# CAFENet: Class-Agnostic Few-Shot Edge Detection Network

Younghyun Park
dnffkf369@kaist.ac.kr
Jun Seo
tjwns0630@kaist.ac.kr
Jaekyun Moon
jmoon@kaist.edu

Korea Advanced Institute
of Science and Technology,
Daejeon, Korea

**Abstract**

We tackle a novel few-shot learning challenge, few-shot semantic edge detection, aiming to localize boundaries of novel categories using only a few labeled samples. Reliable boundary information has been shown to boost the performance of semantic segmentation and localization, while also playing a key role in its own right in object reconstruction, image generation and medical imaging. However, existing semantic edge detection techniques require a large amount of labeled data to train a model. To overcome this limitation, we present Class-Agnostic Few-shot Edge detection Network (CAFENet) based on a meta-learning strategy. CAFENet employs a semantic segmentation module in small-scale to compensate for the lack of semantic information in edge labels. To effectively fuse the semantic information and low-level cues, CAFENet also utilizes an attention module which dynamically generates multi-scale attention map, as well as a novel regularization method that splits high-dimensional features into several low-dimensional features and conducts multiple metric learning. Since there are no existing datasets for few-shot semantic edge detection, we construct two new datasets, FSE-1000 and SBD-$5^i$, and evaluate the performance of the proposed CAFENet on them. Extensive simulation results confirm that CAFENet achieves better performance compared to the baseline methods using fine-tuning or few-shot segmentation.

# 1 Introduction

Semantic edge detection aims to identify pixels that belong to boundaries of predefined categories. Boundary information has been shown to be effective for boosting the performance of semantic segmentation [3, 5] and localization [31, 38]. It also plays a key role in applications such as object reconstruction [11, 43], image generation [18, 32] and medical imaging [1, 22]. Early edge detection algorithms interpret the problem as a low-level grouping problem exploiting hand-crafted features and local information [4, 28]. Recently, there have been significant improvements in edge detection thanks to advances in deep learning. Moreover, beyond previous boundary detection, category-aware semantic edge detection became possible [2, 17, 40]. However, it is still not feasible to train deep neural networks without massive amounts of annotated data.

To overcome the data scarcity issue in image classification, few-shot learning has been actively discussed in recent years [12, 20]. Few-shot learning algorithms train machines to

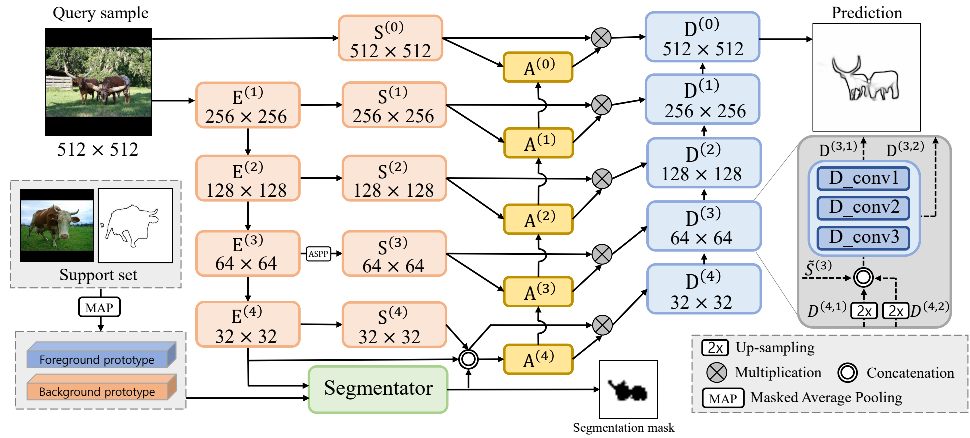


Figure 1: Network architecture overview of proposed CAFENet. ResNet-34 encoders $E^{(1)} \sim E^{(4)}$ extract multi-level semantic features. The segmentator module generates a segmentation prediction using query feature from $E^{(4)}$ and prototypes $P_{FG}, P_{BG}$ from support set features. Small bottleneck blocks $S^{(0)} \sim S^{(4)}$ transform the original image and multi-scale features from encoder blocks to be more suitable for edge detection. Dynamic Attention Modules $A^{(0)} \sim A^{(4)}$ dynamically generate attention map in every position and every resolution. Decoder $D^{(0)} \sim D^{(4)}$ take attentive multi-scale features to give edge prediction.

learn previously unseen classification tasks using only a few labeled examples. More recently, the idea of few-shot learning is applied to computer vision tasks requiring highly laborious and expensive data labeling such as semantic segmentation [9, 30] and object detection [13, 19]. In this paper, we consider a novel few-shot learning challenge, few-shot semantic edge detection, to detect the semantic boundaries using only a few labeled samples. Through experiments, we show that few-shot semantic edge detection can not be simply solved by fine-tuning a pretrained semantic edge detector or utilizing a nonparametric edge detector in a few-shot segmentation setting. To tackle this elusive challenge, we propose a class-agnostic few-shot edge detector (CAFENet) and present new datasets for evaluating few-shot semantic edge detection.

Fig. 1 shows the architecture of the proposed CAFENet. Since the edge labels do not contain enough semantic information due to the sparsity of labels, the performance of the edge detector severely degrades when the training dataset is very small. To overcome this, we jointly train the segmentation module with segmentation labels generated from given boundary labels. Although the previous works of [15, 42] show that joint multi-task learning with segmentation and edge detection can improve the performance, ours is the first attempt to use the segmentator in low-resolution to supplement semantic information for edge detection.

The main contributions of this paper are as follows. 1) We formulate a novel problem: few-shot semantic edge detection that aims to perform semantic edge detection on previously unseen objects using a few training examples. 2) We devise a few-shot semantic edge detector, CAFENet, which jointly trains a metric-based segmentator with an edge detector to effectively exploit the few labeled samples. 3) We propose multi-split matching regularization (MSMR) to regularize the embedding space and the metric-based segmentator. 4) We build a dynamic attention module (DAM) that dynamically generates multi-scale attention maps to effectively fuse the semantic information and local cues to make accurate but category-aware edge prediction. 5) We introduce two new datasets of SBD-$5^i$ and FSE-1000 for few-shot edge detection and show that CAFENet outperforms baselines by large margins.

# 2 Related Work

## 2.1 Few-shot Learning

To tackle the few-shot learning challenge, many methods have been proposed based on meta-learning. Optimization-based methods [12, 25] train the meta-learner which updates the parameters of the actual learner so that the learner can easily adapt to a new task within a few labeled samples. Metric-based methods [27, 29, 36] train the feature extractor to assemble features from the same class together on the embedding space while keeping features from different classes far apart.

## 2.2 Few-shot Semantic Segmentation

Few-shot segmentation aims to perform semantic segmentation using a few labeled samples. OSLSM of [26] adopts a two-branch structure: conditioning branch generating element-wise scale and shift factors and segmentation branch performing segmentation with task-conditioned features. Co-FCN [24] also utilizes a two-branch structure. The globally pooled prediction is generated in the conditioning branch and fused with query features to predict the mask in the segmentation branch. CANet of [41] adopts masked average pooling to generate the global feature vector, and concatenates it with every location of the query feature for dense comparison. PANet of [30] introduces prototype alignment, predicting the segmentation mask of support samples using query prediction results as labels of query samples, for regularization. PMM of [35] utilizes multiple prototypes with Expectation-Maximization (EM) process to effectively leverage the semantic information from the few labeled samples.

## 2.3 Semantic Edge Detection

Semantic edge detection aims to find the boundaries of objects from an image and classify the objects at the same time. The history of semantic edge detection [1, 17] dates back to the work of [23] which adopts the support vector machine as a semantic classifier on top of the traditional canny edge detector. Recently, many semantic edge detection algorithms rely on deep neural networks. CASENET of [39] addresses the semantic edge detection as a multi-label problem where each boundary pixel is labeled into categories of adjacent objects. DFF of [17] proposes a novel way to leverage multi-scale features. The multi-scale features are fused by weighted summation with fusion weights generated dynamically for each image and each pixel. RPCNet of [42] and PGN of [15] propose to jointly train segmentation module with edge detector in original resolution to improve the performance of edge detection. AG-CRFs of [34] considers attention-gated CRF to fuse multi-scale features. BAN of [14] employs a channel-wise attention mechanism to preserve informative features. Our method also utilizes a segmentation module and attention mechanism. However, CAFENet utilizes a metric-based few-shot segmentator in low-resolution and relies on segmentation prediction to supplement semantic information to the edge detector. For the attention mechanism, existing attention methods are only applicable to non-semantic edge detection problems. To solve the arduous semantic edge detection problem, we model a novel attention module that generates a multi-scale spatial-wise attention map to highlight semantically meaningful regions.

# 3 Problem Setup

For few-shot semantic edge detection, we use train set $D_{train}$ and test set $D_{test}$ consisting of non-overlapping categories $C_{train}$ and $C_{test}$. The model is trained only using $C_{train}$, and the

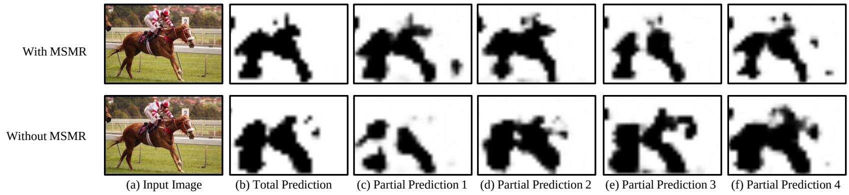


Figure 2: Example results on SBD-$5^i$. Columns from left to right: Input image, total prediction, and partial predictions from each feature split. The total prediction is obtained by matching the high-dimensional prototypes with the high-dimensional feature vectors. To generate partial predictions, we equally split feature vectors into 4 low-dimensional sub-vectors as done in MSMR and match the low-dimensional feature sub-vectors with corresponding prototype sub-vectors.

test categories $C_{test}$ are never seen during the training phase. For meta-training of the model, we adopt episodic training. Each episode is composed of a support set with a few labeled samples and a query set. When an episode is given, the model adapts to the given episode using the support set and detect semantic boundaries of the query set. In this work, we address $N_c$-way $N_s$-shot semantic edge detection. In this setting, each training episode is constructed by $N_c$ classes sampled from $C_{train}$. When $N_c$ categories are given, $N_s$ support samples and $N_q$ query samples are randomly chosen from $D_{train}$ for each class. In evaluation, the performance of the model is measured using test episodes. The test episodes are constructed in the same way as the training episodes, except that $N_c$ classes and corresponding support and query samples are sampled from unencountered $C_{test}$ and $D_{test}$.

# 4 Method

We propose a novel algorithm for few-shot semantic edge detection. Fig. 1 illustrates the network architecture. The proposed CAFENet adopts the semantic segmentation module to compensate for the lack of semantic information in edge labels. The predictive segmentation mask is used to generate attention maps which are applied to multi-scale skip connection features. The final edge prediction is generated using attentive multi-scale features.

## 4.1 Semantic Segmentator

Most previous works on semantic edge detection directly predict edges from the given input image. However, direct edge prediction is challenging when only a few labeled samples are given. To overcome this difficulty, we combine a semantic segmentation module with an edge detector. With the assistance of the segmentation module, CAFENet can effectively localize the target object. For few-shot segmentation, we employ the metric-learning which utilizes prototypes for foreground and background as done in [9, 30]. Given the support set $S = \{x_i^s, y_i^s\}_{i=1}^{N_s}$, the encoder $E$ extracts features $\{E(x_i^s)\}_{i=1}^{N_s}$ from $S$. Also, given support edge labels $\{y_i^s\}_{i=1}^{N_s}$, we generate the dense segmentation mask $\{M_i^s\}_{i=1}^{N_s}$ using a rule-based preprocessor; pixels inside the boundary are considered as foreground pixels. Using down-sampled segmentation labels $\{m_i^s\}_{i=1}^{N_s}$, the prototype for foreground pixels $P_{FG}$ is computed as $P_{FG} = \frac{1}{N_s}\frac{1}{H\times W}\sum_i\sum_j E_j(x_i^s)m_{i,j}^s$ where $j$ indexes the pixel location and $H, W$ denote height and width. Likewise, the background prototype $P_{BG}$ is computed using negative mask $1 - m_{i,j}^s$. The probability that pixel $j$ belongs to foreground for the query sample $x_i^q$ is

$$p(y_{i,j}^q = FG|x_i^q;E) = \frac{exp(-\tau d(E_j(x_i^q), P_{FG}))}{exp(-\tau d(E_j(x_i^q), P_{FG})) + exp(-\tau d(E_j(x_i^q), P_{BG}))} \quad (1)$$

where $d(\cdot,\cdot)$ is squared Euclidean distance between two vectors and $\tau$ is a learnable temperature parameter. With query samples $\{x_i^q\}_{i=1}^{N_q}$ and the down-sampled segmentation labels $\{m_i^q\}_{i=1}^{N_q}$, the segmentation loss $L_{Seg}$ is calculated as the mean-squared error (MSE) loss between predicted probabilities and the down-sized segmentation mask.

$$L_{Seg} = \frac{1}{N_q}\frac{1}{H\times W}\sum_{i=1}^{N_q}\sum_{j=1}^{H\times W}\{(p(y_{i,j}^q = FG|x_i^q;E) - m_{i,j}^q)^2\}. \tag{2}$$

Note that the segmentation mask is generated in a down-sized scale so that any pixel near the boundaries can be classified into the foreground to some extent, as well as the background. Therefore, we regard the problem as regression using the MSE loss.

## 4.2 Multi-Split Matching Regularization

The metric-based few-shot segmentation method utilizes distance metrics between the high-dimensional feature vectors and prototypes. However, this approach is prone to overfitting due to the massive number of parameters in feature vectors. To get around this issue, we propose a novel regularization method: multi-split matching regularization (MSMR). In MSMR, high-dimensional feature vectors are randomly split into several low-dimensional feature vectors, and the metric learning is conducted on each vector split. With the query feature $E(x_i^q) \in \mathbb{R}^{C\times W\times H}$ where $C$ is channel dimension, $E(x_i^q)$ is randomly divided into $K$ sub-vectors $\{E^k(x_i^q)\}_{k=1}^K$ along channel dimension. Each sub-vector $E^k(x_i^q)$ is in $\mathbb{R}^{\frac{C}{K}\times W\times H}$. Likewise, the prototypes are also disassembled into $K$ corresponding sub-vectors $\{P_{FG}^k\}_{k=1}^K$ and $\{P_{BG}^k\}_{k=1}^K$. For the $k^{th}$ sub-vector of query feature $E^k(x_i^q)$, the probability that the $j^{th}$ pixel belongs to the foreground class is computed as follows:

$$p^k(y_{i,j}^q = FG|x_i^q;E) = \frac{exp(-\tau d(E_j^k(x_i^q), P_{FG}^k))}{exp(-\tau d(E_j^k(x_i^q), P_{FG}^k)) + exp(-\tau d(E_j^k(x_i^q), P_{BG}^k))}. \tag{3}$$

The prediction result of $K$ sub-problems are reflected on learning by combining the split-wise losses to original loss in Eq. 2. The final segmentation loss is calculated as

$$L_{Seg} = \frac{1}{N_q\times H\times W}\sum_{i=1}^{N_q}\sum_{j=1}^{H\times W}\{(p_{i,j} - m_{i,j}^q)^2 + \sum_{i=1}^{K}(p_{i,j}^k - m_{i,j}^q)^2\}. \tag{4}$$

where $p_{i,j} = p(y_{i,j}^q = FG|x_i^q;E)$, $p_{i,j}^k = p^k(y_{i,j}^q = FG|x_i^q;E)$.

In Fig. 2, we evaluate the quality of partial predictions generated from 4 low-dimensional sub-vectors to figure out the effect of MSMR. While the model trained with previous metric learning shows inconsistent partial predictions, the model trained with MSMR shows consistent partial predictions and generates a better total prediction as well.

## 4.3 Dynamic Attention Module

In few-shot edge detection, it is important to appropriately utilize semantic information and low-level details together. As shown in Fig. 1, we adopt the nested encoder structure to exploit rich hierarchical features. The multi-scale side outputs from encoder $E^{(1)} \sim E^{(4)}$ are post-processed through bottleneck blocks $S^{(1)} \sim S^{(4)}$. We employ the Atrous Spatial Pyramid Pooling (ASPP) block of [6] in front of $S^{(3)}$. We have empirically found that locating ASPP there shows better performance.

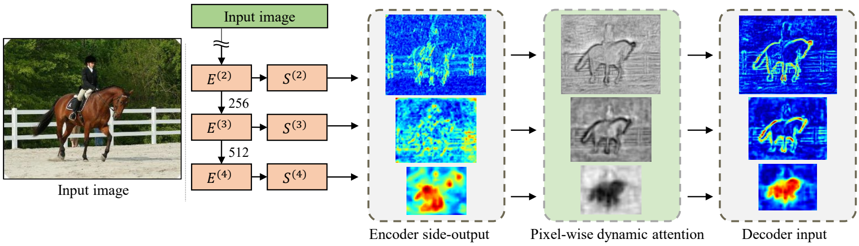


Figure 3: An example of activation map of [37] before and after pixel-wise semantic attention (warmer color has higher value). As seen, the attention mechanism makes encoder side-outputs attend to the regions of the target object (*horse* in the figure).

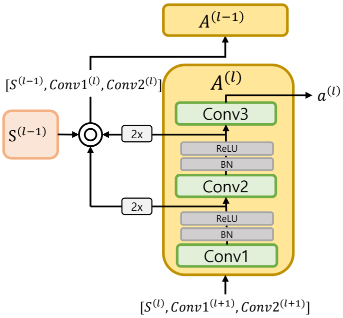


Figure 4: Architecture of DAM

Even though the multi-scale features provide local details, it is difficult to localize the target object from the multi-scale features as shown in Fig. 3. To appropriately fuse semantic information and low-level cues, we propose to build Dynamic Attention Module(DAM) which generates attention values for every pixel location, in every resolution. Using the attention maps generated by DAM, the multi-scale features are refined to support semantic edge detection. As shown in Fig. 4, Our DAM consists of 5 convolutional blocks, and each block $A^{(l)}$ is composed of 3 successive convolutional layers. $A^{(l)}$ takes the feature from $S^{(l)}$ and $A^{(l+1)}$ as inputs, and outputs the attention map $a^{(l)}$ and the feature. Especially, the block $A^{(4)}$ in the lowest scale takes a concatenated feature vector $[E^4(x_i^q), S^4(x_i^q), p(x_i^q)]$ as the input where $p(x_i^q)$ contains predictive distribution for segmentation. To the best of our knowledge, this is the first attempt to adopt a bottom-top attention module to generate multi-scale spatial-wise attention map. For applying the attention map $a^{(l)}$, multi-level side output from $S^{(l)}$ is pixel-wisely weighted by $1+a^{(l)}$ to obtain $\tilde{S}^{(l)}$, selectively highlighting the semantically important region. We visualize the effect of semantic attention of DAM in Fig. 3. Interestingly, the attention map $a^{(l)}$ in low resolution highlights the semantically related region, while the counterpart in high resolution focuses on local details like sudden changes in pixel values.

## 4.4 Semantic Edge Detector

As shown in Fig. 1, the decoder network is composed of five consecutive convolutional blocks. The outputs of decoder blocks $D^{(1)} \sim D^{(4)}$ are bilinearly upsampled by two and passed to the next block. Similar to [10], the up-sampled decoder outputs are then concatenated to the skip connection features from the previous decoder block and the attended multi-scale features $\tilde{S}^{(0)} \sim \tilde{S}^{(4)}$. The hierarchical decoder network in turn refines the outputs of the previous decoder blocks and finally produces the edge prediction $\hat{y}_i^q$ of query samples $x_i^q$. Given a query set $Q = \{x_i^q, y_i^q\}_{i=1}^{N_q}$, the cross-entropy loss is computed as

$$L_{CE} = -\sum_{i=1}^{N_q} \{ \sum_{j \in Y_+} log(\hat{y}_i^q) + \sum_{j \in Y_-} log(1-\hat{y}_i^q) \} \tag{5}$$

where $Y_+$ and $Y_-$ denote the sets of foreground and background pixels. To produce crisp boundaries, cross-entropy loss is combined with Dice loss of [8]

$$L_{Dice} = \sum_{i=1}^{N_q} \{ \frac{\sum_j (\hat{y}_{i,j}^q)^2 + \sum_j (y_{i,j}^q)^2}{2\sum_j \hat{y}_{i,j}^q y_{i,j}^q} \} \tag{6}$$

| Metric | Method (5-shot) | SBD-$5^0$ | | SBD-$5^1$ | | SBD-$5^2$ | | SBD-$5^3$ | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MF (ODS) | DFF + Finetune | 8.87 | 9.25 | 5.54 | 8.63 | 4.91 | 8.08 | 2.95 | 7.83 | 5.57 | 8.45 |
| | PGN + Finetune | 14.83 | 19.11 | 16.89 | 19.95 | 16.13 | 19.24 | 13.94 | 16.81 | 15.45 | 18.78 |
| | PANet + Sobel | 18.13 | 19.47 | 23.17 | 23.33 | 21.04 | 21.04 | 17.75 | 17.78 | 20.02 | 20.41 |
| | PMM + Sobel | 31.18 | 31.73 | 29.23 | 29.99 | 29.38 | 29.91 | 25.65 | 26.03 | 28.86 | 29.42 |
| | CAFENet (Ours) | **34.92** | **39.02** | **40.83** | **42.52** | **34.75** | **38.41** | **32.16** | **35.54** | **35.67** | **38.87** |
| AP | DFF + Finetune | 7.91 | 9.17 | 3.71 | 6.77 | 3.31 | 7.04 | 1.55 | 6.14 | 4.12 | 7.28 |
| | PGN +Finetune | 10.81 | 12.96 | 11.49 | 13.67 | 10.73 | 12.46 | 8.43 | 10.18 | 10.37 | 12.32 |
| | PANet + Sobel | 11.56 | 11.52 | 14.78 | 14.10 | 12.40 | 11.84 | 9.46 | 9.29 | 12.05 | 11.69 |
| | PMM + Sobel | 22.77 | 23.26 | 20.21 | 20.76 | 19.85 | 20.38 | 17.56 | 17.94 | 20.10 | 20.59 |
| | CAFENet (Ours) | **31.29** | **35.41** | **35.95** | **38.92** | **29.32** | **33.41** | **25.89** | **29.73** | **30.61** | **34.37** |

Table 2: Comparison results on SBD-$5^i$. Both 5-shot (right) and 1-shot (left) performances are considered.

where j denotes the pixels of a label. The final loss for meta-training is given by $L_{final} = \lambda_1 L_{Seg} + \lambda_2 L_{CE} + \lambda_3 L_{Dice}$, where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters to balance various losses.

# 5 Experiments

## 5.1 Datasets

**SBD-$5^i$**: based on the SBD dataset of [16] for semantic edge detection, we propose a new SBD-$5^i$ dataset. With reference to the setting of Pascal-$5^i$, 20 classes of the SBD dataset are divided into 4 splits. In the experiment with split $i$, 5 classes in the $i^{th}$ split are used as $C_{test}$, while the remaining 15 classes are utilized as $C_{train}$. The training set $D_{train}$ is constructed with all image-annotation pairs whose annotation includes at least one pixel from the classes in $C_{train}$. For each class, the boundary pixels which do not belong to that class are considered as background. The test set $D_{test}$ is also constructed in the same way as $D_{train}$, using $C_{test}$ this time. We conduct 4 experiments with each split of $i = 0 \sim 3$, and report the performance of each split as well as the averaged performance.

| Metric | Method | 1-shot | 5-shot |
|---|---|---|---|
| MF (ODS) | PANet + Sobel | 38.68 | 39.83 |
| | PMM + Sobel | 32.39 | 36.53 |
| | CAFENet (Ours) | **57.88** | **59.52** |
| AP | PANet + Sobel | 28.37 | 29.28 |
| | PMM + Sobel | 27.82 | 33.45 |
| | CAFENet (Ours) | **58.78** | **60.93** |

Table 1: Comparison results on FSE-1000.

**FSE-1000**: the datasets used in previous semantic edge detection research such as SBD of [16] and Cityscapes of [7] are not suitable for few-shot learning as they have only 20 and 30 classes, respectively. We propose a new FSE-1000 dataset based on FSS-1000 of [33]. FSS-1000 is a dataset for few-shot segmentation and composed of 1000 classes and 10 images per class with foreground-background segmentation annotation. From the images and segmentation masks of FSS-1000, we build FSE-1000 by extracting boundary labels from segmentation masks. For dataset split, we split 1000 classes into 800 training classes and 200 test classes. Detailed class configuration can be found in the Supplementary Material.

## 5.2 Evaluation Settings

We use two evaluation metrics to measure the few-shot semantic edge detection performance of our approach: the Average Precision (AP) and the F-measure (MF) at optimal dataset scale (ODS). In evaluation, we compare the unthinned raw prediction results and the ground truths without Non-Maximum Suppression (NMS) following [2, 40]. For the evaluation of edge detection, we set a matching distance tolerance to be zero which is an error threshold between the prediction result and the ground truth. In addition, we evaluate not only the positive predictions from the area inside an object but also the zero-padded region as false positives, which is stricter than the evaluation protocol in prior works of [16, 39]. 1000 test episodes are randomly sampled for the evaluation. The scores are given in percentages. For more implementation details, see Supplementary Materials.

| Metric | Method | SBD-$5^0$ | | SBD-$5^1$ | | SBD-$5^2$ | | SBD-$5^3$ | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MF (ODS) | baseline | 28.89 | 30.02 | 30.80 | 31.27 | 28.29 | 28.23 | 25.23 | 26.10 | 28.30 | 28.91 |
| | Seg | 34.43 | 37.63 | 39.36 | 41.13 | 33.19 | 35.16 | 30.82 | 33.33 | 34.45 | 36.81 |
| | Seg + DAM | 34.29 | 38.67 | 39.67 | 41.75 | 33.31 | 35.82 | 31.22 | 33.91 | 34.62 | 37.54 |
| | Seg + MSMR | 33.66 | 38.11 | 40.09 | 41.65 | 34.19 | 37.62 | 30.45 | 34.28 | 34.60 | 37.92 |
| | Seg + DAM + MSMR | **34.92** | **39.02** | **40.83** | **42.52** | **34.75** | **38.41** | **32.16** | **35.54** | **35.67** | **38.87** |
| AP | baseline | 25.11 | 26.09 | 26.36 | 26.49 | 21.87 | 22.47 | 20.63 | 21.58 | 23.49 | 24.16 |
| | Seg | 29.98 | 33.04 | 34.56 | 36.69 | 26.52 | 29.66 | 25.02 | 26.64 | 29.02 | 31.51 |
| | Seg + DAM | 30.43 | 34.58 | 34.89 | 37.54 | 27.62 | 29.94 | 25.31 | 27.95 | 29.56 | 32.50 |
| | Seg + MSMR | 29.24 | 33.95 | 34.64 | 37.05 | 28.94 | 32.84 | 23.90 | 28.12 | 29.18 | 32.99 |
| | Seg + DAM + MSMR | **31.29** | **35.41** | **35.95** | **38.92** | **29.32** | **33.41** | **25.89** | **29.73** | **30.61** | **34.37** |

Table 4: Ablation studies on SBD-$5^i$. 5-shot (right) and 1-shot (left) performances are considered.

| Metric | Method | 1-shot | 5-shot |
|---|---|---|---|
| MF (ODS) | baseline | 52.71 | 53.52 |
| | Seg | 56.03 | 57.98 |
| | Seg + DAM | 56.41 | 56.91 |
| | Seg + MSMR | 57.50 | 59.38 |
| | Seg + DAM + MSMR | **57.88** | **59.52** |
| AP | baseline | 53.66 | 54.59 |
| | Seg | 56.82 | 58.21 |
| | Seg + DAM | 57.90 | 59.38 |
| | Seg + MSMR | 58.36 | 60.32 |
| | Seg + DAM + MSMR | **58.78** | **60.93** |

Table 3: Ablation studies on FSE-1000.

To verify the proposed method, we compare CAFENet with two baselines. The first baseline is a fine-tuned semantic edge detection model with only a few labeled samples. A meta-learning strategy is not used for the first baseline. We employ the DFF of [17] and PGN of [15] with the implementation offered by the authors. For each split of SBD-$5^i$, we pretrain a 15-way edge detector with training classes and fine-tune the pretrained edge detector with a few labeled samples for new classes in the test split. During pretraining, we follow the training strategies and hyperparameters of [17] and [15], respectively. In fine-tuning, we randomly initialize some sub-modules that are closely related to the final classification ("side5", "side5-w", and "ada-learner" for [17] and "edge branch", "segmentation branch" and "refinement branch" for [15]) and train them altogether using the support images. The second baseline is constructed by combining a rule-based edge detector with a few-shot segmentation algorithm. It is occasionally believed that semantic edge detection can be replaced by segmentation, but prior works of [2, 21] verify that the semantic edge detector outperforms the segmentator combined with the Sobel operator. In our experiments, we combine PANet [30] and PMM [35] with the Sobel operator based on the implementation provided by the authors. For each split of SBD-$5^i$, we meta-train the PANet and PMM on training classes. In evaluation, we obtain edge predictions by applying the Sobel operator on the segmentation predictions as done in [2]. For a fair comparison, we thoroughly find the best kernel size for the Sobel operator. See Supplementary Material for more details.

We utilize the ResNet-34 backbone for CAFENet and PANet. For PMM, the ResNet-50 backbone is used. We also employ higher shot training in 1-shot experiments for both baselines as done in CAFENet experiments. The results in Tables 2 and 1 show that the proposed CAFENet outperforms all baselines in both MF and AP scores by a significant margin; few-shot semantic edge detector can not be simply substituted by a few-shot segmentator or a fine-tuned semantic edge detector. This is an impressive result because CAFENet wields a ResNet-34 backbone which is smaller than the ResNet-50 backbone of PMM. For FSE-1000, we only experiment with the few-shot segmentation baseline since it is hard to train a semantic edge detector with a large number of training classes. We can see that the proposed CAFENet outperforms the baseline even when the dataset contains more diverse classes.

## 5.3 Ablation Studies on DAM and MSMR

In this section, we show the results of ablation experiments to examine the impact of the proposed DAM and MSMR. The results on SBD-$5^i$ and FSE-1000 are shown in Tables 4 and 3, respectively. The **baseline** method does not utilize a segmentation module. The prototypes for edge and non-edge classes are computed using down-sampled edge labels, and

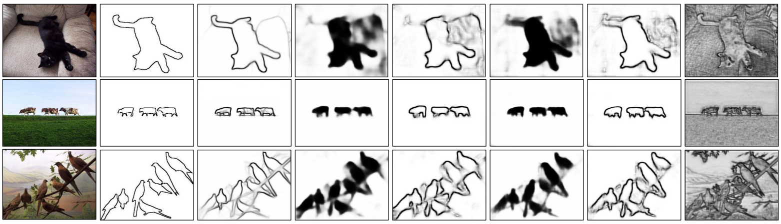


Figure 5: Qualitative Results on the SBD-$5^i$ Dataset.

the edge prediction is done using a metric-based method. The low-scale edge prediction is concatenated with the encoder feature and the skip connection feature, and then passed to the decoder to predict the edge in the original scale. The **Seg** method utilizes the segmentation module, and conducts semantic segmentation in low scale using the segmentation labels generated from the edge labels. As done in the baseline, the segmentation result is concatenated with the encoder feature and skip connection feature and directly passed to the decoder. **Seg + DAM** applies the dynamic attention module with the segmentation module. DAM applies multi-scale and pixel-wise attention to features in the skip architecture. **Seg + MSMR** applies MSMR on the top of the segmentation module, with the auxiliary regularization loss for training. The **Seg + DAM + MSMR** method utilizes both DAM and MSMR. For a fair comparison, all methods use the same network architecture and hyperparameter settings. Tables 4 and 3 demonstrate that segmentation process in **Seg** gives significant performance advantages over baseline for both SBD-$5^i$ and FSE-1000. It is also seen that **Seg + DAM** benefits from the additional usage of DAM and exhibits the better performance than **Seg**. MSMR regularization also gives a performance gain and **Seg + MSMR** outperforms **Seg**. Finally, applying DAM and MSMR together provides extra gains, as seen by the scores associated with **Seg + DAM + MSMR**. Clearly, when compared to **baseline**, our overall approach **Seg + DAM + MSMR** provides large gains.

### 5.4 Qualitative Results

In Fig. 5 and Fig. 6, we illustrate qualitative results of our method as well as the baseline methods for SBD-$5^i$ and FSE-1000, respectively. From the result, we can see that the DFF method succeeds in finding the edges of the objects, but it fails to distinguish the boundary of target object from the boundary of the other objects. On the other hand, PANet + Sobel and PMM + Sobel methods successfully localize the target object, but they fail to refine the correct boundary. In contrast, the proposed CAFENet is capable of localizing the objects from target class and detecting the correct boundary at the same time.

## 6 Conclusion

In this paper, we establish the few-shot semantic edge detection problem. We proposed the Class-Agnostic Few-shot Edge detector (CAFENet) based on a skip architecture utilizing multi-scale features. To compensate for the shortage of semantic information in edge labels, the segmentation module is employed in low resolution. A dynamic attention module generates attention maps from segmentation masks effectively combining the semantic information and local details. The attention maps are applied to multi-scale skip connections to localize the semantically related region. We also present the MSMR regularization method splitting the feature vectors and prototypes into several low-dimension sub-vectors and solving multiple metric-learning sub-problems with the sub-vectors. We built two novel

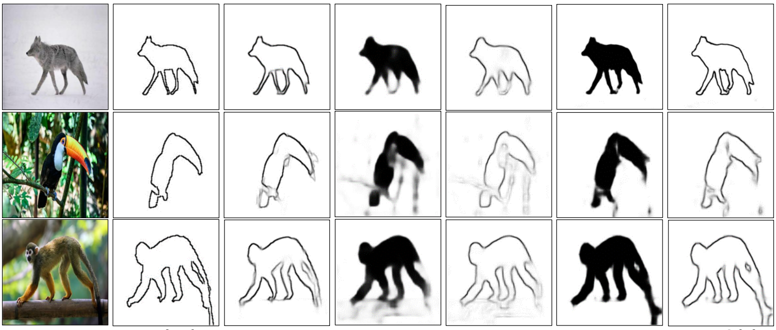


Figure 6: Qualitative Results on the FSE-1000 Dataset.

datasets of FSE-1000 and SBD-$5^i$ well-suited to few-shot semantic edge detection. Experimental results demonstrate that the proposed method significantly outperforms the baseline approaches relying on fine-tuning or few-shot semantic segmentation.

# Acknowledgement

This paper was result of the research project supported by SK hynix Inc.

# References

[1] Husein Hadi Abbass and Zainab Radhi Mousa. Edge detection of medical images using markov basis. *Applied Mathematical Sciences*, 11(37):1825–1833, 2017.

[2] David Acuna, Amlan Kar, and Sanja Fidler. Devil is in the edges: Learning semantic boundaries from noisy annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11075–11083, 2019.

[3] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3602–3610, 2016.

[4] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[5] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4545–4554, 2016.

[6] 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.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] Ruoxi Deng, Chunhua Shen, Shengjun Liu, Huibing Wang, and Xinru Liu. Learning to predict crisp boundaries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 562–578, 2018.

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

[10] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1623–1632, 2019.

[11] Vittorio Ferrari, Loic Fevrier, Frederic Jurie, and Cordelia Schmid. Groups of adjacent contour segments for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 30(1):36–51, 2007.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[13] Kun Fu, Tengfei Zhang, Yue Zhang, Menglong Yan, Zhonghan Chang, Zhengyuan Zhang, and Xian Sun. Meta-ssd: Towards fast adaptation for few-shot object detection with meta-learning. *IEEE Access*, 7:77597–77606, 2019.

[14] Lianli Gao, Zhilong Zhou, Heng Tao Shen, and Jingkuan Song. Bottom-up and top-down: Bidirectional additive net for edge detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020 [scheduled for July 2020, Yokohama, Japan, postponed due to the Corona pandemic]*, pages 594–600, 2020.

[15] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–785, 2018.

[16] 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.

[17] Yuan Hu, Yunpeng Chen, Xiang Li, and Jiashi Feng. Dynamic feature fusion for semantic edge detection. *arXiv preprint arXiv:1902.09104*, 2019.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[19] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2019.

[20] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9258–9267, 2019.

[21] Yun Liu, Ming-Ming Cheng, Deng-Ping Fan, Le Zhang, JiaWang Bian, and Dacheng Tao. Semantic edge detection with diverse deep supervision. *arXiv preprint arXiv:1804.02864*, 2018.

[22] J Mehena. Medical image edge detection using modified morphological edge detection approach. 2019.

[23] Mukta Prasad, Andrew Zisserman, Andrew Fitzgibbon, M Pawan Kumar, and Philip HS Torr. Learning class-specific edges for object detection and segmentation. In *Computer Vision, Graphics and Image Processing*, pages 94–105. Springer, 2006.

[24] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. 2018.

[25] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

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

[27] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.

[28] Kokichi Sugihara. *Machine interpretation of line drawings*. The Massachusetts Institute of Technology, 1986.

[29] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[30] 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 International Conference on Computer Vision*, pages 9197–9206, 2019.

[31] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Lost shopping! monocular localization in large indoor spaces. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2695–2703, 2015.

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[33] Tianhan Wei, Xiang Li, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. *arXiv preprint arXiv:1907.12347*, 2019.

[34] Dan Xu, Wanli Ouyang, Xavier Alameda-Pineda, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. *arXiv preprint arXiv:1801.00524*, 2018.

[35] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. *arXiv preprint arXiv:2008.03898*, 2020.

[36] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. *arXiv preprint arXiv:1905.06549*, 2019.

[37] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[38] Xin Yu, Sagar Chaturvedi, Chen Feng, Yuichi Taguchi, Teng-Yok Lee, Clinton Fernandes, and Srikumar Ramalingam. Vlase: Vehicle localization by aggregating semantic edges. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3196–3203. IEEE, 2018.

[39] Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5964–5973, 2017.

[40] Zhiding Yu, Weiyang Liu, Yang Zou, Chen Feng, Srikumar Ramalingam, BVK Vijaya Kumar, and Jan Kautz. Simultaneous edge alignment and learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 388–404, 2018.

[41] 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 Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019.

[42] Mingmin Zhen, Jinglu Wang, Lei Zhou, Shiwei Li, Tianwei Shen, Jiaxiang Shang, Tian Fang, and Long Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13666–13675, 2020.

[43] Dongchen Zhu, Jiamao Li, Xianshun Wang, Jingquan Peng, Wenjun Shi, and Xiaolin Zhang. Semantic edge based disparity estimation using adaptive dynamic programming for binocular sensors. *Sensors*, 18(4):1074, 2018.