Elimination of Non-Novel Segments at Multi-Scale for Few-Shot Segmentation

Alper Kayabaşı[1,2], Gülin Tüfekci[1,2], İlkay Ulusoy[2]
[1]Research Center, Aselsan Inc
[2]Middle East Technical University, Ankara, Turkey
{alper.kayabasi,gulin.tufekci,ilkay}@metu.edu.tr

Abstract

Few-shot segmentation aims to devise a generalizing model that segments query images from unseen classes during training with the guidance of a few support images whose class tally with the class of the query. There exist two domain-specific problems mentioned in the previous works, namely spatial inconsistency and bias towards seen classes. Taking the former problem into account, our method compares the support feature map with the query feature map at multi scales to become scale-agnostic. As a solution to the latter problem, a supervised model, called as base learner, is trained on available classes to accurately identify pixels belonging to seen classes. Hence, subsequent meta learner has a chance to discard areas belonging to seen classes with the help of an ensemble learning model that coordinates meta learner with the base learner. We simultaneously address these two vital problems for the first time and achieve state-of-the-art performances on both PASCAL-5^i and COCO-20^i datasets.

1. Introduction

Semantic segmentation is a crucial task that classifies each pixel of an image to make sense of the scene with application areas such as autonomous driving [5] and medical imaging [19]. Deep learning pervades semantic segmentation like other tasks of computer vision [2, 15]. Supervised segmentation models are required to employ abundant annotated data belonging to each class in the training set since the generalization capacity of supervised models decreases with scarce labeled data. Therefore, adapting the model to work on unseen classes requires dense annotation of myriad data from novel classes. Shaban *et al*. [20] proposed few-shot segmentation to remove the labeling effort and increase the generalization capacity of a model given few data for the first time.

Few-shot segmentation addresses the problem of making pixel-wise predictions for a target image, called a query, from an unseen class with the guidance of a support image

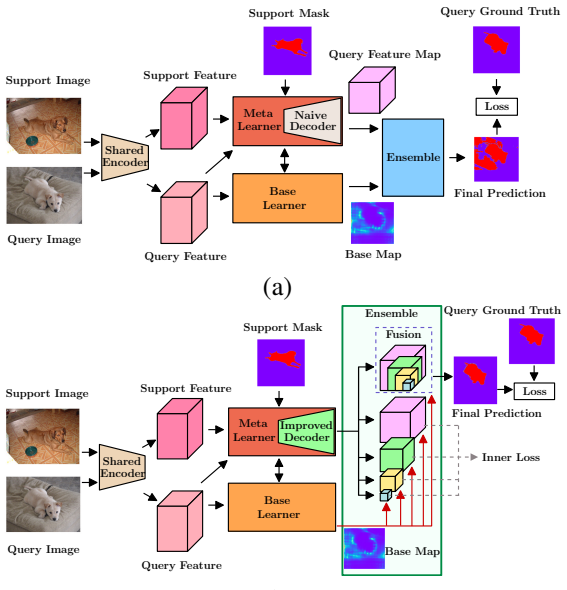

Figure 1. (a) Overview of BAM [10]. Support and query features are used by meta learner to extract support feature map while base learner provides guidance for the base classes and leads the meta learner to focus on novel regions via ensembling. (b) Our proposed method. The decoder for meta learner is improved such that query feature map is obtained at multi-scale. Support feature map is compared with query feature maps at multi-scale to obtain enriched query features. Query predictions obtained from enriched query features at each scale are ensembled with the base map as well as the prediction obtained from the fusion of them. Inner losses are computed at different scale levels and the final prediction is obtained from the ensemble of the base map with the predictions from the fused query feature maps. (Best viewed in zoom)

from the same category. Inspired by the few-shot classification task [23], most methods utilized the episodic training strategy in which the gradients are averaged over tasks named as an episode. Each episode is sampled from a dataset whose classes are disjoint from a test case where only a few data are available. These episodes are used to imitate the test case during training to prevent overfitting. De-

spite this intention, a model trained with this strategy tends to mistake segments from seen training classes, referred as *base classes*, as *novel classes* because of constantly experiencing the same set of classes during training. Hence, the co-occurrence of novel and base classes in the same scene causes entanglement between features of pixels that are part of the novel and base categories.

To prevent this entanglement, prediction of the supervised model which is trained on base classes guides the meta learner, which is responsible for detecting novel areas. Meta learner is directed to areas not occupied by the base classes so that contradiction between the base learner and the meta learner is avoided via the ensemble model entailing both base and meta predictions [10]. On the side of meta learner, candidate objects in query image might not cover as same area as those in support images; so, the model should compare support feature map with query feature map at different resolutions to disentangle adjacent regions around novel segments [22]. As shown in Fig. 1, the ensemble of base and meta learner without improved decoder fails to distinguish background from foreground since naive decoder, which is designed for the supervised scheme, lacks to combine features at different resolutions in favor of complete query prediction. Hence, we transform the naive decoder into an improved decoder such that not only does it correlate the support image with the query image at multi-resolution but also it benefits from merits of base learner at multi-resolution. In this regard, we hypothesize that there are cases where it is not enough that base learner discourages meta learner from base regions at single-scale. Our experiments verify that the improved decoder and ensembling the predictions at multi-scale outperform the decoder equipped with ensembling the prediction at single-scale. Our contributions in this paper are two-fold:

- We alleviate the spatial inconsistency and the bias problems together with the assistance of our proposed decoder that seeks to remove bias at multi-resolution.
- Our proposed method achieves new state-of-the art performance on both PASCAL-5^i (mIoU @ 1-shot: 68.59%, mIoU @ 5-shot: 72.05%) and COCO-20^i (mIoU @ 1-shot: 47.16%, mIoU @ 5-shot: 52.50%) datasets for few-shot segmentation task.

2. Related work

2.1. Semantic segmentation

Fully convolutional neural network (FCN), which is the pioneering work in semantic segmentation field, formulates semantic segmentation as a pixel-wise classification task [15]. In FCN, all fully connected layers at the end of a model are transformed into convolution layers so that the network accepts arbitrary input sizes. Success of FCN accelerates the field and results in outstanding architectures such as UNet [19], PSPNet [30], and Deeplab [2, 3]. PSPNet combines average pooled feature maps at different scales to contain not only global but also local context [30]. Deeplab introduces ASPP module [2] equipped with dilated convolution that increases the receptive field of the network without a decrease in resolution by inserting holes between filter weights.

2.2. Few-shot segmentation

Few-shot segmentation studies can be categorized into four groups according to their objectives: imbalance in details, inter-class gap, spatial inconsistency, and correlation reliability.

The imbalance in details problem emphasizes that there might be details that do not co-exist in both query and corresponding support image. Hence, inconsistent regions between support and query should be detected and eliminated on the support side to prevent redundant details or noise in an adaptive manner. PGNet [28] proposes a network that associates each query pixel to the relevant parts of the support image to remove noise, where relevancy is generally quantified by a similarity metric such as cosine similarity. PANet [24] adds regularization loss to ensure that the network becomes successful if the roles of support and query are swapped. ASGNet [11] aims to find an adaptive number of prototypes and their spatial extents based on image content with a boundary-aware superpixel algorithm so that prototypes represent parts of an object with similar characteristics. Each query pixel utilizes support prototype giving maximum cosine similarity with itself as reference.

Most approaches assume that transferable knowledge constantly exists in base set and tries to segment image from unseen classes during training. This strong assumption loses its validity in proportion to the discrepancy between base and novel dataset, and this problem is referred as an inter-class gap. RePRI [1] shows that the severity of overfitting is exaggerated in few-shot segmentation and adapting novel classes by fine-tuning over support images improves segmentation performance. CWT [16] episodically trains self-attention block that adapts classifier weights of network updated over support image during both training and test. BriNet [27] regards prediction for query as pseudo-mask and switches the roles between query and support to update model with ground truth support mask during test stage until determined mIoU threshold is exceeded for support mask and its prediction.

Architectures designed for supervised cases fail to provide scale-invariance in few-shot scenarios since contextual relationships are not figured out by a handful of data. Methods design control mechanism that provides favorable information exchange between different resolutions [25, 26, 22, 28].

Correlation map determines pixel-wise similarity between support and query images. Problems such as background clutter and occlusion lead to noise in the correlation map; hence, it results in erroneous comparisons and training processes based on misinterpreted correspondences, which is called as a hyper correlation reliability problem. Methods are proposed to check the validity of correspondences based on learnable or engineered criteria so that filtered correlation maps become interpretable. After the elimination of deceptive correspondences, all similarities corresponding to each query pixel from the support image are summed to obtain an activation score that determines the level of association of that query pixel with the foreground of support [17, 8].

3. Preliminaries

In order to relate the specified challenges in few-shot segmentation with our approach and guide the reader through steps, the task is formally defined in Section 3.1, Feature Enrichment Module (FEM) [22] and Base and Meta Learner (BAM) [10] are introduced in Sections 3.2 and 3.3, respectively.

3.1. Task description

Few-shot segmentation task utilizes base dataset containing adequate images with their annotations whose classes, $\mathcal{C}_{base}$, are disjoint from novel classes, $\mathcal{C}_{novel}$, in which dense predictions are fulfilled with few data and their annotations. K number of available data and their annotations belonging to the novel classes constitute a support set $\mathcal{S}$ for testing which are expected to guide a model M to make predictions for query image $\mathcal{I}_q$, which is dubbed as K-shot segmentation. The support set is formally represented as $\mathcal{S} = \{\mathcal{I}_{s_i}, \mathcal{M}_{s_i}\}_{i=1}^{K}$, where $\mathcal{I}_{s_i}$, and $\mathcal{M}_{s_i}$ correspond to i$^{\text{th}}$ support image and its dense ground truth mask. On the side of training, support set for training is sampled from base dataset along with query set which consists of the query image and its ground truth, sharing its class with the chosen support set. The aforementioned classes are treated as novel class during training in order to perform episodic training, where pixels belonging to chosen class are assigned as foreground while pixels from all other classes are considered as background. Query set is formally represented as $\mathcal{Q} = \{\mathcal{I}_q, \mathcal{M}_q\}$, where $\mathcal{I}_q$ and $\mathcal{M}_q$ correspond to the query image and its dense ground truth mask. The model, M, is trained by backpropagating binary cross entropy loss between $\mathcal{I}_q$ and $\mathcal{M}_q$ over tasks, named as episodes involving the selected support set from base dataset with the accompanying query set.

3.2. Revisit feature enrichment module

Multi-scale modules in supervised semantic segmentation generally do not provide mechanism to form independent interaction between masked global average pooled support feature map, called as support prototype, and average pooled query feature map at different scales. For example, conventional multi-scale architectures apply single filtering to combination of query feature map, support prototype, and prior mask that describes likelihood of query pixel being related with at least one pixel in foreground of support [22]. Different from these approaches, inter-source enrichment module of FEM separately applies the filtering to the query feature map at each different scale, which is combined with support prototype and prior mask. Furthermore, inter-scale interaction module of FEM fulfills the information transfer between two consecutive resolutions in top-down path, where top-down path consists of outputs of inter-source enrichment module ordered from high resolution to low resolution. During information transfer, preservation of hierarchical structure allows gradual accumulation of information from higher resolution to lower resolution. In this module, each resolution has direct connection only to its neighbour in the top-down direction. Therefore, there is no connection between any resolution pairs other than the consecutive ones. Hence, the module has a chance to decide on the scale at which the obtained information is sufficient to make a prediction and the following scales would bring redundancy. Following this reasoning, feature maps at different resolutions are fused via information concentration module in FEM.

3.3. Revisit base and meta learner

Typical few-shot segmentation approaches use meta-learning approach such that the knowledge gained from training the model on the base classes is utilized to predict the mask of the query image belonging to a novel class given a support image belonging to the same novel class. This process is called as meta-learning since learning tasks are sampled from the base classes during training in order to simulate the few-shot settings in testing so that the training and testing conditions are matched. However, as [10] states, training on base classes introduce a bias towards them during testing, which prevents the model to work on the novel classes properly. To tackle this bias BAM is introduced, where a base learner, apart from the meta learner, explicitly works on the known classes. When the information related to known classes is used during testing, the recognition of novel classes would be enhanced.

Training BAM consists of two stages, namely base-training and meta-training. Both learners share the same backbone as feature encoder. To leverage the representations at different levels of abstraction, features are obtained from different layers of the encoder. Base learner is trained in a supervised manner so that the ability to make confident predictions regarding base classes is gained. In meta-training stage, the parameters of the base learner are fixed.

The features of support and query images are extracted by the shared encoder and the features obtained after block-2 and block-3 of ResNet-50 [7] are concatenated and transformed with 1×1 convolution layer, which are denoted by $\mathbf{f}_m^s$ and $\mathbf{f}_m^q$ respectively. Query features after block-4 of ResNet-50, $\mathbf{f}_b^q$, are processed by base learner and decoded by Pyramid Scene Parsing Network (PSPNet) [30], which is composed of Pyramid Pooling Module (PPM) and classifier so that the probability map of base classes, $\mathbf{p}_b^f$, is obtained. This step is crucial since the base classes are the background classes for the query image while the novel class is the foreground, which is to be predicted by the meta learner. The support mask, $\mathbf{m}^s$, is used together with $\mathbf{f}_m^s$ in order to obtain the support prototype, $\mathbf{v}_s$. The query features, support prototype and the prior map are concatenated, which is inputted to the meta decoder. At the end of the meta decoder, output background and foreground probability maps, $\mathbf{p}_m^0$ and $\mathbf{p}_m^1$, are obtained.

Low level features for support and query images are obtained from the intermediate levels of the encoder, denoted by $\mathbf{f}_{low}^s$ and $\mathbf{f}_{low}^q$, and the Frobenius norm between their Gram matrices is computed as adjustment factor, ψ. The adjustment factor is leveraged such that the smaller ψ is, the closer the representations of support and query images become. In other words, as ψ gets smaller, the reliability of the prediction of the meta learner increases such that the query features become representative of the support features. Moreover, $\mathbf{p}_m^0$ is ensembled with $\mathbf{p}_b^f$ in order to force the pixels belonging to non-novel regions for the query image to be closer to the base classes. This enhanced information is used such that the corresponding pixels are less likely to be predicted as novel. Resultant ensembled information is concatenated with $\mathbf{p}_m^1$ in order to produce final prediction.

4. Method

As CANet implies [29], we use the middle level features by applying 1x1 convolution to concatenation of feature maps obtained from block-2 and block-3. We represent middle level features belonging to support and querys image respectively as in Eq. 1 and Eq. 2, where Enc symbolizes the middle level feature extractor.

$$\mathbf{f}_m^s = Enc(\mathcal{I}_s) \ \in R^{H\times W\times C} \tag{1}$$

$$\mathbf{f}_m^q = Enc(\mathcal{I}_q) \ \in R^{H\times W\times C} \tag{2}$$

To obtain the prior map in a similar manner to PFENet [22], high level query and support features are reshaped from $\mathrm{R}^{H\times W\times C}$ to $\mathrm{R}^{HW\times C}$ at first. After that, row wise norms for high level query and support pixel features are computed respectively as in Eq. 3 and Eq. 4, where $^\circ$ corresponds to Hadamard root while $diag$ outputs diagonal elements of a matrix as a column vector.

$$\|\tilde{\mathbf{f}}_b^q\| = (diag(\tilde{\mathbf{f}}_b^q \times \tilde{\mathbf{f}}_b^{q\intercal}))^{\circ 1/2} \in R^{HW\times 1} \tag{3}$$

$$\|\tilde{\mathbf{f}}_b^s\| = (diag(\tilde{\mathbf{f}}_b^s \times \tilde{\mathbf{f}}_b^{s\intercal}))^{\circ 1/2} \in R^{HW\times 1} \tag{4}$$

Prior map is calculated by max pooling the cosine similarity matrix between the high level query and support pixels along row wise direction as shown in Eq. 5, where $\oslash$ is Hadamard divison.

$$\mathbf{C}_q = \text{pool}((\tilde{\mathbf{f}}_b^q \times \tilde{\mathbf{f}}_b^{s\intercal}) \oslash (\|\tilde{\mathbf{f}}_b^q\| \times \|\tilde{\mathbf{f}}_b^s\|^\intercal)) \ \in R^{H\times W\times 1} \tag{5}$$

Masked global average pooling is applied to $\mathbf{f}_m^s$ to extract support prototype, $\mathbf{v}_s$, in Eq. 6, where $\mathcal{R}$ downsamples $\mathcal{M}_s$ to the size of $\mathbf{f}_m^s$.

$$\mathbf{v}_s = \text{masked\_avg\_pool}(\mathbf{f}_m^s, \mathcal{R}(\mathcal{M}_s)) \ \in R^{1\times 1\times C} \tag{6}$$

FEM takes $\mathbf{v}_s, \mathbf{C}_q$, and $\mathbf{f}_m^q$ as input and outputs N+1 enriched query feature maps where N of them correspond to enriched auxiliary feature maps at N different scales and the last one is the fusion of them as shown in Eq. 7.

$$\mathbf{X}_q^{s_1}, \mathbf{X}_q^{s_2}, ..., \mathbf{X}_q^{s_N}, \mathbf{X}_q^{fused} = FEM(\mathbf{C}_q, \mathbf{f}_m^q, \mathbf{v}_s) \tag{7}$$

$$\mathbf{C}^{AUX} = \{\mathbf{C}^{aux,1}, \mathbf{C}^{aux,2}, ..., \mathbf{C}^{aux,N}, \mathbf{C}^{aux,fused}\} \tag{8}$$

$\mathbf{C}^{AUX}$ in Eq. 8 represents set of classifiers, where first N classifiers correspond to auxiliary classifiers, which make predictions for the multi-scale features, while the last classifier is responsible for the prediction deduced from the fused feature. By using these classifiers, we obtain background and foreground logit values for enriched query feature maps at each scale and the fused feature map respectively in Eq. 9 and Eq. 11, where $\oplus$ performs concatenation operation.

$$\mathbf{p}_{m,s_i}^0, \mathbf{p}_{m,s_i}^1 = \mathbf{C}^{aux,i}(\mathbf{X}_q^{s_i}) \tag{9}$$

$$\mathbf{p}_{m,s_i} = \mathbf{p}_{m,s_i}^0 \oplus \mathbf{p}_{m,s_i}^1 \tag{10}$$

$$\mathbf{p}_{m,fused}^0, \mathbf{p}_{m,fused}^1 = \mathbf{C}^{aux,fused}(\mathbf{X}_q^{fused}) \tag{11}$$

$$\mathbf{p}_{m,fused} = \mathbf{p}_{m,fused}^0 \oplus \mathbf{p}_{m,fused}^1 \tag{12}$$

$BaseLearner$ in Eq. 13 takes $\mathbf{f}_q^m$ as input and outputs summation of predicted probabilities for all classes except background.

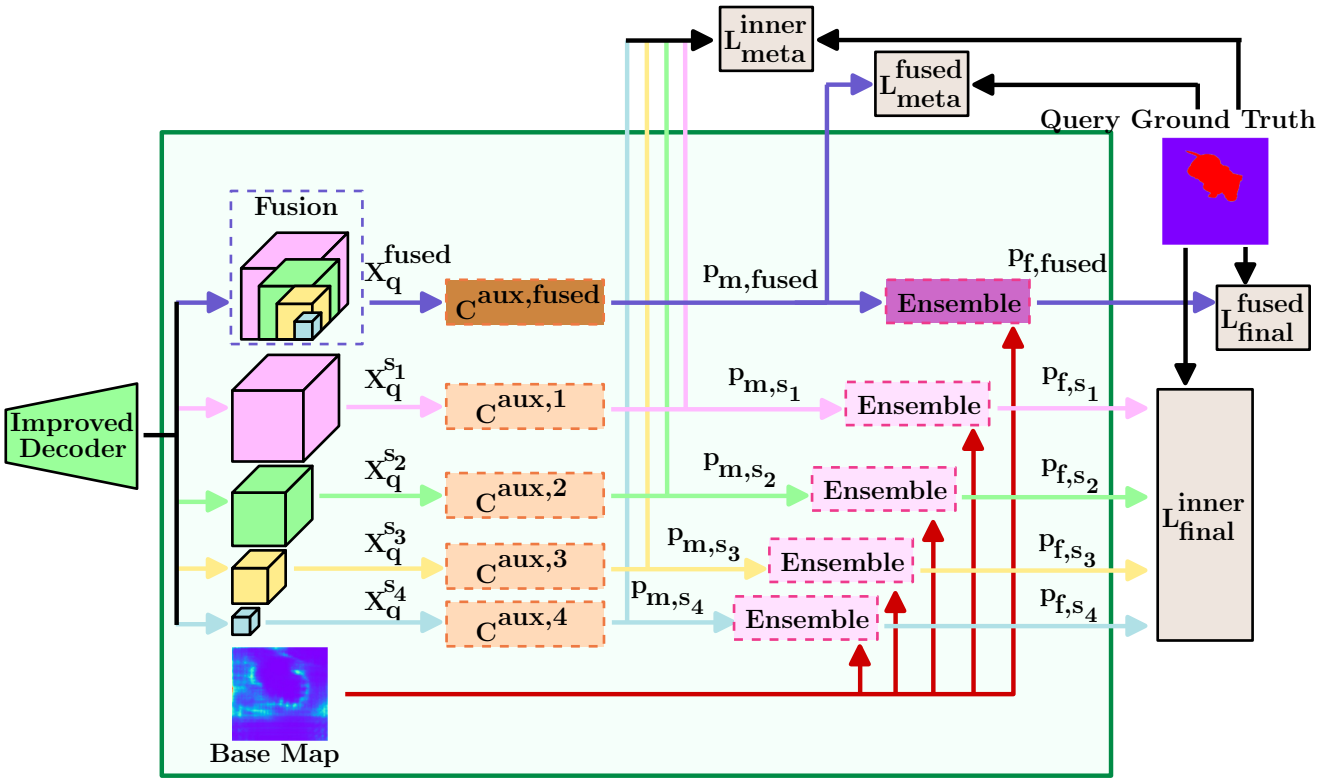

Figure 2. Detailed architecture of the multi-scale ensemble module. Features at multi-scale and the fusion of them are obtained at the end of the improved decoder as $\mathbf{X}_q^{s_i}$ and $\mathbf{X}_q^{fused}$ respectively, which are used by the corresponding auxiliary classifiers. The resultant enriched query feature maps are ensembled with the base map to obtain query predictions at multi-scale, which are denoted by $\mathbf{p}_{f,s_i}$ and $\mathbf{p}_{f,fused}$ respectively. Inner losses are computed from probability maps at intermediate scales $(\mathbf{p}_{m,s_i})$ and predictions at intermediate scales $(\mathbf{p}_{f,s_i})$ while fused losses are computed from fused probability maps $(\mathbf{p}_{m,fused})$ and fused predictions $(\mathbf{p}_{f,fused})$. (Best viewed in color)

$$\mathbf{p}_b^f = BaseLearner(\mathbf{f}_m^q) \tag{13}$$

While Eq. 14, Eq. 15, and Eq. 16 are the same as in BAM [10], we employ separate ensemble models for each auxiliary predictions to render meta-model aware of non-novel regions at each scale, as inspired by BAM, as shown in Eq. 17, Eq. 18, and Eq. 19 as well as in Fig. 2 with pink rectangular boxes covered by dashed lines.

$$\mathbf{p}_{f,fused}^0 = Ens_\phi(\mathbf{p}_b^f, Ens_\psi(\mathbf{p}_{m,fused}^0, \psi)) \tag{14}$$

$$\mathbf{p}_{f,fused}^1 = Ens_\psi(\mathbf{p}_{m,fused}^1, \psi) \tag{15}$$

$$\mathbf{p}_{f,fused} = \mathbf{p}_{f,fused}^0 \oplus \mathbf{p}_{f,fused}^1 \tag{16}$$

$$\mathbf{p}_{f,s_i}^0 = Ens_{\phi,s_i}(\mathbf{p}_b^f, Ens_\psi(\mathbf{p}_{m,s_i}^0, \psi)) \tag{17}$$

$$\mathbf{p}_{f,s_i}^1 = Ens_\psi(\mathbf{p}_{m,s_i}^1, \psi) \tag{18}$$

$$\mathbf{p}_{f,s_i} = \mathbf{p}_{f,s_i}^0 \oplus \mathbf{p}_{f,s_i}^1 \tag{19}$$

$$\mathcal{L}_{meta}^{inner} = \sum_{i=1}^{N} CE(\mathbf{p}_{m,s_i}, \mathcal{M}_q) \tag{20}$$

$$\mathcal{L}_{meta}^{fused} = CE(\mathbf{p}_{m,fused}, \mathcal{M}_q) \tag{21}$$

$$\mathcal{L}_{final}^{inner} = \sum_{i=1}^{N} CE(\mathbf{p}_{f,s_i}, \mathcal{M}_q) \tag{22}$$

$$\mathcal{L}_{final}^{fused} = CE(\mathbf{p}_{f,fused}, \mathcal{M}_q) \tag{23}$$

Eq. 20 and Eq. 21 compute cross entropy losses for the auxiliary predictions and the fused prediction before ensemble respectively while Eq. 22 and Eq. 23 compute cross entropy losses for the auxiliary predictions and the fused prediction after ensemble accordingly.

$$\mathcal{L}_{final}^{total} = \mathcal{L}_{meta}^{inner} + \mathcal{L}_{meta}^{fused} + \mathcal{L}_{final}^{inner} + \mathcal{L}_{final}^{fused} \tag{24}$$

All individuals losses are accumulated to update the network eventually as in Eq. 24.

Backbone	Method	1-shot (%)					5-shot (%)				
		Fold-0	Fold-1	Fold-2	Fold-3	Average	Fold-0	Fold-1	Fold-2	Fold-3	Average
VGG-16	PFENet (TPAMI'20) [22]	56.90	68.20	54.40	52.40	58.00	59.00	69.10	54.80	52.90	59.00
	NTRENet (CVPR'22) [14]	57.70	67.60	57.10	53.70	59.00	60.30	68.00	55.20	57.10	60.20
	DPCN (CVPR'22) [13]	58.90	69.10	63.20	55.70	61.70	63.40	70.70	68.10	59.00	65.30
	BAM (CVPR'22) [10]	63.18	70.77	66.14	57.53	64.41	67.36	73.05	70.61	64.00	68.76
	BAM++ (ours)	**64.67**	**72.11**	**67.83**	**59.47**	**66.02**	**69.40**	**74.35**	**72.77**	**67.19**	**70.93**
ResNet-50	PGNet (ICCV'19) [28]	56.00	66.90	50.60	50.40	56.00	57.70	68.70	52.90	54.60	58.50
	PFENet (TPAMI'20) [22]	61.70	69.50	55.40	56.30	60.80	63.10	70.70	55.80	57.90	61.90
	NTRENet (CVPR'22) [14]	65.40	72.30	59.40	59.80	64.20	66.20	72.80	61.70	62.20	65.70
	ASNet (CVPR'22) [9]	68.90	71.70	61.10	**62.70**	66.10	**72.60**	74.30	65.30	67.10	70.80
	DPCN (CVPR'22) [13]	65.70	71.60	69.10	60.60	66.70	70.00	73.20	70.90	65.50	69.90
	BAM (CVPR'22) [10]	68.97	73.59	67.55	61.13	67.81	70.59	75.05	70.79	67.20	70.91
	BAM++ (ours)	**69.46**	**74.16**	**69.20**	61.54	**68.59**	70.81	**75.34**	**73.04**	**68.99**	**72.05**

Table 1. 1-shot and 5-shot class mIoU results on PASCAL-5^i dataset for VGG-16 and ResNet-50 as backbone, provided for 4 folds and the average. The best results are given in **boldface**. The underlined results show the best performance excluding our method.

Backbone	Method	1-shot (%)					5-shot (%)				
		Fold-0	Fold-1	Fold-2	Fold-3	Average	Fold-0	Fold-1	Fold-2	Fold-3	Average
ResNet-50	NTRENet (CVPR'22) [14]	36.80	42.60	39.90	37.90	39.30	38.20	44.10	40.40	38.40	40.30
	ASNet (CVPR'22) [9]	-	-	-	-	42.20	-	-	-	-	68.80
	DPCN (CVPR'22) [13]	42.00	47.00	43.20	39.70	43.00	46.00	54.90	50.80	47.40	49.80
	BAM (CVPR'22) [10]	43.41	50.59	**47.49**	43.42	46.23	49.26	54.20	**51.63**	**49.55**	51.16
	BAM++ (ours)	**44.43**	**51.98**	47.01	**45.22**	**47.16**	**52.53**	**57.02**	50.97	49.49	**52.50**

Table 2. 1-shot and 5-shot class mIoU results on COCO-20^i dataset for ResNet-50 as backbone, provided for 4 folds and the average. The best results are given in **boldface**. The underlined results show the best performance excluding our method.

5. Experiments

5.1. Details

Datasets. The model is evaluated on two datasets which are commonly used in few-shot segmentation tasks. PASCAL-5^i [20] is the first dataset, containing 20 classes and it is a combination of PASCAL VOC 2012 [4] and the extended annotations obtained from [6]. The second dataset is COCO-20^i [18], which is generated from MSCOCO [12]. COCO-20^i is more challenging when compared to PASCAL-5^i as it consists of images belonging to 80 classes. The datasets are split into 4 folds containing equal number of classes in order to perform cross-validation while 1000 support and query pairs are randomly sampled for each fold. One of the folds is selected for evaluating the performance of the model on unseen classes while the rest of them are used as base classes for training the model. This procedure is repeated for all folds.

Evaluation metrics. In order to compare with previous studies on few-shot segmentation [28, 22, 14, 9, 13, 10], class mean intersection-over-union is used as the evaluation metric, which is calculated as in 25, where C is the number of classes in each fold.

$$mIoU = \frac{1}{C}\sum_{i=1}^{C} IoU_i \tag{25}$$

The foreground-background IoU (FB-IoU) is also calculated as an additional metric.

Implementation details. All experiments are conducted on PyTorch framework with NVIDIA RTX 2080Ti GPUs. As suggested in BAM [10], there are two training stages, namely pre-training and meta-training. Pre-training stage is utilized for learning the base classes while ResNet-50 [7] and VGG-16 [21] are used as backbone for PASCAL-5^i and only ResNet-50 [7] is used as backbone for COCO-20^i. For PASCAL-5^i, PSPNet [30] is trained for 100 epochs as base learner with an initial learning rate of 2.5e-3. For the base learner on COCO-20^i, the model shared by the authors of [10] is used. In meta-training stage, PASCAL-5^i and COCO-20^i are trained for 200 and 50 epochs respectively while the learning rate is set to 5e-2. For both stages, SGD is utilized as optimizer. Random scaling, rotation, horizontal flip, cropping and Gaussian Blur is applied to images. The sizes of the enriched query features at the output of the improved decoder are set to 60, 30, 15, and 8, which makes N = 4 as suggested by [22].

Generalized few-shot segmentation setting. Our method is also evaluated in generalized few-shot segmentation setting, which is defined by [10], where both pixels belonging to novel and base classes are detected. For this setting, novel pixels are predicted as *novel* if their final foreground probabilities exceed a predefined threshold while the pixels predicted as *base* should be assigned to one of the base classes. By this way, the pixels belonging to different base classes are distinguished while the rest of the pixels are classified as *novel* or *background*. This setting requires the calculation of mIoU on base and novel classes and also the combination

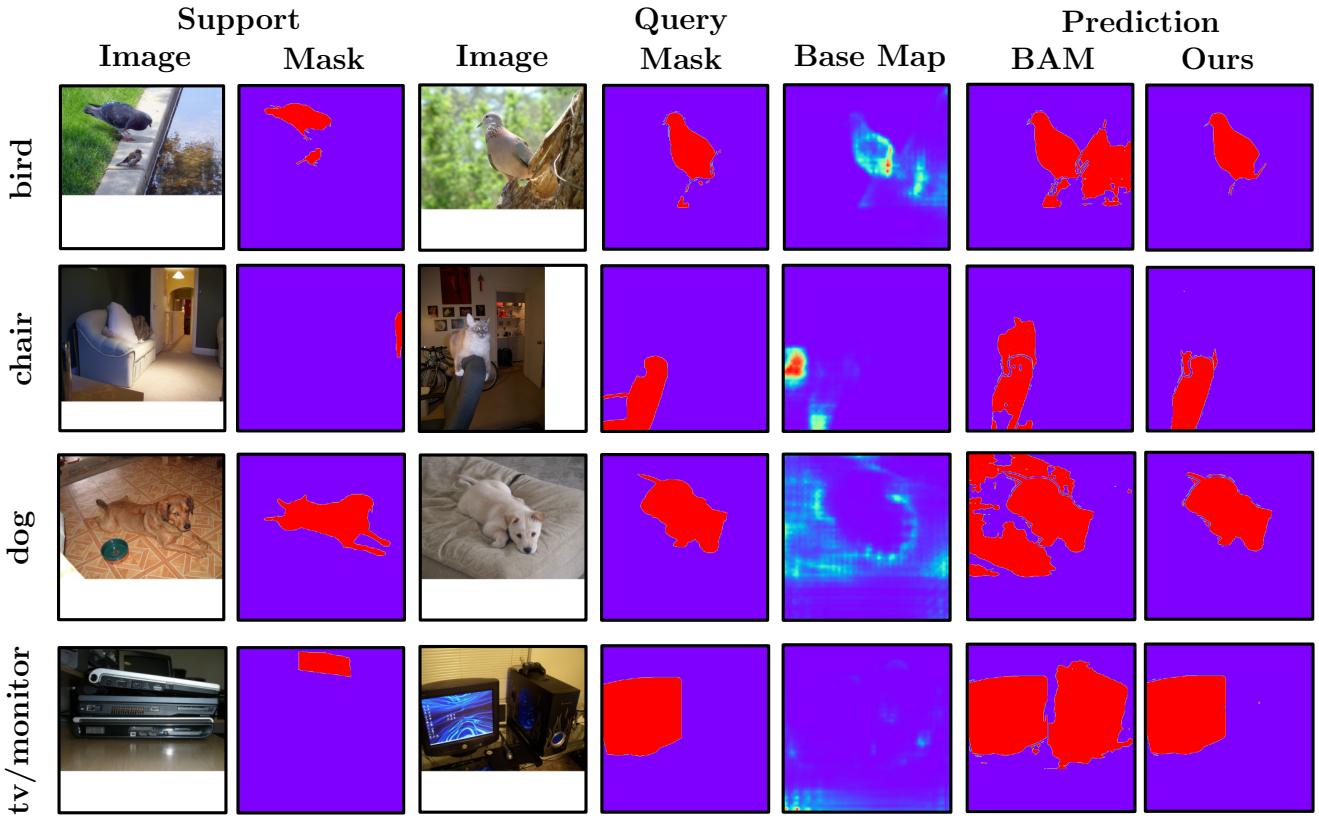

Figure 3. Qualitative 1-shot results on PASCAL-5^i dataset for ResNet-50 backbone. Results for one novel class from each fold are provided in rows. First two columns contain image and mask for support while the following two columns contain image and ground truth for query. Fifth column shows the probability map for query obtained from base learner. Predictions are provided for BAM [10] and our method for comparison in the last two columns. (Best viewed in color)

of them, which are denoted by mIoU_n, mIoU_b and mIoU_a respectively.

5.2. Results

5.2.1 Quantitative results

Table 1 shows the performance comparison between BAM++ and other methods proposed for few-shot segmentation task using ResNet-50 and VGG-16. The mIoU results include 1-shot and 5-shot cases for PASCAL-5^i dataset. BAM++ outperforms the existing methods for both settings. When VGG-16 is utilized as backbone, our method surpasses the state-of-the-art results by 1.61% and 2.17% for 1-shot and 5-shot settings, respectively. When it comes to the model with ResNet-50 as backbone, 0.78% and 1.14% performance gains are achieved for 1-shot and 5-shot settings. The results on COCO-5^i dataset are provided in Table 2 for ResNet-50 as backbone only. BAM++ outperforms the best results by 0.93% and 1.34% under 1-shot and 5-shot settings, respectively. Comparison with state-of-the-art models regarding the FB-IoU scores is provided in Table 3 for both backbones on PASCAL-5^i dataset. The results show that our method performs well in 1-shot setting while exceeds the best result by 0.66% in 5-shot setting for ResNet-50. On the other hand, model with VGG-16 outperforms the previous state-of-the-art by 1.43% and 1.42% for 1-shot and 5-shot settings respectively.

5.2.2 Qualitative results

Qualitative results for PASCAL-5^i dataset under 1-shot setting with ResNet-50 backbone are provided in Fig. 3. The differences between our proposed architecture and BAM can be seen when the predicted masks are analyzed. The main advantage of our model is revealed in cases where there is another object adjacent to the novel target object. In such cases, models generally tend to entangle the objects. In Fig. 3, it is seen that BAM predicts both the monitor and the computer as novel objects although there is only monitor in the support image. Since our model analyzes the features at different scales, it distinguishes the neighboring objects from each other well. Moreover, another faulty case is given in the third row, which is consistent with our hypothesis. Even though base learner discourages meta learner from non-novel regions, i.e. sofa, meta learner of BAM predicts these regions as novel. When ensembling the query predic-

Backbone	Method	1-shot (%)	5-shot (%)
VGG-16	PFENet (TPAMI'20) [22]	72.00	72.30
	NTRENet (CVPR'22) [14]	73.10	74.20
	DPCN (CVPR'22) [13]	73.70	77.20
	BAM (CVPR'22) [10]	77.26	81.10
	BAM++ (ours)	**78.69**	**82.52**
ResNet-50	PFENet (TPAMI'20) [22]	73.30	73.90
	NTRENet (CVPR'22) [14]	77.00	78.40
	ASNet (CVPR'22) [9]	77.70	80.40
	DPCN (CVPR'22) [13]	78.00	80.70
	BAM (CVPR'22) [10]	**79.71**	82.18
	BAM++ (ours)	79.65	**82.84**

Table 3. 1-shot and 5-shot FB-IoU results on PASCAL-5^i dataset for VGG-16 and ResNet-50 as backbone, provided as the average. The best results are given in **boldface**. The underlined results show the best performance excluding our method.

tions at different scales is introduced, such incorrect predictions are eliminated. As it can be seen in the predicted map of our method, only the regions belonging to the dog are considered as foreground. We deduce that ensembling at multi-scale ensures the model to focus on non-novel regions rather than the areas belonging to base classes.

5.2.3 Generalized few-shot segmentation results

Our method surpasses BAM [10] in generalized few-shot segmentation setting for both backbones on PASCAL-5^i dataset as shown in Table 4. The mIoU results validate the superiority of ensembling at multi-scale for both novel and base predictions.

5.3. Ablation study

Ablation study regarding the decision on how to include the inner losses for the multi-scale predictions are performed by considering the following cases: calculation of inner losses before and after the ensembling, without the ensembling, and after the ensembling only. The contributions of $\mathcal{L}_{meta}^{inner}$ in Eq. 20 and $\mathcal{L}_{final}^{inner}$ in Eq. 22 on the final mIoU performance are investigated. Thus, we experimented with the cases where either $\mathcal{L}_{meta}^{inner}$ is inactive, $\mathcal{L}_{final}^{inner}$ is inactive, or both $\mathcal{L}_{meta}^{inner}$ and $\mathcal{L}_{final}^{inner}$ are active for the $\mathcal{L}_{final}^{total}$ calculation in Eq. 24. The results are obtained for PASCAL-5^i dataset under 1-shot setting and provided in Table 5. Activating only $\mathcal{L}_{meta}^{inner}$ reaches an mIoU performance of 68.37% while including $\mathcal{L}_{final}^{inner}$ alone obtains the performance of 68.45%. The last row in Table 5 indicates that when both $\mathcal{L}_{meta}^{inner}$ and $\mathcal{L}_{final}^{inner}$ are used, the highest performance is achieved, which is 68.59%. As consequence, this ablation experiment validates our hypothesis, which emphasizes the weakness of the model implementing ensembling at single scale and the merits of the co-existence of $\mathcal{L}_{meta}^{inner}$ and $\mathcal{L}_{final}^{inner}$.

Backbone	Method	1-shot (%)			5-shot (%)		
		mIoU_n	mIoU_b	mIoU_a	mIoU_n	mIoU_b	mIoU_a
VGG-16	BAM [10]	43.19	67.03	61.07	46.15	67.02	61.80
	BAM++	**43.94**	**67.80**	**61.83**	**47.20**	**67.80**	**62.64**
ResNet-50	BAM [10]	47.93	72.72	66.52	49.17	72.72	66.83
	BAM++	**49.98**	**72.87**	**67.15**	**52.41**	**72.87**	**67.76**

Table 4. Generalized few-shot segmentation results on PASCAL-5^i dataset for VGG-16 and ResNet-50 as backbone. The best results are given in **boldface**.

Method	$\mathcal{L}_{inner}^{meta}$	$\mathcal{L}_{inner}^{final}$	mIoU (%)
BAM++	✓	-	68.37
BAM++	-	✓	68.45
BAM++	✓	✓	68.59

Table 5. Ablation studies on inner losses for the multi-scale predictions regarding the ensembling with the base map under 1-shot setting for PASCAL-5^i. Results show the averaged mIoU over 4 folds.

6. Conclusion

We observed that although ensembling meta prediction with base prediction guides the model by making the meta learner cautious in the regions where objects from base classes exist, meta learner misclassifies non-novel regions by neglecting base learner. This situation arises as a consequence of ensembling the predictions at single-scale. Therefore, we proposed to perform ensembling for predictions at multi-scale as well as the final prediction. By this way, bias existing at non-novel regions is diminished. The experiments on PASCAL-5^i and COCO-20^i verifies our hypothesis and our model achieves new state-of-the-art on few-shot segmentation benchmark.

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? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13979–13988, 2021.

[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, 2017.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[5] Jiaqi Fan, Fei Wang, Hongqing Chu, Xiao Hu, Yifan Cheng, and Bingzhao Gao. Mlfnet: Multi-level fusion network

for real-time semantic segmentation of autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2022.

[6] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] Sunghwan Hong, Seokju Cho, Jisu Nam, and Seungryong Kim. Cost aggregation is all you need for few-shot segmentation. *arXiv preprint arXiv:2112.11685*, 2021.

[9] Dahyun Kang and Minsu Cho. Integrative few-shot learning for classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9979–9990, 2022.

[10] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8057–8067, 2022.

[11] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8334–8343, 2021.

[12] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[13] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2022.

[14] Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11573–11582, 2022.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[16] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8741–8750, 2021.

[17] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6941–6952, 2021.

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

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

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

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[23] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[24] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019.

[25] 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*, pages 5475–5484, 2021.

[26] 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*, pages 7293–7302, 2021.

[27] Xianghui Yang, Bairun Wang, Kaige Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou. Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation. *arXiv preprint arXiv:2008.06226*, 2020.

[28] 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*, pages 9587–9595, 2019.

[29] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019.

[30] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.