcal{C}_{novel} \cup \mathcal{C}_{base}$ for GFSS.

3.2 Overall Framework

A semantic segmentation model are typically a fully convolutional network (FCN), consisting of an encoder, a decoder, and a simple classifier. The encoder gradually reduces the feature map resolution and captures higher semantic information; then, the decoder aggregates features from different stages and recover the spatial information. For simplicity, we denote the segmentation backbone (i.e., the encoder and the decoder) as f_θ. Then given an input image x, we can produce a dense feature $\mathcal{I} = f_\theta(x) \in \mathbb{R}^{H \times W \times D}$, where H and W are the feature resolution and D is the embedding dimension. To make predictions, we use a per-pixel classifier g_ψ to compute the `softmax` normalized segmentation mask $\hat{m} = g_\psi(I) \in \mathbb{R}^{H \times W \times |\mathcal{C}|}$ for each semantic class $c \in \mathcal{C}$, where $\mathcal{C}$ is a category set for some pre-defined classes.

In the context of FSS, transfer learning based approaches tend to follow a two-stage training scheme of pre-training and fine-tuning. In the **pre-training stage**, we train the feature extraction network (i.e., the encoder and the decoder) on the whole base dataset $\mathcal{C}_{base}$. Following previous works [12, 14], we use PSPNet [10] as the segmentation backbone and adopts a standard cross-entropy loss. In the **fine-tuning stage**, we freeze the backbone and train a new classifier with the support set $\mathcal{S}$ to adapt to a new class. Hence, we can make predictions on the query set $\mathcal{Q}$. Notably, the new classifier is 2-way (i.e., the novel and background class) for FSS and $(|\mathcal{C}_{base}|+2)$-way (i.e., the novel and background class, along with all base classes) for GFSS. More details about our training configurations are discussed in Section 4.2.

Figure 2 illustrates the model architectures and workflow we adopt for both FSS and GFSS. Section 3.3 presents the pixel-wise contrastive loss we devise for representation learning. Section 3.4 discusses how we reformulate the per-pixel cross-entropy loss and tackle the problem of background shift after extending to GFSS. Section 3.5 elaborates on the distillation-based regularizer we introduce to preserve the base class knowledge during new task adaptation.

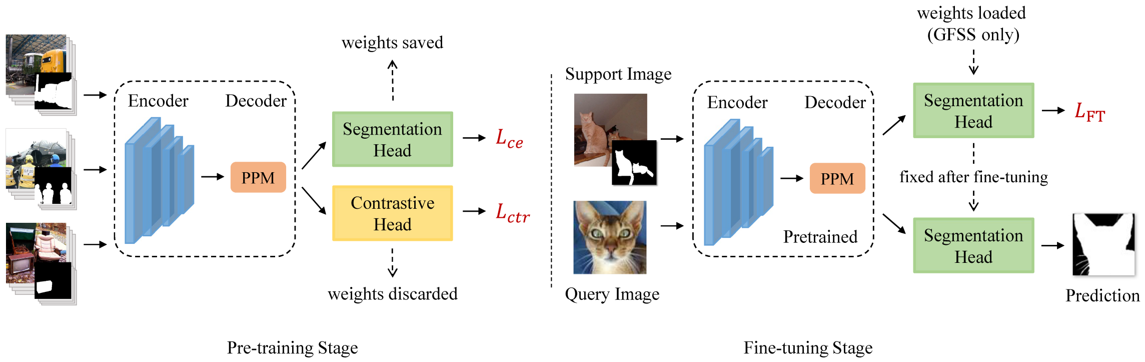

Figure 2: Model architectures for pre-training stage (left) and fine-tuning stage (right).

3.3 FSS: Contrastive Segmentation Pre-training

Pixel-wise Cross-Entropy Loss. Let $\hat{y}_i \in \mathbb{R}^{|\mathcal{C}|}$ be the predicted probability distribution for pixel i, where $\hat{y}_i$ is drawn from the `softmax` normalized segmentation mask $\hat{m}$. Then, the pixel-wise cross-entropy loss is defined as:

$$\mathcal{L}_i^{\mathrm{CE}} = -\log q_x(i, y_i), \tag{1}$$

where y_i is the ground truth label for pixel i and $q_x(i, c) = \hat{y}_i[c]$. As noted by [5], this training objective suffers from two potential limitations: 1) it penalizes pixel-wise predictions independently but fails to model the correlations between pixels; 2) due to the entailed `softmax` function in Eq.1, the supervision on the feature representations is indirect since the loss only depends on the relative relation between logits. Accordingly, introducing a pixel-wise contrastive loss is an effective way of tackling both limitations, as it models pixel-to-pixel relations and provides direct supervision on the feature representations.

Pixel-to-Pixel Contrastive Loss. Inspired by recent semantic segmentation works that incorporates contrastive learning into standard supervised training [5, 6], we introduce a pixel-to-pixel contrastive loss to complement the pixel-wise corss-entropy loss and produce more discriminative feature embeddings. Given the dense feature $\mathcal{I} = f_\theta(x) \in \mathbb{R}^{H \times W \times D}$ for an image x, we can derive the pixel-wise embedding $\boldsymbol{i} \in \mathbb{R}^D$ for all $\boldsymbol{i} \in \mathcal{I}$. Then, the InfoNCE loss [25] can be defined as:

$$\mathcal{L}_i^{\text{NCE}} = \frac{1}{|\mathcal{P}_i|} \sum_{\boldsymbol{i}^+ \in \mathcal{P}_i} -\log \frac{\exp(\boldsymbol{i} \cdot \boldsymbol{i}^+ / \tau)}{\exp(\boldsymbol{i} \cdot \boldsymbol{i}^+ / \tau) + \sum_{\boldsymbol{i}^- \in \mathcal{N}_i} \exp(\boldsymbol{i} \cdot \boldsymbol{i}^- / \tau)}, \tag{2}$$

where $\boldsymbol{i}$ serves as the anchor point, $\mathcal{P}_i$ and $\mathcal{N}_i$ corresponds to two set of pixel-wise embeddings for the positive and negative samples, respectively. The intuition behind Eq.2 is that we pull positive sample pairs closer and push negative sample pairs apart. As we are tackling image data with a considerable amount of pixels, it is impractical to make contrast between every possible pairs. Instead, we need an effective sampling strategy for anchor points and hard examples. In this work, we adapt the sampling strategy[1] proposed by [6] to the context of FSS pre-training. Specifically, the clasifier in the pre-training stage generally consists of $|\mathcal{C}| + 1$ classes, where the extra one denotes the background. We exclude the background pixels from anchor point sampling, as it may belong to any unlisted semantic categories. Note that the sampling is performed within a mini-batch, so a sample pair may include pixels drawn from different images. A pixel sample falls into $\mathcal{P}_i$ if its semantic category is identical to the anchor point, and it falls into $\mathcal{N}_i$ otherwise.

Combining the pixel-wise cross-entropy loss in Eq.1 and the pixel-to-pixel loss in Eq.2, the pre-training loss can be computed as follows:

$$\mathcal{L}_{\text{PT}} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{ctr}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (\mathcal{L}_i^{\text{CE}} + \lambda \mathcal{L}_i^{\text{NCE}}), \tag{3}$$

where λ is a positive coefficient balancing the two terms. By regularizing the embedding space

[1] For simplicity, we refer the readers to the original work for more details.

through $\mathcal{L}_{\text{ctr}}$, the segmentation backbones explores the global structures of the data and produces more generic embeddings. We present our ablation studies in Section 4.3 to validate it efficacy.

3.4 GFSS: Reformulating Cross-Entropy Loss

We expand our work into GFSS to cater the need for base class knowledge after novel class adaptation. The shift of the task introduces a major change in the fine-tuning stage, that is, we need to load the weights of the $(|\mathcal{C}_{base}|+1)$-way classifier (i.e., base classes and background) from the pre-training stage and form a $(|\mathcal{C}_{base}|+2)$-way classifier with the novel class. Most existing GFSS works follow a convention proposed by [22], where every base and novel class instances in a ground truth mask are exhaustively labeled. We argue that such a training scheme is by no means practical since gathering such comprehensive annotations can be costly and pointless. Considering the fact that we are only concerned about the regions of a single query class in most cases, we propose to adopt unified binary ground truth masks for both FSS and GFSS.

However, this raises a problem on fine-tuning since the ground truth mask only indicates the region of either class of interest (c^*) or others. To cope with this issue, we aggregate the output probabilities for other classes (i.e., $\mathcal{C} \setminus \{c*\}$) and reformulate the task to a binary classification problem. On top of Eq.1, the modified per-pixel cross-entropy loss can be defined as:

$$\mathcal{L}_i^{\text{CE}'} = -\log \tilde{q}_x(i, y_i), \text{ where } \tilde{q}_x(i,c) = \begin{cases} \hat{y}_i[c] & \text{if } c = c^* \\ \sum_{k\in\mathcal{C}\setminus\{c*\}} \hat{y}_i[k] & \text{otherwise} \end{cases}. \tag{4}$$

Hence, the training objective for our GFSS baseline can be computed as:

$$\mathcal{L}_{\text{ce}'} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathcal{L}_i^{\text{CE}'}. \tag{5}$$

3.5 GFSS: Background Modeling via Distillation

Following the training objective defined in Eq.4 and Eq.5, the model is trained to distinguish the foreground and background regions, where the foreground corresponds to the class of interest and the background is a collection of all other semantic categories. Though the model masters a novel class after the fine-tuning stage, the performance drops on the base classes can be quite significant due to the monotonous focus. This is generally termed as catastrophic forgetting. Drawn ideas from incremental learning [26], introducing a distillation loss is a common strategy to preserve the knowledge from previous tasks. In the context of GFSS, we aims to introduce a regularizer that enforces the model to make predictions that is close to the classifer in the pre-training stage. A standard distillation loss can be defined as:

$$\mathcal{L}_i^{\mathrm{KD}} = - \sum_{c \in y_i^{base}} q_x^{PT}(i,c) \log q_x^{FT}(i,c), \tag{6}$$

where $y^{base} \in \mathbb{R}^{|\mathcal{C}_{base}|+1}$ is the ground truth label in the pre-training stage, q_x^{PT} and q_x^{FT} give the predicted probability in the Pre-Training (PT) and Fine-Tuning (FT) stage respectively. Let $\hat{y}^{base} \in \mathbb{R}^{|\mathcal{C}_{base}|+1}$ be the prediction drawn from the base classifier, we have $q_x^{PT}(i,c) = \hat{y}^{base}[c]$. Given the background class c_b and the novel class c_n in the fine-tuning stage, we have:

$$q_x^{FT}(i,c) = \begin{cases} \hat{y}_i[c_b] + \hat{y}_i[c_n] & \text{if } c = c_b \\ \hat{y}_i[c] & \text{otherwise} \end{cases}. \tag{7}$$

The intuition behind Eq.6 is that we regularize the output probability distributions from the new classifier with the base classifier predictions. Eq.7 tackles the inconsistency of classes (i.e., background shift), as the background in the pre-training stage includes both the background and the novel class in the fine-tuning stage. After introducing the distillation loss, the training objective can be formulated as:

$$\mathcal{L}_{\mathrm{FT}} = \mathcal{L}_{\mathrm{ce}'} + \gamma \mathcal{L}_{\mathrm{distil}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (\mathcal{L}_i^{\mathrm{CE}'} + \gamma \mathcal{L}_i^{\mathrm{KD}}), \tag{8}$$

where γ is a positive coefficient balancing the two terms. By enforcing a non-forgetting learning objective through $\mathcal{L}_{\mathrm{distil}}$, the problem of catastrophic forgetting on base classes is alleviated. We conduct extensive experiments to investigate its impact in Section 4.4.

4 Experiment

4.1 Datasets

We evaluate the performance of the proposed methods on two widely-used FSS benchmarks, PASCAL-5^i [15] and COCO-20^i [27]. PASCAL-5^i is built from PASCAL VOC [1] and contains 20 semantic classes that are evenly divided into 4 folds. COCO-20^i is built from MS-COCO [2] and consists of 80 semantic classes that are evenly divided into 4 folds. For both datasets, the model is trained on 3 folds and tested on the remaining one in a cross-validation manner.

4.2 Implementation Details

Pre-training. We build our model based on PSPNet [10] with a ResNet-50 [28] backbone. We adopt the standard supervised learning to train the segmentation backbone on each fold of the FSS dataset, which consists of 16/61 classes (background included) for PASCAL-5^i and COCO-20^i respectively. We train the model for 100 epochs on PASCAL-5^i and 20 epochs for COCO-20^i with a per-pixel cross-entropy loss (Eq.1). The balancing coefficient λ for the contrastive loss (Eq.2) is searched within the hyper-parameter space $\{0.01, 0.05, 0.1, 0.2, 0.4\}$ and we pick the best-performed one for each data fold. We use a SGD optimizer with an initial learning rate of 2.5e$-$3 with a cosine learning rate decay, a momentum of 0.9, and a weight decay of 1e$-$4. We set the batch size to 12 and resize input images to 473×473. Label smoothing is used with the smoothing parameter $\epsilon = 0.1$. We only apply random mirror flipping for data augmentations.

Episodic Inference. The transfer learning baseline we adopt in this work doesn't need to be meta-learned. In other words, we can perform inference (i.e., fine-tuning with the support set $\mathcal{S}$) on a given episode right after the pre-training stage. We use a SGD optimizer with a learning rate of 1e$-$1 and perform 100 optimization steps. We adopt a standard per-pixel cross-entropy loss for FSS, and we use the proposed training objectives (Eq.8) for GFSS.

Evaluation Metrics. We adopt the widely used mean Intersection over Union (mIoU) as the evaluation metric. It is computed by averaging the IoU values of all classes in a fold. Following the convention in FSS, the model is validated on 20000 randomly sampled episodes. For GFSS, we randomly sample an extra base set $\mathcal{B} = \{(x^b, m^b)\}$ to evaluate the performance on $\mathcal{C}_{base}$.

4.3 FSS: Ablation Studies

We incorporate the proposed pixel-to-pixel contrastive loss (Eq.2) into a transfer learning baseline to evaluate its efficacy. Specifically, we first pre-train a baseline backbone with standard per-pixel cross-entropy loss (Eq.1) and then pre-train another backbone that introduces pixel-to-pixel contrastive loss (Eq.2). Then we evaluate the embeddings produced by different backbones through an FSS baseline model. Following the convention, we conduct experiments on two classic benchmarks, PASCAL-5^i and COCO-20^i [1, 2], and report the mean IoU on novel classes under 1-shot and 5-shot settings. As shown in Table 1, the proposed contrastive loss exhibits sizable performance gains on COCO-20^i (+1.2% for 1-shot and +0.5% for 5-shot), whereas the performance degrades on PASCAL-5^i (-1.2% for 1-shot and -2.3% for 5-shot). Such results can be expected as contrastive learning heavily relies on the variety of classes. On that note, the

20 semantic classes on PASCAL-5^i may be inadequate for the model to distinguish between the numerous semantic classes in practice, whereas the 80 semantic classes on COCO-20^i entail much richer semantic information and thus result in more generic feature embeddings.

Dataset	Method	1-shot Novel mIoU					5-shot Novel mIoU				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
PASCAL-5^i	Baseline	54.5	62.2	60.4	46.4	55.9	61.4	70.5	71.6	59.6	65.8
	w/ $\mathcal{L}_{\text{ctr}}$	53.6	62.3	58.0	45.0	54.7	59.8	69.4	67.9	56.7	63.4
	Δ	-0.9	+0.1	-2.4	-1.4	-1.2	-1.6	-1.1	-3.8	-2.9	-2.3
COCO-20^i	baseline	30.2	33.4	30.3	32.4	31.6	40.6	42.8	38.1	42.4	41.0
	w/ $\mathcal{L}_{\text{ctr}}$	31.6	35.2	31.5	32.7	32.8	39.1	44.5	40.2	41.9	41.4
	Δ	+1.4	+1.8	+1.2	+0.3	+1.2	-1.5	+1.8	+2.1	-0.5	+0.5

Table 1: Ablation studies of contrastive pre-training on PASCAL-5^i and COCO-20^i.

4.4 GFSS: Ablation Studies

In the context of GFSS, we load the base classifier weights from the pre-training stage to perform predictions on base classes. Interestingly, we found that this simple expansion of classifer is also beneficial to the performance on novel classes. As shown in Table 2, we observe performance gains of +1.8% and +2.6% on PASCAL-5^i and COCO-20^i under 1-shot setting. We suppose this is because the base classifier helps filter out some confusing regions under extremely limited data. This gain effect becomes progressively weaker as the number of training samples increases. Notably, the performance degrades on COCO-20^i under 5-shot settings, and we suspect that the misclassification of novel classes as base classes can be accumulated, which could result in a more pronounced impact when there are more possible semantic categories.

Dataset	CLS Width	1-shot Novel mIoU					5-shot Novel mIoU				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
PASCAL-5^i	2	54.5	62.2	60.4	46.4	55.9	61.4	70.5	71.6	59.6	65.8
	$\lvert\mathcal{C}_{base}\rvert + 2$	58.8	64.8	59.5	47.8	57.7	62.7	71.7	72.1	60.9	66.9
	Δ	+4.4	+2.7	-1.0	+1.4	+1.8	+1.3	+1.2	+0.5	+1.4	+1.1
COCO-20^i	2	30.2	33.4	30.3	32.4	31.6	40.6	42.8	38.1	42.4	41.0
	$\lvert\mathcal{C}_{base}\rvert + 2$	31.1	37.5	33.5	34.6	34.1	40.7	42.0	35.8	41.9	40.1
	Δ	+0.9	+4.1	+3.2	2.2	+2.6	+0.0	-0.8	-2.3	-0.5	-0.9

Table 2: Comparison of FSS results using 2-way classifier and ($\lvert\mathcal{C}_{base}\rvert + 2$)-way classifier. Their training objectives are $\mathcal{L}_{\text{ce}}$ (Eq.1) and $\mathcal{L}_{\text{ce}'}$ (Eq.4), respectively. The classifier weights for base classes are directly drawn from the pre-training stage.

We investigate the efficacy of the proposed loss (Eq.8) for the fine-tuning stage using different distillation weights (γ). As a regularizer that preserve the knowledge on base classes, enlarging the distillation weight generally increases the performance on base classes and may degrade the performance on novel classes if the training samples are extremely limited (e.g., under 1-shot setting). According to Table 3, we can observe a consistent performance improvement on base classes as we enlarge the distillation weight. Meanwhile, the 1-shot novel mIoU gradually decreases accordingly. We also notice that the effect of distillation varies across folds, for example, introducing distillation always boosts novel mIoU on Fold-2 but usually degrades the performance on Fold-3. This suggests that the correlations between base and novel categories play an important role in terms of the efficacy of distillation. In particular, we find that distillation tends to perform well if one of the base classes is similar to a query (novel) class (e.g., bus and train), as it helps the model clarify the confusing semantic categories.

NS	γ	Base mIoU					Novel mIoU				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
1	0	52.3	43.8	47.7	52.0	48.9	31.1	37.5	33.5	34.6	34.1
	5	54.9 (+2.6)	45.1 (+1.3)	49.5 (+1.8)	52.7 (+0.7)	50.5 (+1.6)	31.5 (+0.4)	37.6 (+0.1)	33.8 (+0.3)	34.7 (+0.1)	34.4 (+0.3)
	10	55.0 (+2.7)	45.7 (+1.9)	50.0 (+2.3)	52.9 (+0.9)	50.9 (+2.0)	31.0 (-0.1)	37.2 (-0.3)	33.7 (+0.2)	34.5 (-0.1)	34.1 (+0.0)
	20	55.2 (+2.9)	46.3 (+2.5)	50.4 (+2.7)	53.0 (+1.0)	51.2 (+2.3)	30.0 (-1.1)	36.5 (-1.0)	33.5 (+0.0)	34.1 (-0.5)	33.5 (-0.6)
5	0	50.8	41.6	46.3	50.8	47.4	40.7	42.0	35.8	41.9	40.1
	5	53.1 (+2.3)	42.6 (+1.0)	47.2 (+0.9)	52.0 (+1.2)	48.7 (+1.3)	41.2 (+0.5)	42.1 (+0.1)	36.2 (+0.4)	41.4 (-0.5)	40.2 (+0.1)
	10	53.5 (+2.7)	43.3 (+1.7)	47.9 (+1.6)	52.3 (+1.5)	49.3 (+1.9)	41.0 (+0.3)	42.2 (+0.2)	36.5 (+0.7)	41.3 (-0.6)	40.2 (+0.1)
	20	53.9 (+3.1)	44.2 (+2.6)	48.7 (+2.4)	52.7 (+1.9)	49.9 (+2.5)	40.9 (+0.2)	42.1 (+0.1)	36.7 (+0.9)	41.5 (-0.4)	40.3 (+0.2)

Table 3: Comparison of GFSS results with different distillation weights (γ) on COCO-20^i, where NS refers to number of shots. $\gamma = 0$ indicates the baseline with no distillation loss.

NS	MT	Base mIoU on PASCAL-5^i					Novel mIoU on PASCAL-5^i				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
1	B	44.8	54.9	61.6	56.9	54.5	58.8	64.8	59.5	47.8	57.7
	DL	65.4 (+20.6)	70.4 (+15.5)	71.4 (+9.8)	71.4 (+14.5)	69.6 (+15.1)	59.1 (+0.3)	61.0 (-3.8)	56.4 (-3.1)	47.1 (-0.7)	55.9 (-1.8)
5	B	35.6	55.1	61.3	54.6	51.6	62.7	71.7	72.1	60.9	66.9
	DL	59.9 (+24.3)	70.7 (+15.6)	71.6 (+10.3)	72.0 (+17.4)	68.6 (+17.0)	62.6 (-0.1)	72.2 (+0.5)	71.1 (-1.0)	61.0 (+0.1)	66.7 (-0.2)
NS	MT	Base mIoU on COCO-20^i					Novel mIoU on COCO-20^i				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
1	B	52.3	43.8	47.7	52.0	48.9	31.1	37.5	33.5	34.6	34.1
	B†	48.2 (-4.1)	47.0 (+3.2)	53.6 (+5.9)	53.0 (+1.0)	50.5 (+1.6)	29.9 (-1.2)	34.0 (-3.5)	32.2 (-1.3)	33.6 (-1.0)	32.4 (-1.7)
	DL	55.0 (+2.7)	45.7 (+1.9)	50.0 (+2.3)	52.9 (+0.9)	50.9 (+2.0)	31.0 (-0.1)	37.2 (-0.3)	33.7 (+0.2)	34.5 (-0.1)	34.1 (+0.0)
	DL†	49.5 (-2.8)	48.3 (+4.5)	55.0 (+7.3)	54.1 (+2.1)	51.7 (+2.8)	28.1 (-3.0)	32.0 (-5.5)	31.7 (-1.8)	31.8 (-2.8)	30.9 (-3.2)
5	B	50.8	41.6	46.3	50.8	47.4	40.7	42.0	35.8	41.9	40.1
	B†	46.2 (-4.6)	46.4 (+4.8)	52.1 (+5.8)	51.4 (+0.6)	49.0 (+1.6)	39.1 (-1.6)	45.4 (+3.4)	40.7 (+4.9)	42.9 (+1.0)	42.0 (+1.9)
	DL	53.5 (+2.7)	43.3 (+1.7)	47.9 (+1.6)	52.3 (+1.5)	49.3 (+1.9)	41.0 (+0.3)	42.2 (+0.2)	36.5 (+0.7)	41.3 (-0.6)	40.2 (+0.1)
	DL†	47.9 (-2.9)	47.2 (+5.6)	53.5 (+7.2)	53.0 (+2.2)	50.4 (+3.0)	40.0 (-0.7)	45.4 (+3.4)	41.3 (+5.5)	42.6 (+0.7)	42.3 (+2.2)

Table 4: GFSS results on PASCAL-5^i and COCO-20^i, where NS refers to number of shots, MT refers to method, B refers to baseline, DL refers to baseline with distillation loss ($\gamma = 10$), † refers to using backbone weights from contrastive pre-training.

Table 4 presents the GFSS results on PASCAL-5^i and COCO-20^i, where all the proposed techniques have been integrated. As discussed in Section 4.3, contrastive pre-training tends to exhibit performance gains when the size of semantic categories is large, so we only adopt it on COCO-20^i. The experiment results on PASCAL-5^i demonstrate the significant performance gains (+15.1% under 1-shot setting and +17.1% under 5-shot setting) after introducing the distillation loss, but it inevitably degrades the performance on novel classes (-1.8% under 1-shot setting and -0.2% under 5-shot setting). As shown by the experiments on COCO-20^i, all the proposed techniques offer performance gains on base classes. With properly-chosen hyper-parameters (i.e., λ and γ), we can also improve the mean IoU on novel classes, especially when more training samples are available. By combining contrastive pre-training and background knowledge distillation, our model well balances and enhances the performance on GFSS benchmarks.

4.5 Qualitative Results

We also conduct ablation studies qualitatively by visualizing the prediction masks in Figure 3. As shown in the left figure, the proposed contrastive loss (Eq.2) allows the backbone to learning more generic feature embedding that help it distinguish between various semantic categories and segment the boundary of different instances. As demonstrated in the right figure, the distillation loss (Eq.6) preserves the knowledge on base classes and alleviates catastrophic forgetting.

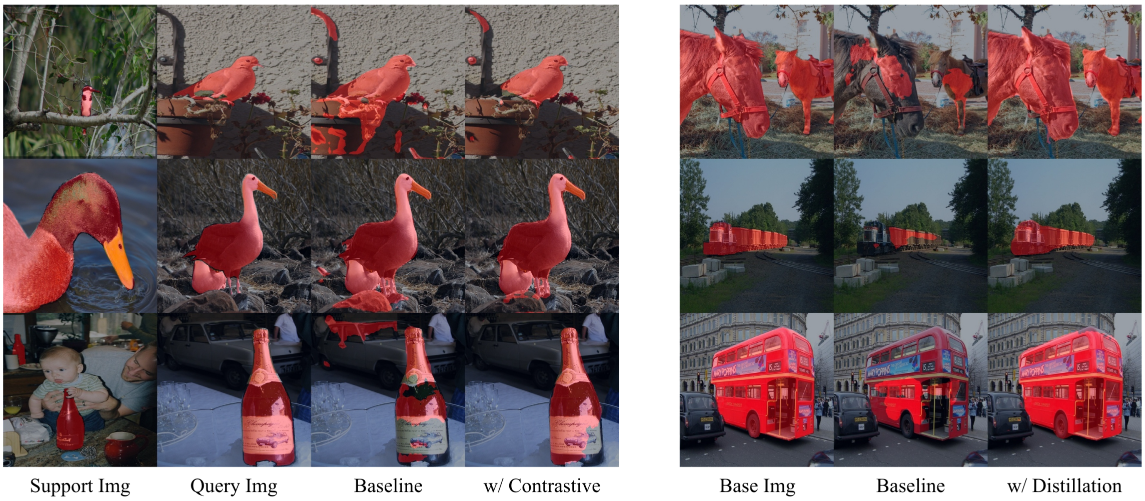

Figure 3: Qualitative Ablation on Contrastive Loss (left) and Distillation Loss (right).

5 Discussion & Conclusion

This work presents a practical GFSS training scheme along with two plug-and-play techniques. In the pre-training stage, we complement standard per-pixel cross-entropy loss by introducing a contrastive loss that models pixel-to-pixel correlations. It helps produce more discriminative feature embeddings and offers sizable performance gains, especially when adequate semantic categories are provided. Besides, we propose to adopt unified binary mask for both FSS and GFSS, which no longer requires annotating novel images exhaustively and is closer to real-world scenarios. On top of this training scheme, we reformulate the cross-entropy loss and devise a distillation loss to model background information and avoid catastrophic forgetting. Their efficacy has been shown by extensive experiments on PASCAL-5^i and COCO-20^i.

Remarkably, the proposed contrastive and distillation loss are not restricted to the context of FSS and GFSS. Since the pixel-to-pixel contrastive loss improves the generalization ability of the feature representations, its core idea is highly portable to any dense prediction tasks in computer vision. Furthermore, the distillation loss is also widely applicable to any data-efficient learning tasks (e.g., few-shot recognition & detection) to alleviate the problem of catastrophic forgetting.

Last but not least, there are many possibilities for us to refine the approaches proposed in this work. For example, we can follow [5] and introduce a memory bank to store high-quality samples for contrastive learning. Besides, we may also borrow ideas from constraint-based fine-tuning [29] to mitigate the performance drops on novel classes during distillation.

References

[1] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. European Conference on Computer Vision, September 2014.

[3] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. Ben Ayed, and J. Dolz, "Few-shot segmentation without meta-learning: A good transductive inference is all you need?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 979–13 988.

[4] Z. Lu, S. He, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang, "Simpler is better: Few-shot semantic segmentation with classifier weight transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8741–8750.

[5] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7303–7313.

[6] X. Zhao, R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu, "Contrastive learning for label efficient semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 623–10 633.

[7] X. Yan, Z. Chen, A. Xu, X. Wang, X. Liang, and L. Lin, "Meta r-cnn: Towards general solver for instance-level low-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9577–9586.

[8] X. Wang, T. Huang, J. Gonzalez, T. Darrell, and F. Yu, "Frustratingly simple few-shot object detection," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9919–9928.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[12] Z. Lu, S. He, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang, "Simpler is better: Few-shot semantic segmentation with classifier weight transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8741–8750.

[13] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1050–1065, 2022.

[14] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. B. Ayed, and J. Dolz, "Few-shot segmentation without meta-learning: A good transductive inference is all you need?" in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 974–13 983.

[15] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, September 2017, pp. 167.1–167.13.

[16] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.

[17] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[18] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 798–21 809, 2020.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[21] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.

[22] Z. Tian, X. Lai, L. Jiang, S. Liu, M. Shu, H. Zhao, and J. Jia, "Generalized few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 563–11 572.

[23] K. Dong, W. Yang, Z. Xu, L. Huang, and Z. Yu, "Abpnet: Adaptive background modeling for generalized few shot segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2271–2280.

[24] J. Myers-Dean, Y. Zhao, B. Price, S. Cohen, and D. Gurari, "Generalized few-shot semantic segmentation: All you need is fine-tuning," *arXiv preprint arXiv:2112.10982*, 2021.

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] F. Cermelli, M. Mancini, S. R. Bulo, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9233–9242.

[27] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 622–631.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] K. Guirguis, A. Hendawy, G. Eskandar, M. Abdelsamad, M. Kayser, and J. Beyerer, "Cfa: Constraint-based finetuning approach for generalized few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4039–4049.