MaskSplit: Self-supervised Meta-learning for Few-shot Semantic Segmentation

Mustafa Sercan Amac*[1], Ahmet Sencan*[2], Orhun Bugra Baran[2], Nazli Ikizler-Cinbis[3], and Ramazan Gokberk Cinbis[2]

[1]MonolithAI, [2]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, [3]Department of Computer Engineering, Hacettepe University, Ankara, Turkey, sercanamac@gmail.com, {sencan.ahmet,bbaran,gcinbis}@metu.edu.tr, nazli@cs.hacettepe.edu.tr

Abstract

Just like other few-shot learning problems, few-shot segmentation aims to minimize the need for manual annotation, which is particularly costly in segmentation tasks. Even though the few-shot setting reduces this cost for novel test classes, there is still a need to annotate the training data. To alleviate this need, we propose a self-supervised training approach for learning few-shot segmentation models. We first use unsupervised saliency estimation to obtain pseudo-masks on images. We then train a simple prototype based model over different splits of pseudo masks and augmentations of images. Our extensive experiments show that the proposed approach achieves promising results, highlighting the potential of self-supervised training. To the best of our knowledge this is the first work that addresses unsupervised few-shot segmentation problem on natural images.

1. Introduction

Semantic segmentation is the task of assigning labels to pixels of a given image. There has been tremendous progress in semantic segmentation with the developments in architectures [40, 25, 37, 23, 3, 55]. However, these approaches typically require large amounts of training data for each class of interest to achieve accurate results. The manual effort needed for collecting segmentation annotations greatly limits the scalability of such supervised approaches.

Aiming to mimic the human ability to recognize and segment novel object classes with just a few examples, few-shot semantic segmentation has gained popularity over the past few years. In contrast to supervised approaches, *few-shot semantic segmentation* (FSS) aims to estimate the mask for an input image, *i.e*. the *query* image, with the help of just a few training images and their groundtruth masks, *i.e*. the *support* samples.

*Equal contribution. Listing order is random.

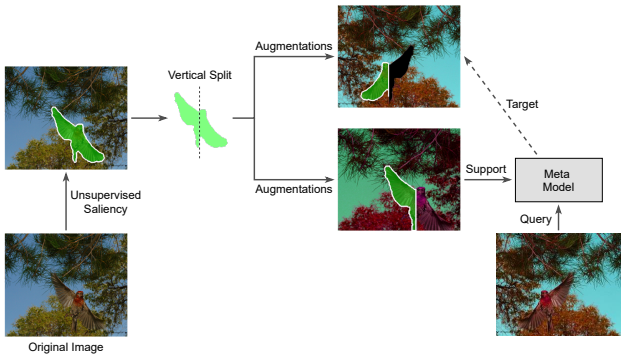

Figure 1: Illustration of the proposed self-supervised meta-learning framework. We split saliency masks to create artificial support and query training pairs, using which we train the few-shot segmentation model. Thanks to this self-supervised learning setup, no manual segmentation annotations are needed during training.

The recent work on FSS focuses on modeling one-to-one correspondence between support and query pixels to create correlation scores [27, 44, 53]. Another line of research is learning *class prototypes* to be used in deciding whether each pixel belongs to the object or the background [49, 22, 52, 42, 54, 45, 41]. To build models that are trained to generalize over few training samples, recent work widely adopts the *meta learning* paradigm [27, 44, 53, 49, 22, 52, 42, 54, 45] . In meta learning, the main idea is to create a series of tasks based on the training set simulating the few-shot segmentation problem and train the meta model over these tasks. As an alternative, it has also been recently shown that carefully designed transductive inference mechanisms combined with simple pretraining and fine-tuning strategies can yield state-of-the-art results without involving meta learning procedures [1].

Even though the few-shot setting enables the models to make predictions for the new classes of objects with just a

few annotated images, these models still need to be trained on a large number of base training classes to achieve strong few-shot generalization. Thus, just like the fully-supervised approaches, these models still rely on large amounts of training examples for training the meta model (or pretraining the base model). As a result, even the few-shot learning approaches end up being practically limited due to the difficulties of collecting rich segmentation training data sets. Since a meta (base) few-shot segmentation model is effectively a deep segmentation architecture, the limitations in base training data is likely to affect the generalization abilities of the learned meta model to a great extent.

To study ways to reduce the annotation dependency, we explore the problem of self-supervised learning of few-shot segmentation models. The goal is to learn a class-agnostic meta model without using any segmentation annotations during training. Once the self-supervised meta model is learned, the segmentation model of a novel class is obtained on-the-fly using a single sample, just as in standard FSS. Since meta model learning requires no manual annotations, such an approach, in theory, can allow leveraging arbitrarily large-scale and rich unlabelled image collections. With a similar motivation, a self-supervised FSS approach for medical images is first proposed by [33]. We generalize the problem to natural images with complex scenes.

In order to by-pass the dependency on supervised segmentation data for the base classes and realize self-supervised meta-learning for FSS, we propose a saliency-based scheme for creating a training pair from each training image. More specifically, we create episodic training tasks by (i) estimating an unsupervised saliency mask at each image, (ii) splitting the mask stochastically into two parts, and (iii) treating the splitted mask pairs, after image augmentations, as support and query pairs for training the FSS model. The combination of mask splitting and strong augmentations effectively create challenging one-shot segmentation tasks. Figure 1 illustrates the proposed approach.

There has been significant progress in self-supervised learning over the past few years [20, 4, 5, 31, 8, 34, 21, 28]. In particular, contemporary methods obtain representations that are competitive with their supervised counterparts when applied to recognition tasks such as image classification, object detection or semantic segmentation. In contrast, in this paper we learn the few-shot segmentation model itself in a self-supervised manner, and to the best of our knowledge, ours is the first one in this direction for natural images.

We conduct extensive experiments on the widely used FSS benchmarks based on the MS-COCO [24] and PASCAL [10]. We also present a detailed ablative study, where we experimentally analyze the model over a variety of supervision settings and model component configurations.

In the remainder of the paper, we first present an overview of the recent developments on (few-shot) semantic segmentation and self supervised learning. We then explain our proposed method, which consists of the problem setup, the novel self-supervised training procedure and the architecture of our FSS network. Finally, we present a detailed experimental analysis of the proposed framework and conclude the paper.

2. Related work

Fully-supervised & few-shot semantic segmentation. In fully-supervised semantic segmentation, a central challenge is obtaining high-resolution segmentation results by efficiently modeling both contextual and local information. To incorporate the contextual information efficiently, [2, 50] introduce *dilated convolution*, which allows the enlargement of the receptive field of a convolutional kernel, without increasing the number of trainable parameters. To tackle the same problem, pooling mechanisms, such as *global average pooling* [25], *pyramid pooling module* [57], and *atrous spatial pyramid pooling* [2], also offer powerful modeling tools. Encoder-decoder like architectures are similarly widely used to design efficient and effective semantic segmentation networks, *e.g*. [40, 37, 13, 23, 3]. Attention mechanisms are also used to improve long range interactions across regions, *e.g*. [58, 55, 56, 12, 17].

For few-shot segmentation, the pioneering work of [39] proposes a two-branch solution where the conditioning branch predicts the task parameters using the support set, and then these parameters guide the segmentation branch in predicting pixel-wise labels. Follow-up works can be categorized as prototype-based, graph-based and meta-learning free approaches. In prototype-based approaches, the aim is to obtain features from support examples that summarize classes with pooling and those support features are typically matched with query features via a distance metric [9, 45, 48, 54, 36]. In order to create stronger connections between the support and query features, graph-based approaches [53] [44] [46] are also used. Since most prototype-based approaches rely on global average pooling, the aim of graph based approaches is to establish more local-to-local connections between the support and query features. Another approach with the same objective is [47] where a memory network is trained with different query feature resolutions. In contrast to these meta-learning approaches, [1] first trains a segmentation representation over the base classes using supervised training, and then uses transductively regularized fine-tuning to obtain task-specific models.

Self-supervised learning. Self-supervised learning focuses on extracting supervision from the structure of data. Recently, it has been shown that high-level semantic visual representations can be successfully learned by using self-

supervision and directly used in downstream tasks such as classification, detection and segmentation tasks [20, 4, 5]. Some pretext tasks for self-supervision involve solving jigsaw puzzles [31, 8], rotation prediction [20], emptied pixel prediction [34], and order prediction [21, 28].

Self-supervision based training signals have recently been used to improve FSS models. In particular, [59] proposes to get more refined support features by using self-supervision on support images through inner gradient optimization. [52] similarly aims to obtain improved support features with self-supervision over the support masks. Our work fundamentally differs from both of these approaches as we aim to formulate a purely self-supervised approach to learn the meta-model in an unsupervised manner, as opposed to defining an auxiliary self-supervised training loss to improve the model in a traditional FSS setting.

A related problem is *unsupervised semantic segmentation*, where the goal is to cluster pixels across images into semantic semantic groups. This problem has recently been tackled using end-to-end [18, 32], and two-staged approaches [43]. Our work fundamentally differs in terms of both the problem definition and the overall approach. First, instead of unsupervisedly learning a segmentation model of a fixed number of classes, we aim to learn a class-agnostic meta model that can synthesize semantic segmentation of an arbitrary novel class based on a single training example, in an unsupervised manner. Second, as opposed to the clustering based approaches in unsupervised semantic segmentation, our approach relies on episodic meta-learning over self-supervisedly generated pseudo-groundtruths.

Finally, we should note that few works have recently explored self-supervised meta-learning in a few-shot classification context: [16] and [19] proposes clustering (and augmentation) based task creation schemes for learning few-shot classification models. To the best of our knowledge, ours is the first work to propose and study self-supervised learning of a few-shot segmentation model.

3. Method

In this section, we first formally define the self-supervised few-shot segmentation problem, and then the proposed training methodology. Finally, we define our network architecture realizing the proposed approach.

3.1. Preliminaries and problem definition

Traditional few-shot segmentation. The ultimate goal of few-shot segmentation is to obtain a meta model that can yield an accurate segmentation model of a novel class, given just one or few samples for the novel class. In the standard FSS scenario, the FSS model itself is meta-learned (or pretrained) over a supervised training set D_{train} over classes C_{train}, and evaluated over a test set D_{test} over classes C_{test}. Since the goal is to learn an FSS model that generalizes well to novel classes, D_{train} and D_{test} consists of distinct classes, *i.e.* $C_{test} \cap C_{train} = \emptyset$. In these data sets, each example corresponds to a triplet (x, m, y) where x is the image, m is the groundtruth binary mask and y is the class label corresponding to the mask.

Episodic training. Meta-learning of an FSS model is typically formulated in terms of *episodic training*. In episodic training, the meta model is trained over a series of training batches consisting of support set S and query set Q examples, sampled from D_{train}. On each query example with some class y, the corresponding binary mask is estimated using the meta model according to the support samples provided for the same class. The meta model is iteratively updated over the episodes to minimize a semantic segmentation loss, *e.g.* pixel-wise cross entropy loss, evaluated on the query samples. Once the training is over, the model is tested by sampling random episodes from D_{test}, where the groundtruth masks for the support samples of novel classes are provided as one(few)-shot guidance to the meta model and those of the query samples are used only for evaluating the resulting FSS outputs on the corresponding queries.

Self-supervised few-shot segmentation. We now define the self-supervised few-shot segmentation problem, using the same notation as above. Similar to the standard FSS problem, in self-supervised FSS, we are given datasets D_{train} and D_{test}. Unlike standard FSS, however, here D_{train} consists of only unlabeled images, with no masks or class labels. Therefore, it is not immediately clear how to define an episodic training procedure, as we can neither provide support samples with class-specific masks, nor sample support and query image pairs from D_{train} such that the support and query images are known to belong to the same class. Once the model is trained, we use the same evaluation protocol of standard FSS to evaluate the learned meta model on large number of few-shot segmentation tasks.

In this work, we intentionally focus on the one-shot segmentation problem to study the problem isolated from the orthogonal concerns regarding the fusion of guidance provided by multiple support samples during evaluation.

3.2. Proposed approach

Saliency-driven self-supervised training. As explained above, in the proposed self-supervised FSS problem, the training set lacks masks and labels. To address the first problem, *i.e.* the lack of groundtruth masks, we propose to use unsupervised saliency to define pseudo groundtruth masks. For this purpose, we adapt the saliency mask estimation approach used in [43]: we train an unsupervised saliency model on the MSRA dataset [7] using the DeepUSPS approach [29]. We then train a BAS-NET model [35] from scratch using the saliency estimations

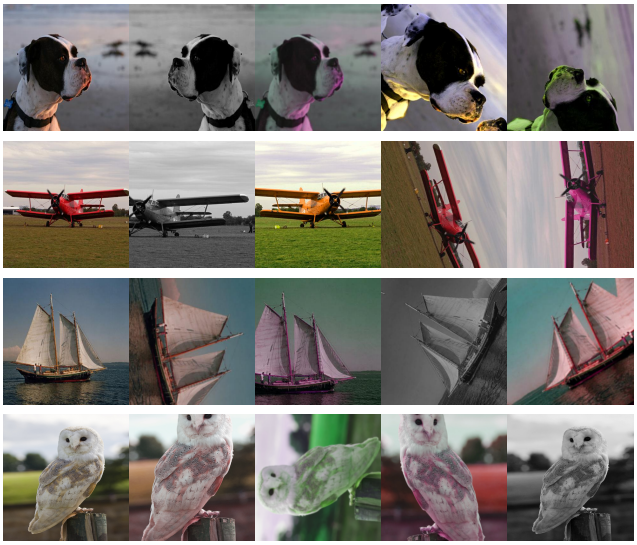

Figure 2: Augmentation samples. The first column show the original images and the remaining four columns contain images with different augmentations.

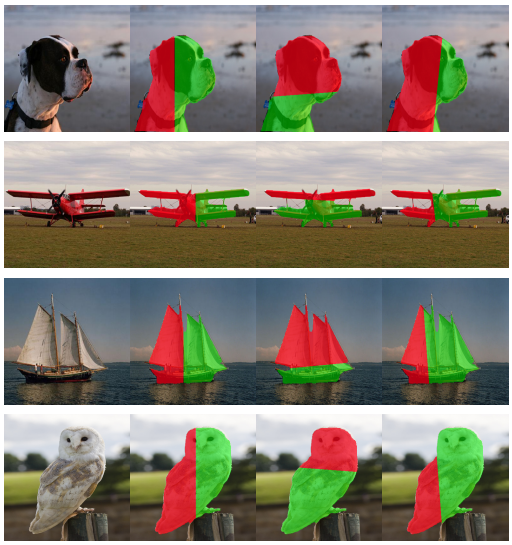

Figure 3: Splitting samples. The images on the first column are original input images and the remaining three columns contain images that demonstrate Vsplit, Hsplit with slope and Vsplit with slope splits overlaid, respectively.

given by the unsupervised saliency model. Then BAS-NET model trained on pseudo-masks is used to obtain object mask proposals from Pascal and COCO datasets.

While the use of unsupervised saliency estimates provides an alternative to manual groundtruth masks, it still does not address the second main problem, *i.e*. the lack of class labels in D_{train}, which are also required to form same class support and query pairs. To remedy this problem, we propose to create support and query pairs from each individual image, by adapting contemporary self-supervised representation learning practices and leveraging the spatial nature of the segmentation task.

More specifically, to define an episode from a single image, we need to create distinct support and query samples. To achieve this, we propose to first apply a set of random augmentations twice to each training image. To this end, we adapt the augmentations used in the SimCLR [4], and utilize the `grayscale`, `color jitter`, `horizontal flip`, `vertical flip`, `rotate` and `random resize` augmentations. We emphasize again that while the SimCLR method aims to learn a global image representation based on contrastive learning over augmented image patches, our goal is to episodically meta-learn a one-shot segmentation model by forming support and query pairs that differ significant enough from each other. As can be observed in Figure 2, the utilized augmentations are able to produce a wide range of variants of a single image.

While augmentations are effectively used in self-supervised learning of image-wide representations, here we target a more structured and arguably detailed task, *i.e*. to learn the few-shot learning meta-model for producing pixelwise predictions. Hence, we seek for more powerful ways to construct training pairs. For this purpose, we propose a method that we call *MaskSplit* in general. The main idea is rooted in the observation that different parts of a single object visually differ significantly. Based on this insight, we propose to split each saliency mask approximately in half, randomly treat one side as the *support* foreground mask and the other side as the *query* foreground mask. By using these masks over the separately augmented versions of an image, we effectively construct support and query pairs, and by using these pairs episodic training can be formed with no groundtruth masks or class labels (Figure 1).

We consider three main variants of the MaskSplit framework: vertical splitting (*Vsplit*), horizontal splitting (*Hsplit*), and their combination (*MixedSplit*). To avoid potential unwanted biases caused by always using axis-aligned mask splits, we split masks along an oriented line. We refer to the basic axis-aligned versions of these schemes as splitting *without slope*. We present visual examples corresponding to Vsplit without slope, Hsplit with slope and Vsplit with slope in Figure 3, and provide more technical details in the following paragraphs.

Vsplit. To divide images in half using Vsplit, we find the line l_0 parallel to y-axis of the image such that l_0 divides the foreground pixels in half. We then assign right side to be the support's foreground and the left side to be the query's foreground. To obtain more and different combinations of this splitting procedure we propose *alternating* vertical split with slope. Here, alternating means that randomly assigning left or right side of saliency mask to support and query

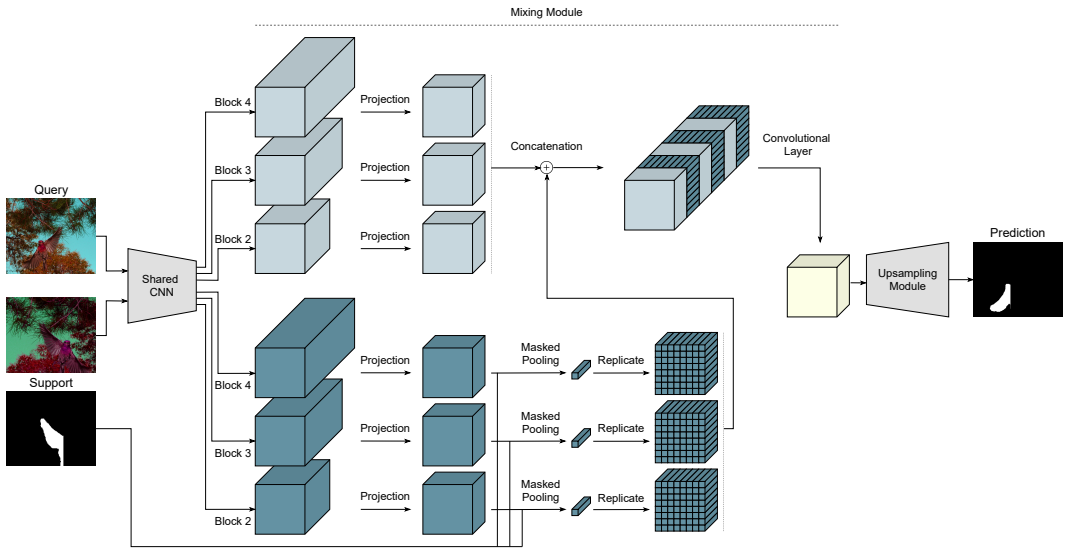

Figure 4: Architecture of our prototype based network, trained with the proposed self-supervised meta-learning approach.

instead of assigning them to same-side in every episode. In order to split the foreground with slope, we shift the line by a number of pixels on both top of the image and the bottom of the image. The shift operation is done in opposite directions for the top and bottom intersection of the image with the line. The default shift range is $(-40, 40)$ in our experiments.

Hsplit and MixedSplit. We also perform experiments with horizontal splitting with slope, where we find the line l_0 parallel to x-axis instead of y-axis. The remaining procedure is the same as in Vsplit. In MixedSplit, we apply Vsplit to the fifty percent of the episodes and apply Hsplit to the rest.

An important implementation detail is the definition of areas over which the loss function is evaluated. More specifically, it shall be observed that part of the non-foreground region in a query image corresponds to the support pixels. Therefore, a naive implementation of the query loss might enforce negative predictions over those support pixels on the query image. To avoid this problem, we ignore the query pixels corresponding to the support area by not calculating the loss function in those background pixels.

3.3. Network architecture

We define a simple prototype-based model inspired from the FSS architecture of [54] to explore different self-supervised strategies. Our network consists of three parts, a **backbone network** to extract features from query and support images, a **mixing module** where support and query features are mixed together, and an **upsampling network** where the mixed features are used to predict the segmentation mask. Figure 4 illustrates the structure of our network.

Backbone. We use ResNet-101 [15] as our backbone to extract features from query and support images. ResNet-101 consists of 4 main blocks that are composed of several convolutional layers. It has been shown that in the first layers of a ConvNet, the model focuses on low-level geometrical features, whereas final layers focus on higher-level semantic features [51]. In our work, we want to take advantage of both geometric and semantic cues by using all blocks, except $block1$. Normally, after each block, the spatial resolution is decreased by strided convolutions. To preserve the spatial resolution we use dilated convolutions [2] instead of strided convolutions after $block2$ so that all feature maps after block2 has a fixed size of 1/8 of the original image. To be able to compare to the existing work on FSS, we use a backbone model pretrained on the ImageNet dataset [38].

Mixing module. To mix different levels of cues we propose a feature mixing module to both mix semantic and geometric features and shape the query features for the prediction using support prototype. The mixing module consists of two sub-modules: projection module to project channel dimensions of different level of features onto a fixed channel size and shared convolution module to mix support-aware query features. There are three projection layers for each block, respectively $block2, block3, block4$. After both query and support features are projected onto a fixed channel dimension, for each different feature level, the query features are concatenated with their corresponding global-average pooled support prototypes. Then all these different levels are concatenated into a feature map and passed onto a shared convolution layer. The output of this module is used by the upsampling module to make predictions.

Upsampling Module. We use a simple upsampling module to decode support-aware query features into a segmentation

mask. It consists of 4 blocks, the first 3 blocks composed of 2 convolutional layers and a bilinear upsampling operation. The last block is composed of one 3×3 convolutional layer followed by a 1×1 convolutional layer.

4. Experiments

Datasets. To evaluate our approach, we use two standard few-shot semantic segmentation datasets: PASCAL-5^i and COCO-20^i. Pascal-5^i dataset is proposed in OSLSM [39] and is based on a combination of PASCAL VOC 2012 [10], and the extra annotations from [14]. This dataset contains 20 classes that are evenly divided into four folds. In the literature, for each fold i, the other three folds are used to train the model, and fold i is used as the target fold. However, in the case of unsupervised meta learning, it is not necessary to divide the images into groups based on their class information. We can utilize the entire training data without any class information to learn our self-supervised representation. Since the training relies only on unsupervised extraction of saliency maps, this enables us to use the images from all folds to train our model. We evaluate our model with and without fold-specific training.

The second dataset, COCO-20^i, is proposed by FWB [30] and is based on COCO dataset [24]. This dataset consists of 80 classes, divided evenly into four folds. Our results for this dataset are obtained by unsupervised training on images from all folds. For both datasets, we use mean intersection-over-union (mIoU) and report one-shot segmentation results in all experiments.

Implementation Details. For the backbone network we use ResNet-101. We extract features from $block2, block3, block4$, and then these features are passed onto the mixing module to mix query and support features at different levels. Both query and support images are resized to a fixed spatial size of 400×400 and features extracted from the ResNet-101 backbone have a fixed size of 50×50. The overall architecture is implemented in PyTorch and PyTorch-Lightning. We use Adam optimizer with a learning rate of 10^{-4} and weight decay 10^{-5} with pixel-wise cross-entropy loss. All layers of ResNet-101 are kept frozen. We train each model for 100 epochs on PASCAL dataset and 20 epochs on COCO dataset with a batch size of 16, on a single Nvidia V100 GPU. To test the models, we use 5 runs with 2500 tasks each.

4.1. Ablation Study

We first conduct extensive ablative studies to understand the effects of major components of our approach such as `Alternate`, `Slope`, `Vsplit` and `Aug`, on the Pascal-5^i dataset. We also experiment with different splitting techniques `Hsplit`, `MixedSplit` and also with no split. Finally, we show the results of different supervision levels.

MaskSplit variants. Here, we look at the effect of different components of our approach. These components are augmentations (denoted with `aug`), slope (denoted with `slope`), alternating support and query sides (denoted with `alternate`), probability of applying the split (denoted with `prob`), the type of split (vertical split `Vsplit` or horizontal split `Hsplit`). When splits are are applied with 0.3 probability, we apply MaskSplit 30 percent of the time for the given configuration, and for the remaining, we only apply augmentations and use the saliency mask as both support's and query's foreground.

The results are summarized in Table 1. The first row of Table 1 corresponds to the case when SimCLR [4] augmentations are used together with our base prototype network. This self-supervised version, makes use of only the SimCLR augmentations instead of MaskSplit based training. This yields a mIoU of 48.6 on average. When we do not use any augmentations, but use the Vsplit, we see that the average mIOU score is 47.8. When Vsplit is used together with the augmentations (using `aug`, `slope`, `alternate` components), we achieve a average mIoU of 53.0. This proves that augmentations are necessary to create visually different query and support examples, and including different augmentations within the self-supervised learning procedure enhances the training quite significantly. This significant performance increase from 48.6 to 53.0 also indicates that proposed splitting mechanism is effective for the self-supervised segmentation task.

We also observe in Table 1 that removing slope option causes the mIoU to drop by 2 points, which means that slope improves the possibility of creating variety of segments in episodes. Making Vsplit probabilistic reduces mIoU approximately one percent. mIoU scores show that applying Vsplit only thirty percent of the time is also effective, if not as effective as applying it all the time.

Lastly, removing the `alternate` option yields only 0.3 point reduction in average mIoU. Here, applying augmentations probably render the sides of the splits to be different enough at each episode. Therefore, not alternating between sides does not seem to have a major effect; nevertheless, it still is a useful part of the procedure.

Different Split Techniques. Table 1 also presents the results with using different split techniques. When no split is used, there is a decrease of almost 4.5 percent in mIoU. This result highlights that proposed splitting procedure adds significant recognition power to the self-supervised process. The comparison between results of `Hsplit` and `Vsplit` show that the positioning of objects in the images are most appropriate to be used with `Vsplit`. For the rest of the experiments, we use MaskSplit with `Vsplit`, `aug`,`slope`,`alternate` options, which achieves the best performance in these ablation studies.

aug	slope	alternate	Vsplit	Hsplit	prob	5^0	5^1	5^2	5^3	mean
✓					0.0	51.5	51.1	52.1	40.0	48.6
	✓	✓	✓		1.0	50.7	50.1	49.7	40.8	47.8
✓		✓	✓		1.0	53.2	54.5	54.0	42.4	51.0
✓	✓	✓		✓	1.0	54.9	55.9	54.7	43.9	52.3
✓	✓	✓	✓	✓	1.0	54.5	55.5	52.5	42.7	51.3
✓	✓		✓		1.0	53.7	57.1	55.4	44.7	52.7
✓	✓	✓	✓		0.3	**54.4**	54.2	**55.4**	43.7	51.9
✓	✓	✓	✓		1.0	54.1	**57.1**	54.9	**46.1**	**53.0**

Table 1: Ablation study of our proposed MaskSplit framework. We report the mIoU scores achieved by different variants of our model on the PASCAL dataset. `Prob` stands for the probability of applying the selected image splitting method.

supervision	mask source	train	5^0	5^1	5^2	5^3	avg
Self-sup.	groundtruth	fold	49.4	52.8	40.6	37.6	45.1
Self-sup.	saliency	fold	51.5	55.2	52.5	44.4	50.9
Supervised	groundtruth	fold	54.9	65.4	47.9	42.2	52.6
Self-sup.	saliency	all	54.1	57.1	54.9	46.1	53.0
Self-sup.	groundtruth	all	59.0	59.0	62.5	49.0	57.3

Table 2: Comparison of different supervision levels and effect of using fold-based training vs all train set training using the PASCAL dataset. For supervised, we use our base prototype model trained in the standard supervised few-shot setting, using the groundtruth masks from the training folds.

Different Supervision Levels. We conduct further experiments to see the effect of using different supervision levels and different amount of training data. The first two rows of Table 2 compare the results of MaskSplit using regular fold-based training; *i.e*. using groundtruth masks during training vs masks obtained in the unsupervised fashion using saliency. Since the number of training images is relatively low in fold-based training, using groundtruth masks only is limiting the context information learned for the self-supervised learner. On the contrary, saliency maps could also include objects that are not in the groundtruth masks. As a result, self-supervised learning appears to be positively affected by the increased variety.

For the supervised counterpart, we train our base prototype network using traditional fold-based few-shot semantic segmentation setting. We observe that this supervised training of the model yields less superior results compared to the self-supervised version that is trained over the entire set of training images (52.6 mIoU vs 53.0 mIoU). The self-supervised nature of training enables us to use all the training images, and this leads to a better learning of segmentation in an unsupervised way.

The last row of Table 2 shows the oracle results of our model, which uses ground truth segmentation maps instead of masks extracted by unsupervised saliency model. This gives us a upper limit on the performance of our proposed framework. The results demonstrate that when we use groundtruth masks instead of saliency, 57.3 percent mIoU score can be achieved.

Method	5^0	5^1	5^2	5^3	avg
Supervised meta-learning (upper-bounds) with ResNet50					
PANet [45]	44.0	57.5	50.8	44.0	49.1
PGNet [53]	56.0	66.9	50.6	50.4	56.0
PFENet [42]	61.7	69.5	55.4	56.3	60.8
SCL(PFENet) [52]	63.0	70.0	56.5	57.7	61.8
RePRI [1]	59.8	68.3	62.1	48.5	59.7
SAGNN [46]	64.7	69.6	57.0	57.2	62.1
CMN [47]	64.3	70.0	57.4	59.4	62.8
Supervised meta-learning (upper-bounds) with ResNet101					
FWB [30]	51.3	64.5	56.7	52.2	56.2
PPNet [26]	52.7	62.8	57.4	47.7	55.2
DAN [44]	54.7	68.6	57.8	51.6	58.2
MLC [49]	60.8	71.3	61.5	56.9	62.6
HSNet [27]	67.3	72.3	62.0	63.1	66.2
Unsupervised approaches					
Saliency* [43]	51.5	49.1	48.1	39.0	46.9
MaskContrast* [43]	53.6	50.7	50.7	39.9	48.7
Ours	**54.1**	**57.1**	**54.8**	**46.1**	**53.0**

Table 3: mIoU scores produced by our method on PASCAL dataset in comparison with two unsupervised approaches and the supervised state-of-the-art few-shot semantic segmentation models. (*) corresponds to the results acquired by adapting the methods to FSS evaluation.

4.2. Comparison to existing work

On Table 3 and Table 4 we compare our approach with the state-of-the-art supervised few-shot segmentation and unsupervised semantic segmentation approaches. The supervised state-of-the-art models are trained on the usual fold-based training using groundtruth segmentation masks, whereas unsupervised models are trained using all the training images with no groundtruth.

In the bottom part of Table 3, we present comparisons to the unsupervised semantic segmentation methods. First is unsupervised saliency, for which we use the version that is optimized by [43]. In evaluation, for each test episode, we take the IoU of the proposed saliency mask with the ground truth. Second unsupervised approach is the recent state-of-the-art unsupervised semantic segmentation method, namely MaskContrast [43]. We use the publicly available model, which was initialized using MoCo v2 [6]. To obtain the few-shot results, we take the masks produced

after the k-means clustering. These masks contain cluster assignments instead of ground truth classes. [43] use the Hungarian matching algorithm to match ground truth classes with cluster assignments. However, this is not directly comparable with unsupervised few-shot segmentation methods, since the Hungarian matching algorithm requires the use of ground truth masks of validation set images. Additionally, episode creation in few-shot segmentation causes a different test set distribution. In order to make all methods directly comparable, we experiment with two different settings: (i) For each episode, we compare the cluster assignments in query and support masks. If they match, we take the IoU of the mask and the ground truth. This first evaluation yields 25.9 mIoU. (ii) For each episode, we take the IoU of the mask and the ground truth without requiring cluster assignments in support and query to match. This second evaluation results in a mIoU of 48.7. We use this second favorable result for comparison. According to the results in Table 3, our model outperforms both baselines by at least a margin of 4% on Pascal-5^i dataset.

The comparisons to supervised approaches in Table 3 show that the proposed self-supervised approach performs comparable to several recent supervised approaches, and the performance gap against the state-of-the-art is arguably not drastic. The results highlight the potential of the self-supervised learning of FSS models.

In Table 4, we also compare our model with a baseline model and state-of-the-art few-shot segmentation models on COCO-20^i. We do not report any result for MaskContrast, since there is not a publicly available model that is trained on COCO dataset. Our model is again able to outperform the saliency method, yet the performance increase is relatively lower. We think that this is due to the larger amount of noise in the generated saliency masks on COCO images.

We have also tried to adapt the self-supervised super-pixel based training strategy from [33], originally proposed for medical images. At each training image, we extract super-pixels via [11], randomly select a super-pixel, and apply augmentations using our pipeline to obtain query and support regions. Despite our efforts, however, we have not been able to get a meaningful baseline in our setting.

In Figure 5, we present some qualitative results that demonstrate the challenges of the task and overall success of the proposed MaskSplit approach.These results show that our method is able create accurate few-shot segmentation results, even when the saliency maps significantly differ from the groundtruth masks. In addition, the presented results also illustrate the importance of various components of our approach.

5. Conclusion

In this work, we define and study the problem of unsupervised few-shot semantic segmentation. Our work aims to remove the need for supervised segmentation examples for meta model training and enable utilization of arbitrarily large unsupervised image collections. We propose a novel self-supervised way to create training episodes, which is based on unsupervised saliency and augmentations. Extensive experiments show that this setting is able to achieve few-shot generalization and we obtain significant performance improvements over our baselines. We believe that our work will stimulate further study on unsupervised learning of few-shot segmentation models.

Method	20^0	20^1	20^2	20^3	avg
Supervised meta-learning (upper-bounds) with ResNet50					
PPNet [26]	28.1	30.8	29.5	27.7	29.0
PFENet [42]	36.5	38.6	34.5	33.8	25.8
RePRI [1]	32.0	38.7	32.7	33.1	34.1
HSNet [27]	36.3	43.1	38.7	38.7	39.2
CMN [47]	37.9	44.8	38.7	35.6	39.3
Supervised meta-learning (upper-bounds) with ResNet101					
FWB [30]	17.0	18.0	21.0	28.9	21.2
DAN [44]	-	-	-	-	24.4
MLC [49]	50.2	37.8	27.1	30.4	36.4
SAGNN [46]	36.1	41.0	38.2	33.5	37.2
PFENet [42]	36.8	41.8	38.7	36.7	38.5
HSNet [27]	37.2	44.1	42.4	41.3	41.2
Unsupervised approaches					
Saliency* [43]	**22.7**	24.3	20.4	22.2	22.4
Ours	22.3	**26.1**	**20.6**	**24.3**	**23.3**

Table 4: Comparison of mIoU scores produced by our method and the state-of-the-art models on COCO dataset. (*) corresponds to the results acquired by adapting the methods to FSS evaluation.

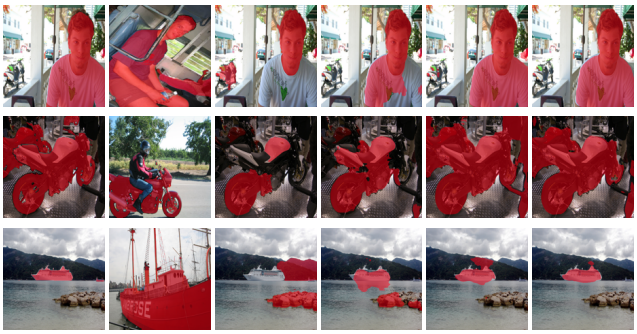

Figure 5: Qualitative results. Columns correspond to query, support, saliency map, result when trained using no augmentations, result when trained without splitting, and result of the proposed MaskSplit approach, respectively.

6. Acknowledgement

This work was supported in part by the TUBITAK Grant 119E597.

References

[1] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need? *arXiv e-prints*, page arXiv:2012.06166, Dec. 2020.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834—848, April 2018.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 833–851, Cham, 2018. Springer International Publishing.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255, 2020.

[6] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[7] Ming Ming Cheng, Guo Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi Min Hu. Global contrast based salient region detection. In *2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 409–416, United States, 2011. IEEE Computer Society.

[8] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[9] Nanqing Dong and E. Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, 2018.

[10] M. Everingham, S. Eslami, L. Gool, Christopher K. I. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2014.

[11] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[12] J. Fu, J. Liu, Haijie Tian, Z. Fang, and H. Lu. Dual attention network for scene segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3141–3149, 2019.

[13] Golnaz Ghiasi and Charless C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 519–534, Cham, 2016. Springer International Publishing.

[14] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 297–312, Cham, 2014. Springer International Publishing.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[16] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *International Conference on Learning Representations*, 2019.

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, Humphrey Shi, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 603–612, 2019.

[18] Xu Ji, A. Vedaldi, and João F. Henriques. Invariant information clustering for unsupervised image classification and segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9864–9873, 2019.

[19] Siavash Khodadadeh, Ladislau Bölöni, and Mubarak Shah. Unsupervised meta-learning for few-shot image and video classification. *CoRR*, abs/1811.11819, 2018.

[20] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, Vancouver, Canada, Apr. 2018.

[21] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequence. In *IEEE International Conference on Computer Vision*, 2017.

[22] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. *CoRR*, abs/2104.01893, 2021.

[23] Guosheng Lin, A. Milan, Chunhua Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5168–5177, 2017.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[25] W. Liu, Andrew Rabinovich, and A. Berg. Parsenet: Looking wider to see better. *ArXiv*, abs/1506.04579, 2015.

[26] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. *CoRR*, abs/2007.06309, 2020.

[27] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. *CoRR*, abs/2104.01538, 2021.

[28] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. In *ECCV*, 2016.

[29] Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction with self-supervision. *CoRR*, abs/1909.13055, 2019.

[30] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 622–631, 2019.

[31] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[32] Yassine Ouali, Celine Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020.

[33] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. *CoRR*, abs/2007.09886, 2020.

[34] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[39] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. 2017.

[40] Evan Shelhamer, J. Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:640–651, 2017.

[41] Mennatullah Siam and Boris N. Oreshkin. Adaptive masked weight imprinting for few-shot segmentation. *CoRR*, abs/1902.11123, 2019.

[42] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *CoRR*, abs/2008.01449, 2020.

[43] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *International Conference on Computer Vision*, 2021.

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

[45] K. Wang, J. Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9196–9205, 2019.

[46] Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5475–5484, June 2021.

[47] Guo-Sen Xie, Huan Xiong, Jie Liu, Yazhou Yao, and Ling Shao. Few-shot semantic segmentation with cyclic memory network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7293–7302, October 2021.

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

[49] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. Mining latent classes for few-shot segmentation. *CoRR*, abs/2103.15402, 2021.

[50] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[51] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[52] Bingfeng Zhang, Jimin Xiao, and Terry Qin. Self-guided and cross-guided learning for few-shot segmentation. *CoRR*, abs/2103.16129, 2021.

[53] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[54] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5212–5221, 2019.

[55] H. Zhang, K. Dana, J. Shi, Zhongyue Zhang, Xiaogang Wang, A. Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[56] H. Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 548–557, 2019.

[57] Hengshuang Zhao, J. Shi, Xiaojuan Qi, Xiaogang Wang, and J. Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017.

[58] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[59] Kai Zhu, Wei Zhai, Zheng-Jun Zha, and Yang Cao. Self-Supervised Tuning for Few-Shot Segmentation. *arXiv e-prints*, page arXiv:2004.05538, Apr. 2020.