Beyond Data Augmentations: Generalization Abilities of Few-Shot Segmentation Models

Muhammad Ahsan[1][a], Guy Ben-Yosef[2][b] and Gemma Roig[1][c]
[1] *Institute of Computer Science, Goethe University Frankfurt, Germany*
[2] *GE Research, U.S.A.*

Keywords: Machine Vision, Deep Neural Networks, Meta-Learning, Few-Shot Learning, Semantic Segmentation.

Abstract: Few-shot learning in semantic segmentation has gained significant attention recently for its adaptability in applications where only a few or no examples are available as support for training. Here we advocate for a new testing paradigm, we coin it half-shot learning (HSL), which evaluates model's ability to generalise to new categories when support objects are partially viewed, significantly cropped, occluded, noised, or aggressively transformed. This new paradigm introduces challenges that will spark advances in the field, allowing us to benchmark existing models and analyze their acquired sense of objectness. Humans are remarkably exceptional at recognizing objects even when partially obstructed. HSL seeks to bridge the gap between human-like perception and machine learning models by forcing them to recognize objects from incomplete, fragmented, or noisy views - just as humans do. We propose a highly augmented image set for HSL that is built by intentionally manipulating PASCAL-5^i and COCO-20^i to fit this paradigm. Our results reveal the shortcomings of state-of-the-art few-shot learning models and suggest improvements through data augmentation or the incorporation of additional attention-based modules to enhance the generalization capabilities of few-shot semantic segmentation (FSS). To improve the training method, we propose a channel and spatial attention module (Woo et al., 2018), where an FSS model is retrained with attention module and tested against the highly augmented support information. Our experiments demonstrate that an FSS model trained with the proposed method achieves significantly a higher accuracy (approximately 5%) when exposed to limited or highly cropped support data.

1 INTRODUCTION

Deep convolutional neural networks (CNNs) have driven significant progress in various computer vision tasks like image classification, semantic segmentation, and object detection during the past several years (Lu et al., 2021). Recent advancement in CNNs cover progress in layer design (Srivastava et al., 2015), activation and loss functions (Janocha and Czarnecki, 2017), regularization (Moradi et al., 2020), optimization and computational speed (Cheng et al., 2018). However, gathering enough labeled data is notoriously tedious particularly for dense prediction tasks like instance segmentation and semantic segmentation (Gu et al., 2018). Few-shot learning was introduced to mitigate this frequent lack of annotated data.

The goal in few-shot learning (FSL) is to learn a new concept representation from only a few annotated examples. This is achieved by learning feature representations via meta-learning, thus being able to generalize the new unseen classes (Hu et al., 2018). For few-shot segmentation (FSS), the input to the model includes a query image Q as well as k support images $\{S_i\}$ and k masks $\{M_i\}$ in which a given single object class C is annotated. The model then returns a segmentation mask of the class C in the query image Q. Typically, the class C is not seen during training, namely the set of dataset classes S is split to two disjoint sets, S^{train} (seen classes) and S^{test} (unseen classes), and during inference $C \in S^{test}$. The goal of our work is to explore the limitations of current FSS models, and to gain insights for further developing novel improved architectures. To achieve our goal, we are deep-diving into a few of the recent approaches suggested for FSS tasks, namely prototype learning to segment an object (Liu et al., 2020), learning through

[a] https://orcid.org/0009-0000-5982-1979
[b] https://orcid.org/0000-0002-4368-0750
[c] https://orcid.org/0000-0002-6439-8076

Ahsan, M., Ben-Yosef, G. and Roig, G.
Beyond Data Augmentations: Generalization Abilities of Few-Shot Segmentation Models.
DOI: 10.5220/0013179200003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 430-438
ISBN: 978-989-758-728-3; ISSN: 2184-4321

fixed background for different foreground objects and vice versa (Lang et al., 2022), a model training technique to make new-class adaptation more manageable with the class weight transformer (Lu et al., 2021), image segmentation learning with task-specific edge detection (Chen et al., 2016), and prototype alignment networks (Wang et al., 2019a). To evaluate generalization capabilities of FSS models, we introduce HSL which tests the ability of a model to generalise to unseen categories when the support information is highly augmented or limited (see Fig.1 and Sec. 3.1). Based on insights from HSL, we propose a novel training method to improve the generalization ability of FSS models when exposed to highly augmented support information. We integrate FSS model with a channel & spatial attention module CBAM (Woo et al., 2018), which benefits when highly augmented or partial object information is used as a support for training. Our primary contributions can be summarized as fellows:

- We propose a challenging testing paradigm, called HSL, for FSS models to evaluate their ability to learn from partial object information.
- We propose a training method incorporating a channel & spatial attention module (CBAM) to improve the models' performance in HSL scenarios.
- We use Grad-CAM, a visualization technique that leverages gradients to identify the importance of spatial locations within convolutional layers.

2 RELATED WORK

The challenge of FSL has been an active area of research for many years. In this work, we explore several key components crucial to our study. First, we discuss FSL, a paradigm that enable model to generalize well from a limited number of training examples where acquiring large labeled dataset is impractical or expensive. Next, we delve into FSS, a specialized application of FSL focused on accurately segmenting different objects in images using only a few annotated examples. Finally we examine the attention module, a mechanism that improve the model performance by allowing it to focus on most relevant input features.

2.1 Few-Shot Learning

FSL is the task of training models to generalize from a small number of labeled examples to correctly classify or segment unseen samples. The main focus of FSL is developing machine learning models suitable for the real-world scenarios where obtaining a large dataset is impractical or expensive. Most of the current approaches in the FSL domain are based on a meta-learning framework, where a base learner adapts to new learning tasks derived from a base dataset to simulate few-shot scenarios (Wang et al., 2019b).

In real-world applications, we are often confronted with incomplete or imperfect data. While FSL aims to address scenarios with limited examples, it still assumes that the available data are reasonably complete and high-quality. In contrast, HSL introduces the notion of training and testing models with significantly imperfect or partial data.

2.2 Few-Shot Segmentation

FSS addresses the challenge of segmenting new classes with limited annotated data, crucial in domains like medicine and agriculture. (Catalano and Matteucci, 2024). In FSS, a model learns to identify pixels in a query image that belong to a specific object class, guided by the segmentation masks from only a small number of support images (Li et al., 2021). Traditional semantic segmentation models typically rely on a significant amount of labeled data to achieve good results and generally struggle to adapt to unseen classes without additional fine-tuning. In response, several robust network architectures have been developed, incorporating key techniques like SegGPT as a generalist segmentation model that unifies various segmentation tasks into an in-context learning framework (Wang et al., 2023a) dilated convolutions (Yu and Koltun, 2015), encoder-decoder frameworks (Ronneberger et al., 2015), multi-level feature aggregation (Lin et al., 2017), and attention modules (Huang et al., 2019). Previous studies typically approach FSS as a guided segmentation task. For instance in (Hu et al., 2018), a base learner (support branch) processes the support information to generate parameters that guide the meta-learning framework in predicting the mask for query images. (Zhang et al., 2020) introduced masked average pooling to extract support features, which became the foundational technique in FSS tasks. Due to the success of prototypical networks, (Zhang et al., 2020) propose a dense prototype learning for segmentation and query mask prediction. In our work we analyze that how training with augmented support samples influences robustness and generalization abilities. For example, in (Hu et al., 2018) a late fusion is proposed where the support image branch predicts the weights of the top layer of the query image branch. The Prototypical learning approach is another method used for FSS which aims to predict foreground and background classes by their similarity to learned prototypes (Wang et al., 2019a). The PPNet model (Liu et al., 2020) performs prototype learning based on a decomposition of the holistic object class into a set of part-aware prototypes. The BAM model (Lang et al., 2022) introduces a new parallel branch base learner to the meta learner which is to identify base classes and distinguishes the regions of base classes from novel classes that do not need to be segmented during inference.

2.3 Attention Module

CNN models are tried to be improved through multiple approaches, like developing a specialized optimizer (Rakelly et al., 2018), introducing adversarial training methods (Wang et al., 2023b), or designing specialized meta-architectures (Hu et al., 2018). Another approach is to use attention blocks to enhance performance by re-calibrating channel-wise feature responses through modeling channel interdependencies. Another lightweight module, called bottleneck attention module, is designed to enhance performance by introducing the attention along both channels and spatial axes. We proposed to incorporate and adapt the CBAM (Woo et al., 2018) in FSS models, which adjusts weights based on the features of the input data.

3 METHODS

For assessing HSL we perform a series of transformations on the support samples of two benchmark PASCAL-5^i, COCO-20^i datasets. In this section we presents the datasets, augmentations, training and testing paradigms that we proposed to assess the attention module and HSL task.

3.1 Half-Shot Learning

HSL introduces a more challenging scenario, where the model is exposed to only a portion of the support objects, which are either partially visible, heavily cropped or aggressively transformed. While state of the art FSS models CWT (Lu et al., 2021), BAM (Lang et al., 2022), PPNet (Liu et al., 2020), and PANet (Wang et al., 2019a) have shown significant progress in standard benchmarks, they are still sensitive to modifications in the support information. In this work, we focus on exploring the reduction of support information, its effects, and the robustness of existing models. For investigating HSL the task of semantic segmentation is a natural choice, since segmentation aims to divide an image into segments or regions, each of which represents a separate object or part of the image. The HSL task allows us to ask whether and how much the model really learns to represent transformable knowledge about general object structure, or if it rather has a strong focus on alignment of the support mask to the image. We illustrate the HSL semantic segmentation in Figure 1. HSL goes beyond few-shot learning by simulating real-world scenarios, where objects are rarely fully visible or perfectly captured. Models trained in HSL are evaluated on their ability to:

- Ignore spurious correlations (e.g., background noise) and focus on core object features.
- Generalize effectively when object data is incomplete or adversarial.
- Handle occlusions and environmental noise, making them more robust and practical for deployment.

HSL provides a more challenging benchmark for evaluating a model's robustness and ability to learn meaningful, core features under difficult, imperfect conditions.

3.2 Datasets and Augmentations

PASCAL-5^i and COCO-20^i are benchmarks used for evaluation of FSS models. Both datasets are derived from larger, well-known datasets (PASCAL VOC and COCO), and they are restructured into subsets specifically designed for FSL tasks.

3.2.1 Datasets

PASCAL-5^i is an extension of PASCAL VOC and also contains annotations from the Simultaneous Detection and Segmentation (SDS) dataset. The train set and test set contains 5,953 and 1,449 images, respectively. The 20 categories in the PASCAL-5^i dataset are divided into four folds (0, 1, 2, 3), and each fold contains 5 disjoint classes. Data instances from three folds are used for model training, and testing performed using the fourth fold in a cross-validation fashion. COCO-20^i is larger and a more challenging dataset designed for different tasks like segmentation, key-point detection, and captioning dataset. The dataset is divided into 4 splits (COCO-20i, where i = (0, 1, 2, 3). The object categories are divided into four folds, each containing 20 distinct classes. It provide 82,081 and 40,137 images for training and evaluation respectively.

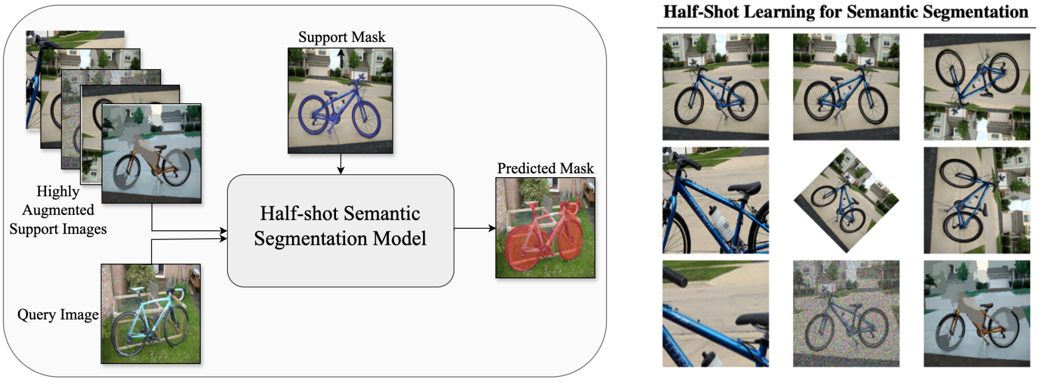
Support Mask
Highly Augmented Support Images
Half-shot Semantic Segmentation Model
Predicted Mask
Query Image
Half-Shot Learning for Semantic Segmentation

Figure 1: Illustrate the HSL for semantic segmentation. In typical FSL scenarios, the support set consists of complete, clean, and fully observable examples of each class. HSL introduces a more difficult scenario, where the model is exposed to half of the support objects that are either partially viewed, significantly cropped, or noised. The goal is to test the model's ability to generalize under adversarial conditions. We propose to use a set of augmentations and perturbations in the support image and mask, while keeping the query in the original form.

3.2.2 Augmentations

We apply the following augmentations as shown in Figure 1 to test HSL on both datasets PASCAL-5^i and COCO-20^i:

- Flip: Both Horizontal-Flip and Vertical-Flip used with probabilities like p=0.5 & 1 that mirrored the objects along the horizontal, vertical or both axis.
- Rotate: We randomly apply affine transformations to scale, translate and rotate the input images. By rotating images in various angles, the dataset becomes more diverse, helping models generalize better by learning rotational invariance. Specifically, we apply four rotate_limits with angles 10°, 20°, 45° and 90° respectively.
- Crop: The center-crop operation focuses on the center of the image, assuming that the most important or relevant information is likely to be in the middle of the image. We apply the center-crop to support images and labels with four different variations, including 20%, 40%, 60%, and 80%.
- Noise: Noise reduces image clarity, making it harder to distinguish details. Gaussian Noise actually sampled the complete noise with all channels of the images. So, it is still imperfect information for the model to learn about the object class.
- Superpixels: With Superpixels we transformed the input images to their superpixel representation partially or completely with p = 0.5 or 1.
- Irrelevant Support: We provide irrelevant support images to the model like support samples have irrelevant category, different from the query image, this also results in poor generalization to new data.

3.3 Models

FSS is a deep learning technique that makes a pre-trained model capable of segmentation of new categories of data that are unseen to the model. We chose the following four FSS models:
CWT: Classifier Weight Transformer model developed to make new-class adaptation more manageable through concentrate on classifier-part rather than to meta-learn the entire complex model (Lu et al., 2021).
BAM: A new perspective on FSS to identify the regions that do not need to be segmented, they proposed an additional branch namely base learner to specifically predict the base class regions (Lang et al., 2022). So, the irrelevant objects in the query images can be concealed significantly. Gram Matrix used to differentiate the image scenes and extend the proposed approach to a setting namely, i.e., generalized FSL, which simultaneously identifies the targets of base and novel classes.
PPNet: Part-aware Prototype Network decompose the holistic class representation into a set of part-aware prototypes. The network consists of three parts, first is Embedding Network to compute the convolutional feature maps of the images, second is Prototypes Generation Network that extracts a set of part-aware prototypes and third is Part-aware Mask Generation Network that generates the final semantic segmentation of the query images (Liu et al., 2020).
PANet: The PANet, prototype alignment network, where they learn class-specific high quality prototypes with non-parametric metric learning from a few support samples (Wang et al., 2019a). They also present a prototype alignment regularization among support and query images.

3.4 Training / Testing Paradigms

CWT, BAM, PPNet, and PANet as our baseline FSS models to evaluate their performance on the HSL task. Experiments performed with pre-trained and retrained models with our customized settings, incorporating data augmentations.

3.4.1 Train Normal and Test with Augmentation

Models are trained using default configurations and reproduced the original results for all of the tested standard FSS models. We subsequently applied various augmentations to the support images to evaluate the robustness and generalization capabilities of FSS models.

3.4.2 Train and Test with Augmentations

Models are trained using highly augmented support images, which provide limited information for the models to learn from and also testing them against the augmented dataset.

3.4.3 Train with Attention Module

To enhance the resilience and ability to learn from limited support information, we suggest incorporating the Convolutional Block Attention Module, CBAM (Woo et al., 2018), which adjusts weights based on the input features. This boost the representation capacity by using attention modules, emphasizing key features while reducing the focus on less relevant ones (Xu et al., 2015; Gregor et al., 2015). CBAM has two sequential components, channel and spatial, which dynamically refine the intermediate feature map at every convolutional block. CBAM infers a 1D channel attention map $M_c \in R^{C\times1\times1}$ and a 2D spatial attention map $M_s \in R^{1\times H\times W}$ when given an intermediate feature map $F \in R^{C\times H\times W}$ as input as illustrated below:

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F' \end{aligned} \tag{1}$$

where F'' is the final refined output and $\otimes$ denotes element-wise multiplication. Spatial information of a feature map combined by using both average-pooling and max-pooling: F^c_{avg} and F^c_{max}, and compute the output feature vectors using element-wise summation. The channel attention is computed as:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_o(F^c_{avg})) + W_1(W_o(F^c_{max}))) \end{aligned} \tag{2}$$

where W_o, W_1 are the MLP weights, ReLU activation function is followed by W_o and σ denotes the sigmoid function. Channel information of a feature map aggregated by using two pooling operations performed to generate 2D maps: F^s_{avg} and F^s_{max}, these maps are concatenated and convolved by a standard convolution layer to prodcuce 2D spatial attention map. The spatial attention is computed as:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{3}$$

where σ denotes the sigmoid function and $f^{7\times7}$ represents a convolution operation.

4 EXPERIMENTAL RESULTS

Experimental results involved training and testing paradigms to investigate augmentation influence using the benchmark datasets PASCAL-5^i and COCO-20^i described in details.

4.1 Train Normal and Test with Augmentation

Models are trained using standard configurations and tested against HSL method and presented results in Table 1.

All four models performed well in the flip experiments; however, PPNet and PANet experienced accuracy losses of 6% & 9% respectively. In the Shift-Scale-Rotate experiments, CWT and BAM performed better, while PPNet and PANet decline in accuracy as the rotation angles increased from 10° to 90°. All FSS models experienced significant accuracy losses with cropped support images. The BAM model, which outperformed the other FSS models, lost approximately 14% accuracy. The models unable to learn sufficiently from the partial or incomplete support information. All FSS models experienced a drop in performance when exposed to noisy data. Among them, BAM lost 15% accuracy; however, it still outperformed the other models. Compared to noise, all models demonstrated better performance with the superpixel representation of the support images. Models tend to be more confused by the irrelevant support but perform better when provided with partial views.

4.2 Train and Test with Augmentation

Training and testing performed with highly augmented support information alongside standard query images. Table 1. Models retrained with augmentations demonstrate only a slight increase in accuracy of 1-2% in some cases compared to those that were not

Table 1: Performance comparison of different Few-Shot Segmentation (FSS) models on highly augmented PASCAL-5i data. The white column represents the mIoU performance when the models are tested on the augmented dataset, while the pink column shows the performance when the models are both trained and tested on the augmented dataset.

#	Augmentation	CWT	CWT w/ Aug	BAM	BAM w/ Aug	PPNet	PPNet w/ Aug	PANet	PANet w/ Aug
0	Baseline	56.40	– –	67.81	– –	55.16	– –	48.10	– –
1	Hor-Flip(p=0.5)	56.34	56.32	67.76	67.80	51.62	52.60	46.46	46.57
2	Hor-Flip(p=1)	55.30	56.24	67.56	67.61	46.88	46.54	38.08	38.41
3	Ver-Flip(p=0.5)	56.18	56.20	67.76	67.72	51.42	52.50	46.34	46.43
4	Ver-Flip(p=1)	55.48	56.71	66.39	67.62	44.39	44.61	37.28	38.74
5	Sh-Rot(L=10)	55.16	56.85	67.60	67.70	46.26	46.11	36.73	36.17
6	Sh-Rot(L=20)	54.50	55.81	66.49	67.50	45.67	45.72	35.46	35.50
7	Sh-Rot(L=45)	54.84	55.65	66.22	67.11	44.21	44.83	34.52	35.82
8	Sh-Rot(L=90)	53.71	55.62	64.40	66.83	44.19	44.27	35.11	36.22
9	C-Crop(20%)	37.20	39.19	53.59	55.82	26.13	26.71	23.34	23.54
10	C-Crop(40%)	46.41	48.76	64.00	65.29	41.56	41.37	31.66	32.91
11	C-Crop(60%)	51.80	53.80	66.83	67.91	42.77	42.61	36.05	37.16
12	C-Crop(80%)	55.49	56.29	67.66	67.41	42.65	42.51	37.0	37.84
13	GaussNoise(p=.5)	28.96	29.65	52.26	52.02	19.54	19.17	21.90	22.82
14	Superpixels(p=.5)	40.53	41.37	63.72	65.18	35.72	36.27	31.39	32.63
15	Irrel-Support	31.73	31.16	47.86	48.61	23.27	23.48	24.62	24.42

retrained with augmentations. An interesting insight is that the models do not effectively utilize partial support views even when trained to do so as shown in Figure. 2 as well.

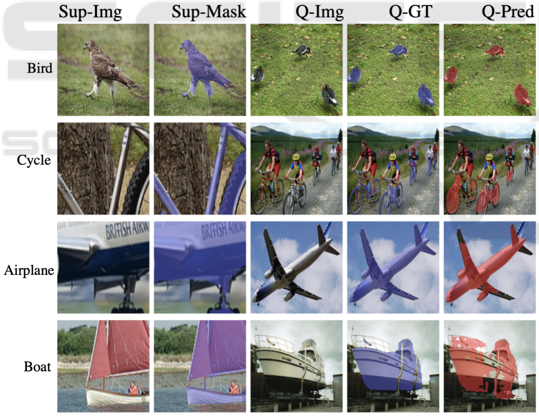

Figure 2: Segmentation results using the proposed method under HSL settings. The method is applied with the BAM model on the PASCAL-5^i dataset. First row displays samples without any augmentation and remaining rows shown results obtained with highly cropped dataset where airplanes and boats indicates that predictions are less accurate compared to the ground truth.

Table 2: Training and sesting of FSS models on augmented COCO-20^i dataset.

#	Augmentation	CWT	CWT w/ Aug	BAM	BAM w/ Aug
0	Baseline	32.90	– –	46.23	– –
1	Hor-Flip(p=0.5)	31.64	32.21	45.27	46.22
2	Hor-Flip(p=1)	31.42	31.55	45.19	46.13
3	Ver-Flip(p=0.5)	32.12	32.43	45.20	46.26
4	Ver-Flip(p=1)	28.74	29.48	44.06	46.58
5	Sh-Rot(L=10)	30.50	30.13	45.36	46.29
6	Sh-Rot(L=20)	29.18	30.27	45.23	46.38
7	Sh-Rot(L=45)	29.55	29.52	45.07	46.87
8	Sh-Rot(L=90)	28.38	29.31	44.51	45.31
9	C-Crop(20%)	19.63	19.86	34.72	36.38
10	C-Crop(40%)	25.18	26.63	36.25	38.56
11	C-Crop(60%)	28.96	29.19	42.27	44.17
12	C-Crop(80%)	31.14	31.73	44.81	47.96
13	GaussNoise(p=.5)	22.06	22.19	36.70	38.47
14	Superpixels(p=.5)	25.10	25.41	37.36	38.82
15	Irrel-Support	17.66	17.89	28.89	30.31

Table 2 presents the performance of CWT and BAM models with another benchmark COCO-20^i. Experiments demonstrate that both models exhibit varying degrees of robustness when subjected to different types of augmentations. In Flip and Rotate, models demonstrated better performance, with only 2% decrease in accuracy as the rotation angle varied from 10° to 90°. Exp#9-12 demonstrate that BAM consistently outperforms CWT when dealing with highly cropped data. Models similarly struggled in the Gaussian Noise experiment, CWT performs poorly with an IoU of 22.19, while BAM shows a stronger performance with an IoU of 38.47. With superpixel representations, CWT achieves an IoU of 25.41, while BAM again outperforms with an IoU of 38.82. This shows that both models can handle superpixel data better than noisy data, but BAM maintains a clear advantage. CWT with irrelevant support, achieving an IoU of only 17.89, while BAM performs better with an IoU of 30.31. This indicate that both models are confused by irrelevant support, but BAM is more robust in these challenging conditions.

4.3 Train with Attention Module

Upon analysis, we observed that BAM consistently demonstrates superior performance compared to the other models, particularly when tested against highly cropped images. This indicates BAM's better generalization abilities towards incomplete or partial support information. Therefore, for further experimentation, we selected BAM model to extend with the integration of an attention module, CBAM, which serves as a simple yet effective weight adjustment mechanism based on the features of the input data. Attention modules have proven to be effective in various visual tasks such as image classification, object detection, and semantic segmentation (Woo et al., 2018). Models utilizing VGG/ResNet as a backbone, such as BAM, can jointly train the combined CBAM-enhanced networks, integrated CBAM with the Res-Blocks in ResNet (He et al., 2016). Representation capacity enhanced through attention modules which prioritizing important features while minimizing irrelevant ones (Xu et al., 2015; Gregor et al., 2015). CBAM adaptively refines the intermediate feature map at each convolutional block of deep networks (see Fig 3). Table 3 clearly depicts generalization capability of the BAM model enhanced against augmented support information. Training of BAM with attention module achieves approximately 2% - 5% higher accuracy when tested with rotated and highly cropped data (exp# 4,5 in Table 3).

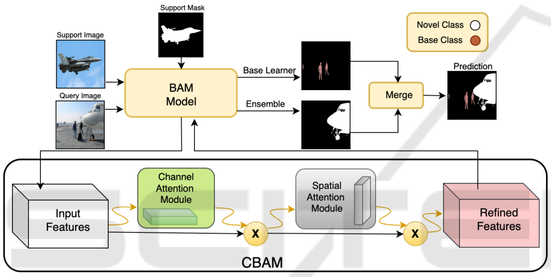

Figure 3: BAM model with an attention module. The module consists of two sequential sub-modules: channel and spatial. The intermediate feature map is adaptively refined through the CBAM module at each convolutional block of deep networks.

We employ Grad-CAM (Selvaraju et al., 2020) as a visualization technique that leverages gradients to assess the importance of spatial locations within convolutional layers. Grad-CAM activation highlights specific regions of the input image using an attention heatmap, indicating the areas that are most crucial for detecting a particular class of interest (Selvaraju et al., 2020). For qualitative analysis, we compare the visualization results of baseline BAM and BAM with the attention module. Figure 4 illustrates the results of Grad-CAM, demonstrating BAM model integrated with the attention module provides slightly better insightful explanations.

Table 3: Performance comparison of BAM model and BAM with Attention module on the highly augmented PASCAL-5i dataset. The first column shows the mIoU performance of the BAM model when tested on the augmented dataset, while the second pink column presents the improved performance of the BAM+Attention model on the same test data.

#	Augmentation	BAM	BAM w/ Attention
0	Baseline	67.81	– –
1	VerticalFlip(p=1)	66.39	67.81
2	Hor-Flip(p=1)	67.59	67.96
3	Sh-Rot(L=20)	66.53	67.29
4	Sh-Rot(L=90)	64.41	67.74
5	C-Crop(20%)	53.57	58.69
6	GaussNoise (p=1)	52.41	53.48
7	Superpixels (p=1)	63.39	65.71

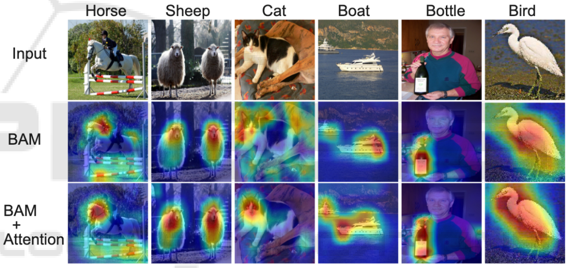

Figure 4: Examples of feature importance visualizations.

5 CONCLUSION

Design and evaluation of the HSL task, where support images contain only partial object information. In comparison to control case, where no support image is provided namely Irrelevant-Support, all tested models exhibited improved performance. This indicates that all models were able to leverage support information, even when it was significantly limited. Variations in model performance were observed due to differences in model architectures and the various backbones employed in the implementation. In our case, the selected models have variations with the backbones from VGG-16 to Resnet-101, which may have different capability to generalize from partial to full objects. In several experiments (e.g., Center-Crop(20%), Noise, and Irrelevant-Support), the tested models struggled to accurately identify the confusing areas in the support images, which hindered the improvement of the meta-learner's predictions. The part-based and prototype models tested here were struggling to extract robust prototypes from the support set with less or incomplete information to learn about the object class. A new training paradigm, referred to as BAM with attention, has been proposed. In this approach, BAM model with an attention module is re-trained in conjunction with attention module and evaluated using highly augmented support images. Although, it still face challenges in extracting robust features from the support set, it demonstrates less confusion and greater capacity for generalization compared to other models. We believe that our findings can illuminate future investigations into the issues of bias or semantic ambiguity problems.

REFERENCES

Catalano, N. and Matteucci, M. (2024). Few shot semantic segmentation: a review of methodologies, benchmarks, and open challenges.

Chen, L.-C., Barron, J. T., Papandreou, G., Murphy, K., and Yuille, A. L. (2016). 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.

Cheng, J., Wang, P.-s., Li, G., Hu, Q.-h., and Lu, H.-q. (2018). Recent advances in efficient computation of deep convolutional neural networks. *Frontiers of Information Technology & Electronic Engineering*, 19:64–77.

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., and Wierstra, D. (2015). Draw: A recurrent neural network for image generation. cite arxiv:1502.04623.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W. (2019). Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612.

Janocha, K. and Czarnecki, W. M. (2017). On loss functions for deep neural networks in classification. *arXiv preprint arXiv:1702.05659*.

Lang, C., Cheng, G., Tu, B., and Han, J. (2022). 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 (CVPR)*, pages 8057–8067.

Li, G., Jampani, V., Sevilla-Lara, L., Sun, D., Kim, J., and Kim, J. (2021). 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.

Lin, G., Milan, A., Shen, C., and Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934.

Liu, Y., Zhang, X., Zhang, S., and He, X. (2020). Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer.

Lu, Z., He, S., Zhu, X., Zhang, L., Song, Y.-Z., and Xiang, T. (2021). Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*.

Moradi, R., Berangi, R., and Minaei, B. (2020). A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986.

Rakelly, K., Shelhamer, E., Darrell, T., Efros, A. A., and Levine, S. (2018). Few-shot segmentation propagation with guided networks. *arXiv preprint arXiv:1806.07373*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. *Advances in neural information processing systems*, 28.

Wang, K., Liew, J. H., Zou, Y., Zhou, D., and Feng, J. (2019a). Panet: Few-shot image semantic segmentation with prototype alignment. In *The IEEE International Conference on Computer Vision (ICCV)*.

Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., and Huang, T. (2023a). Seggpt: Segmenting everything in context.

Wang, Y., Luo, N., and Zhang, T. (2023b). Focus on query: Adversarial mining transformer for few-shot segmentation. *Advances in Neural Information Processing Systems*, 36:31524–31542.

Wang, Y.-X., Ramanan, D., and Hebert, M. (2019b). Meta-learning to detect rare objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9925–9934.

Woo, S., Park, J., Lee, J., and Kweon, I. S. (2018). CBAM: convolutional block attention module. *CoRR*, abs/1807.06521.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zhang, X., Wei, Y., Yang, Y., and Huang, T. S. (2020). Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE transactions on cybernetics*, 50(9):3855–3865.