SimPropNet: Improved Similarity Propagation for Few-shot Image Segmentation

Siddhartha Gairola[2*], **Mayur Hemani**[1 †], **Ayush Chopra**[1 †] and **Balaji Krishnamurthy**[1]
[1]Media and Data Science Research Lab, Adobe Experience Cloud
[2]IIIT Hyderabad
siddhartha.gairola@research.iiit.ac.in, {mayur, ayuchopr, kbalaji}@adobe.com

Abstract

Few-shot segmentation (FSS) methods perform image segmentation for a particular object class in a target (query) image, using a small set of (support) image-mask pairs. Recent deep neural network based FSS methods leverage high-dimensional feature similarity between the foreground features of the support images and the query image features. In this work, we demonstrate gaps in the utilization of this similarity information in existing methods, and present a framework - *SimPropNet*, to bridge those gaps. We propose to jointly predict the support and query masks to force the support features to share characteristics with the query features. We also propose to utilize similarities in the background regions of the query and support images using a novel foreground-background attentive fusion mechanism. Our method achieves state-of-the-art results for one-shot and five-shot segmentation on the PASCAL-5^i dataset. The paper includes detailed analysis and ablation studies for the proposed improvements and quantitative comparisons with contemporary methods.

1 Introduction

Semantic image segmentation assigns class labels to image pixels. It finds applications in image editing [1, 20, 21], medical diagnosis [8, 13, 18], automated driving [7] etc. Supervised deep neural network methods such as [2, 3, 4, 5, 6, 23, 27] enable highly accurate image segmentation. However, they work for only a small number of fixed object classes, and require a large number of image-mask pairs for training which are hard to manually annotate. In several practical scenarios, including online commerce and design, the images may exist in a large number of sparsely populated classes (for instance, images of products). In such cases obtaining image-mask pairs for all possible classes to train a supervised method may be infeasible. Thus, segmentation methods that generalize to new classes with scant training data are of significance. Few-shot image segmentation methods, like [15, 24, 26], are class-agnostic and alleviate the need for a large number of image-mask pairs. These methods utilize additional images and their masks of a particular class in predicting the binary segmentation mask for a given image of the same class. This work proposes a new few-shot segmentation framework that seeks to alleviate a fundamental limitation in existing techniques and significantly improves upon the state-of-the-art.

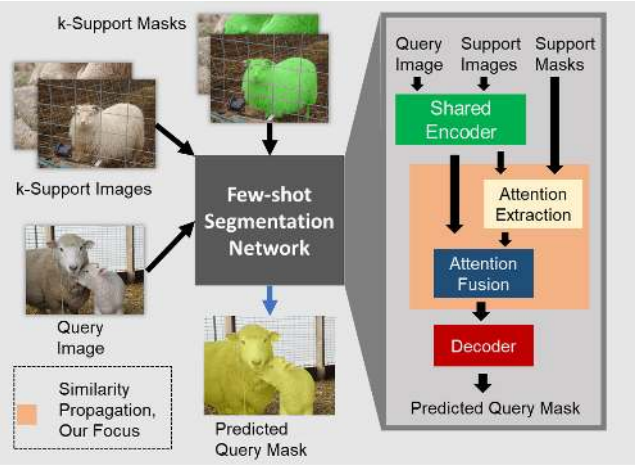

Figure 1: Few-shot Image Segmentation: Broad architecture of contemporary methods ([24, 25, 26]). Features from the support images (in the support mask regions) are processed to obtain a probe representation and fused with features from the query image, and decoded to predict the query mask. Improving similarity propagation between the support and query branches is the focus of this work. The yellow mask in the diagram is a (1-shot) result from our method.

Few-shot segmentation (FSS) methods typically take as input - a *query* image for which the segmentation mask is required, a set of *support* images, and their segmentation masks (*support masks*). One-shot segmentation is the extreme setting where only a single support image-mask pair is available. Our method achieves state-of-the-art performance in both one-shot and few-shot settings (about 5% and 4% gain respectively).

Recent deep neural network based FSS methods (like [16, 24, 25, 26]) employ variations of the following broad process (Figure 1):

1. Query and support image features are extracted using

*work done as part of Adobe MDSR internship program
†equal contribution

shared, pre-trained (on ImageNet [14]) network layers.
2. A probe representation for regions of attention in the support image is obtained from the support features and mask(s).
3. The attention representation is fused with the query features.
4. The fused features are decoded to obtain the query mask.

The attention extraction and fusion modules leverage the high-dimensional feature similarity between the query and the support images to selectively decode the query features to output the segmentation mask. The focus of this work is to demonstrate gaps in the propagation of this similarity information in existing methods and to bridge those gaps.

Experiments with two state-of-the-art methods - [16] and [26] (see Section 3.1) reveal that FSS methods make errors in visually similar regions, across image pair samples of identical classes. They perform poorly when support is identical to query inputs as well. These results indicate that the class and visual similarity information is not propagated optimally across the support and query branches .

We predict the support mask from the support features to endow the features with specificity with respect to the target class, which in turn aids in similarity propagation between the support and query. We also leverage the similarities in the background regions of the support and query images through a background attention vector computed from the support features and inverse mask, and fuse it with the query features. Finally, to prevent the network from overfitting on training class-conditional similarities, we employ an input channel averaging input augmentation for the query input. With these improvements we achieve state-of-the-art performance on PASCAL 5^i dataset for both one-shot and five-shot segmentation tasks.

The contributions of this work can be summarized as follows:

1. It highlights gaps in existing FSS methods in fully utilizing the similarity between the support and query images.
2. It introduces *SimPropNet*, a few-shot segmentation framework that bridges those gaps and achieves state-of-the-art performance on the PASCAL 5^i dataset in both 1-shot and 5-shot evaluation settings. The framework employs:
   a. A dual-prediction scheme (DPr) where the query and support masks are jointly predicted using a shared decoder, which aids in similarity propagation between the query and support features.
   b. A novel foreground-background attentive fusion (FBAF) mechanism for utilizing cues from background feature similarities between the query and support images.

The next section places the proposed method in the context of previous work related to the problem.

2 Related Work

Recent methods for few-shot semantic segmentation (FSS) build a framework for one-shot segmentation and subsequently construct methods to use the framework for k-shot segmentation ($k > 1$). The proposed work follows the same methodology.

As described in Section 1, most FSS methods employ a dual branched neural network model with a support branch and a query branch. This model was introduced by Shaban *et al.* [15] where the support branch is conditioned on the support input to predict the weights of the last layer in the query branch which then predicts the query mask.

Rakelly *et al.* [12] improve upon [15] by employing a *late fusion* strategy for the support mask and support feature maps to segment the query image. Late fusion adopts a parametric module in a two branch setting, which fuses information extracted from the support set with the features of the query image to produce the segmentation mask. Dong and Xing [9] combine the late fusion methodology with the idea of prototypical networks introduced in [19] to learn a metric space where distances to *prototypes* represent class-level similarity between images. Their method uses the prototypes as guidance to fix the query set segmentation rather than obtaining segmentation directly from metric learning. Zhang *et al.* [25] take a different approach to late fusion. They introduce *masked average pooling* (MAP) that pools support features of the regions of interest in the support images, and fuses them with the query features using vector cosine-similarity. This *attention* map is then decoded to predict the query's segmentation maps. Building on [9], Wang *et al.* [24] combine Prototypical Learning [19], and MAP [25] to incorporate the support information into the segmentation process. Zhang *et al.* [26] adopt a learnable method through an attention mechanism to fuse support information from multiple support examples along with iterative refinement. Their method also uses MAP, but instead of fusing its output with the query features using cosine-similarity, they concatenate it with the query features and then decode the output.

These methods for few-shot segmentation depend on the support set to produce the segmentation for the query image, but fail to fully utilize the support information efficiently. We present evidence for this gap in Section 3.1 and subsequently propose SimPropNet, a few-shot segmentation framework that seeks to bridge this gap and improve performance.

3 Method

In this section, we first establish firm ground to motivate our work, and then elucidate the proposed method in the one-shot segmentation setting. Subsequently, we demonstrate how we adapt the method for k-shot segmentation setting.

3.1 Premise Validation

We first present experimental evidence validating the premise that there are gaps in similarity propagation between the support and the query images. This argument is expressed as follows:

1. FSS methods make errors for image regions that may not be inherently hard to segment.
2. The regions of error in the support and query images are class-conditionally similar.
3. Even with maximally similar support and query images, current FSS methods are unable to produce good predictions.

We conduct experiments to corroborate this argument which

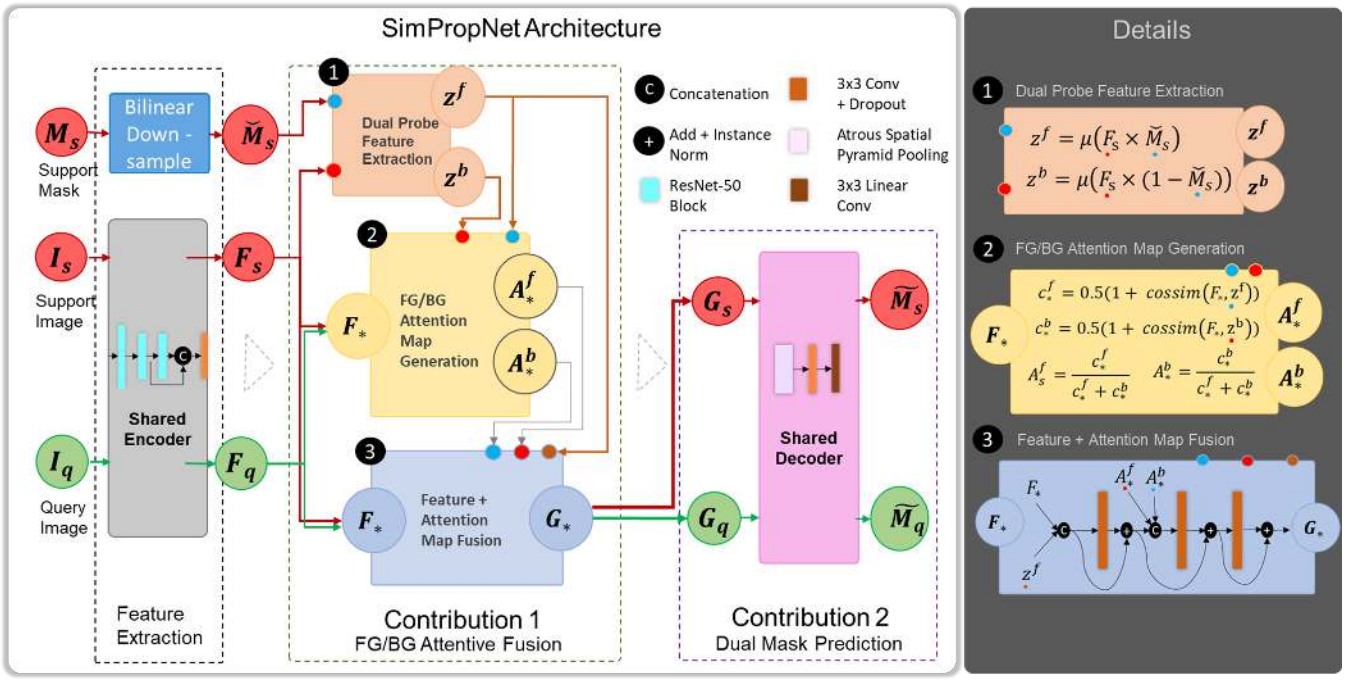

Figure 2: SimPropNet Architecture: The support and query images are encoded to F_s, F_q respectively, with pre-trained ResNet-50 layers and a single convolutional layer (as described in [26]). The support features and the support mask are used to compute the foreground(FG) and background(BG) MAP vectors (Z^f, Z^b) where μ is average pooling. They are used to compute the four attention maps A_s^f,A_s^b, A_q^f, and A_q^f (FG and BG for both support and query). The feature+attention fusion module combines (F_s, Z^f, A_s^f, and A_s^b) to obtain G_s, and identically, (F_q, Z^f, A_q^f, and A_q^b) to obtain G_q. These fused features (G_s, G_q) are finally decoded with the atrous spatial-pyramid pooling and convolutional layers to obtain the predicted masks $\tilde{M}_s$ and $\tilde{M}_q$ respectively.

indicates an opportunity to improve similarity sharing between the support and query images.

We also provide evidence to show that the background regions of the support and query images may be similar and could hold cues for improved segmentation of the query. The experiments use author-provided implementations of the recent state-of-the-art methods [16] and [26].

Errors by FSS versus Supervised Methods We measure the overlap (Table 1) between the error regions of output masks from two FSS methods ([16] and [26]) with the error regions of masks produced by DeepLab v3+ [6] (a state-of-the-art supervised method). The overlap between the error regions for FSS methods and DeepLab v3+ is very small, while the gap in the mean-IoU for correct predictions (DeepLab v3+ versus FSS methods) is large. This indicates that there are image regions where supervised methods like DeepLab v3+ do not fail (but FSS methods do), and that they are not characteristically difficult to segment.

Similarity of mispredicted regions To determine the similarity between regions of the query and support images where [26] makes errors, we compute the Masked Average Pooling vectors (as described in [25]) using pre-trained VGG-16 [17] features. This is done for several pairs of images of identical classes. for the following two regions of each image:

a. the pixels covered by the ground-truth masks (Z^g), and
b. the pixels present in the regions mispredicted by the FSS method (Z^e).

Method	FN Overlap	FP Overlap	TP Gap
CANet	18.32	9.11	23.28
AMP	10.99	19.06	39.41

Table 1: Percentage (%) error overlap between Deeplab v3+ (DLv3+) [6], and FSS Methods (CANet [26], and Adaptive Masked Proxies (AMP) [16]). FN = False Negative (missed regions), FP = False Positives (incorrectly predicted as part of the mask). TP Gap = Overall gap in true prediction made by Deeplab v3+ and FSS methods. Low error overlap and high prediction gap indicates that FSS methods make different mistakes than DLv3+.

For each image pair (A, B) of a class, the ratio of the inner-products of the MAP vectors for the error regions and the ground-truth mask regions - $(Z^e(A) \cdot Z^e(B))/Z^g(A) \cdot Z^g(B)$ - is a measure of the similarity of the corresponding error regions relative to the similarity of the ground-truth mask regions. We measure the ratio over a 1000 pairs of images from the PASCAL VOC dataset [10] and find the average to be 0.87 (std. dev. = 0.25). The high value of relative similarity of error regions substantiates the claim that errors are committed in regions of similarity that could have possibly been propagated from the support to the query.

Identical Inputs: (Support = Query) Table 2 reports results from a third experiment with [16] and [26] where the same image is given as input for both support and query (including the ground-truth mask for the support). This con-

stitutes a basic test for similarity propagation in the network. Ideally, the network must produce the exact mask as provided in the support input, because the query and the support are as similar as possible. However, both methods ([16] and [26]) perform poorly for these inputs indicating the loss of similarity information in the networks.

Method	split-1	split-2	split-3	split-4
CANet	54.51	63.98	48.20	52.76
AMP	54.41	69.34	64.79	60.02

Table 2: Percentage (%) mean-IoU measured for FSS methods (CANet [26] and AMP [16]) for the test images from PASCAL 5^i dataset when query image (I_q) = support image (I_s).

Cues from Background Similarity Figure 3 indicates the degree of similarity measured as the cosine-similarity between the MAP vectors of foreground and background regions for the pairs of images. The figure indicates that images of identical object classes have higher degree of feature similarity in their background features, than in the actual foreground features. This represents an opportunity to collect more information for accurate segmentation.

The results of these experiments indicate the gaps in similarity propagation in existing FSS methods that we propose to exploit. Next, we describe the organization of the proposed network and how it addresses the similarity propagation problem.

3.2 Network Organization

Figure 2 illustrates the architecture of SimPropNet. The network is organized into two branches (support and query) with a shared encoder, a fusion module and a shared decoder. For our experiments, we use a ResNet-50 [11] based feature extractor, and atrous spatial pyramid pooling based decoder as in [26]. The encoder comprises of three layers from a ResNet-50 network pre-trained on ImageNet [14], and a single 3×3 dilated (rate = 2) convolutional layer with 256 filters. The ResNet-50 layers are frozen during training. The decoder comprises of an atrous spatial pyramid pooling layer (as introduced in [4]), and two convolutional layers. The last layer has linear activation and produces a 2-channel output $(1/8)^{th}$ the input size. The output from the last layer is resized to the expected mask size with bilinear interpolation. The predicted query and support masks are compared to their respective ground-truths using the cross-entropy loss.

3.3 Support and Query Mask Prediction (DPr)

The network is trained to predict both the support and the query masks using the same encoder and decoder. We submit that this dual prediction requirement forces the query and support features from the shared encoder to share greater and more target-specific similarity. For instance, if the support image (and mask) is of a car, the network must be able to recover back the support mask (i.e. the entirety of the car) from the foreground support features. Because the encoder is shared between the support and the query images, the query features share this characteristic with the support features. This is reflected in the effective gain in the mIoU as discussed in Section 5.1.

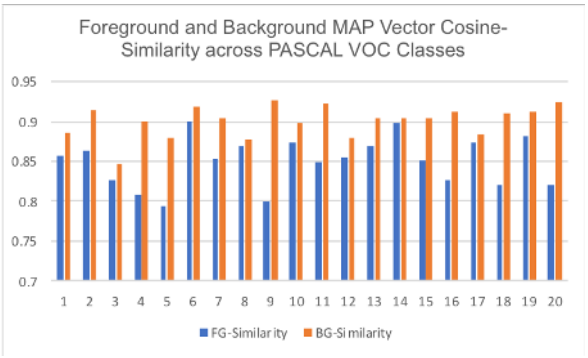

Figure 3: Foreground and Background cosine-similarity based on MAP vectors computed using ResNet-50 (layer2+layer3) features for 4000 image pairs from the PASCAL VOC dataset. Background similarity has a greater average magnitude than foreground similarity across all classes. This indicates the opportunity to obtain cues from similar background regions.

3.4 Foreground/Background Attentive Fusion (FBAF)

Late fusion methods (like [12, 24, 25, 26]) fuse the foreground features from the support branch into the query branch to locate the regions of class-conditional similarity. The foreground features are obtained using the masked average pooling (MAP) operation that computes a channel-wise weighted average of the support features where the weights are the support mask values at each position. We submit that fusing the background feature information from the support branch has the effect of suppressing similar background features in the query features. The fusion process can be concisely stated using the following three equations, each representing a step of the process:

1. Dual Probe Feature Extraction: The foreground and the background MAP vectors are computed using the support mask and the support features.

$$\begin{aligned} Z^f &= \mu_c(F_s * \breve{M}_s) \\ Z^b &= \mu_c(F_s * (1 - \breve{M}_s)) \end{aligned} \tag{1}$$

   Here, μ_c is the average pooling operation with a kernel size equal to the feature map size, F_s are the support features, and $\breve{M}_s$ is the support mask down-sampled to the height and width of F_s.

2. FG/BG Attention Map Generation: Four attention maps, two (foreground and background) each from the support and query features are computed.

$$\begin{aligned} C(F, Z) &= (1 + cossim(F, Z))/2 \\ N(A, B) &= (A/(A+B),\ \ B/(A+B)) \\ A_s^f,\ \ A_s^b &= N(C(F_s, Z^f),\ \ C(F_s, Z^b)) \\ A_q^f,\ \ A_q^b &= N(C(F_q, Z^f),\ \ C(F_q, Z^b)) \end{aligned} \tag{2}$$

Method	1-shot					5-shot				
	split-1	split-2	split-3	split-4	mean	split-1	split-2	split-3	split-4	mean
SimPropNet(ours)	**54.86**	**67.33**	**54.52**	**52.02**	**57.19**	**57.20**	**68.50**	**58.40**	**56.05**	**60.04**
CA-Net [26]	50.41	63.02	46.09	47.46	51.75	52.27	65.29	47.15	50.50	53.81
PA-Net [24]	42.40	58.00	51.10	41.20	48.13	51.80	64.60	59.80	46.50	55.70
SG-One [25]	40.20	58.40	48.40	38.40	46.30	41.90	58.60	48.60	39.40	47.10
AMP [16]	41.90	50.20	46.70	34.70	43.40	41.80	55.50	50.30	39.90	46.90
co-FCN [12]	36.70	50.60	44.90	32.40	41.10	37.50	50.00	44.10	33.90	41.40
OSLSM [15]	33.60	55.30	40.90	33.50	40.80	35.90	58.10	42.70	39.10	43.90

Table 3: Percentage (%) mean-IoU of one-shot segmentation on the PASCAL 5^i dataset using random partitions. The best results are highlighted in bold. SimProp-Net achieves state-of-the-art performance on both the 1-shot and 5-shot settings.

Here $cossim$ is the cosine-similarity measure and F_q are the query features.

3. Feature and Attention Map Fusion: The attention maps and the features for the query and support are fused to two feature vectors that are finally decoded into the query and support mask predictions respectively.

$$\begin{aligned} G^0_* &= F_* \oplus Z^f \\ G^1_* &= IN(Conv_{3\times3}(G^0_*) + G^0_*) \\ G^2_* &= IN(Conv_{3\times3}(G^1_* \oplus A^f_* \oplus A^b_*) + G^1_*) \\ G_* &= IN(Conv_{3\times3}(G^2_*) + G^2_*) \end{aligned} \tag{3}$$

Here G_* is for both support (G_s) and query (G_q) features, and $\oplus$ is the concatenation operation and IN is the Instance-Normalization operation [22].

Because the fusion process combines information from both the foreground and the background support features, we term this *FG/BG Attentive fusion* (FBAF). The analysis in Section 5.1 demonstrates the effective increase in prediction accuracy by employing FBAF.

3.5 k-Shot Inference

The network is not specifically trained for k-shot training ($k > 1$). To incorporate more than one support example pairs in inference, the MAP vectors computed in the Dual Probe Feature Extraction step are averaged over the support pairs:

$$Z^f_{kshot} = \frac{\sum_k Z^f_k}{k}, \quad Z^b_{kshot} = \frac{\sum_k Z^b_k}{k} \tag{4}$$

These MAP vectors are used to compute the foreground and background attention maps, and are fused with the query features (Section 3.4) to predict the query segmentation mask.

3.6 Implementation Details

The training is done on virtual machines on Amazon AWS with four Tesla V100 GPUs and 16-core Xeon E5 processors. The standard SGD optimization algorithm is used, learning rate is kept at (2.5×10^{-3}) and the batch size is 8 for all training. Training with still higher learning rate yields even better results than reported in the paper, but the training is not always stable and may decay considerably in later training steps. Training for each split is run for 180 epochs and the checkpoint with the highest validation score (mIoU) is retained. To prevent the network from overfitting on the training classes, we also use an input regularization technique called **Input Channel Averging** (**ICA**) where the query RGB image is mapped to a grayscale input (after normalizing) with a *switch probability* (initialized at 0.25 for our experiments) that decays exponentially as training progresses. The particular benefit of using ICA is discussed in Section 5.1.

Figure 4: Qualitative One-shot Segmentation Results from SimPropNet. Notice that providing a more similar support image helps to improve the segmentation (Cow's horn in the top row, and marbles in the bottom row).

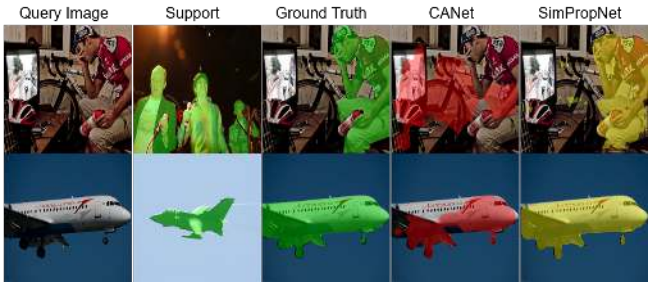

Figure 5: One-shot segmentation results compared to CANet[26] output. Row 1 indicates improved class-specificity - the mask is localized on the correct target class. Row 2 demonstrates better utilization of similar background context.

4 Results

In this section, we present the evaluation metrics and reporting criteria for few-shot segmentation, the quantitative results for SimPropNet, and comparison with other state-of-the-art methods. For brevity, we include very few qualitative results highlighting the key benefits our work.

4.1 Metric and Reporting

For consistency of comparison with baseline techniques, we follow the evaluation protocol established by Shaban *et al.* [15]. Different instances of the network are trained with the different training splits of the PASCAL-5^i dataset, each with 15 classes of objects and their masks. Testing is done on the images with objects of the 5 withheld classes. The mean intersection-over-union (mIoU) metric for output masks with respect to the ground-truth mask is computed. We report the mean IoU values for random support and query pairs (images of the same class) in the standard 1-shot and 5-shot ($k = 5$ support images) settings. To ensure fairness in comparison, we use author-provided code for baselines and evaluate performance for three sets of random support-query test pairs for each of the splits and report the average performance. These splits will be released along with the paper upon acceptance.

4.2 SimPropNet Results

Quantitative Results We compare our performance using the method described in [15] that employs random query and support image pairs (partitions) from each test class in PASCAL-5^i splits. The mean intersection-over-union (mIoU) values are reported in Table 3. Our method, SimPropNet, achieves state-of-the-art performance in both the 1-shot and the 5-shot setting. SimPropNet outperforms CANet [26], the current state-of-the-art in 1-shot setting, by 5.44% in 1-shot and 6.23% in the 5-shot evaluation. Further, SimPropNet outperforms PANet [24], the current state-of-the-art in 5-shot setting, by 9.06% in the 1-shot evaluation and 4.34% in the 5-shot evaluation task.

Qualitative Results Qualitative one-shot segmentation results also indicate significant improvements in output. Figure 4 highlights how providing a more similar support image and mask helps in improving the segmentation, and is expected in a practical scenario. Figure 5 presents two sample results for comparison with CANet [26]. The first row illustrates how CANet may overfit on training classes (target test class (person) is from split-2, and bicycle is in the training set) or its features lack specificity to the target class, and how SimPropNet overcomes this issue. Row 2 of the figure illustrates the benefit of providing support images with similar backgrounds.

5 Analysis

In this section, we first analyse the particular benefit of each component contribution of SimPropNet individually. Next, we probe the effectiveness of SimPropNet in improving similarity propagation. We present these evaluations in the one-shot setting and compare against CANet [26], the existing state-of-the-art.

5.1 Ablation Study of Components

We study the effectiveness of each of our contributions individually - dual prediction (DPr) of the support and query masks, and foreground-background attentive fusion (FBAF). Table 4 reports the mIoU values over the different splits of the PASCAL 5^i dataset. As highlighted by the results, both DPr and FBAF used individually demonstrate clear gains over the baseline (CANet [26]) of 4.14% and 3.75% in mIoU respectively. FBAF alone performs better than DPr in three of the four splits, but has a slightly worse mean because of its sharp decline in performance in split-3. The combination of DPr and FBAF achieves an improvement of 5.34% over [26]. Additionally using ICA during training further improves mIoU on three of the splits and increases mean mIoU to 57.19%.

Method	split-1	split-2	split-3	split-4	mean
Baseline (CANet)	50.41	63.02	46.09	47.46	51.75
DPr	52.69	66.57	53.1	51.23	55.89
FBAF	54.16	66.71	49.11	52.00	55.50
DPr + FBAF	54.08	67.29	54.05	**52.93**	57.09
SimPropNet	**54.86**	**67.33**	**54.52**	52.02	**57.19**

Table 4: Ablations for the different components of SimPropNet (DPr + FBAF + ICA). DPr is the joint prediction of query and support masks, FBAF is the FG/BG Attentive Fusion module, and ICA is the input channel averaging regularization.

5.2 Measuring Gain in Similarity Propagation

We evaluate the performance of the proposed network on identical support and query images (as reported for [26] and [16] in Table 2 in Section 3.1). The performance of a one-shot segmentation method on identical inputs for the query and the support is the upper bound on the degree of similarity propagation in the network. Results of this experiment for SimPropNet (Table 5) show an average gain of 21.5% over CANet [26], and 14.23% over AMP [16], over all splits. This indicates clearly that the network is utilizing the similarity information between the query and support images better.

Method	split-1	split-2	split-3	split-4	mean
SimPropNet	71.24	82.09	74.97	77.15	76.36
CANet	54.51	63.98	48.2	52.76	54.86
ΔCANet	**16.73**	**18.11**	**26.77**	**24.39**	**21.50**
AMP	54.41	69.34	64.79	60.02	62.14
ΔAMP	**16.83**	**12.75**	**10.18**	**17.13**	**14.23**

Table 5: Percentage (%) mean-IoU measured for FSS methods as compared with the proposed SimPropNet for test images from PASCAL 5^i dataset when query image (I_q) = support image (I_s)

6 Conclusions and Future Work

The paper presents a rigorous argument that similarity propagation in existing few-shot image segmentation networks is sub-optimal. It proposes SimPropNet, a deep neural network with two strategies for bridging this gap - predicting the support mask besides the query mask with a shared decoder, and fusing background features into the feature fusion module. The network achieves a new state-of-the-art as shown by a comprehensive set of experiments. Class-conditional similarity matching can only match pixels with a similar class-mix between the query and the support images. In future work, we focus on exploring the "objectness" aspect of the target image to improve few-shot segmentation.

References

[1] Yağız Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37(4):72:1–72:13, 2018.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495, 2017.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[4] 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*, 40:834–848, 2016.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38:1116–1126, 2018.

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

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

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

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

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[15] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 167.1–167.13. BMVA Press, September 2017.

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

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[18] Ashish Sinha and José Estevan Dolz. Multi-scale guided attention for medical image segmentation. *ArXiv*, abs/1906.02849, 2019.

[19] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. 03 2017.

[20] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. Soft color segmentation and its applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(9), September 2007.

[21] Jianchao Tan, Jyh-Ming Lien, and Yotam Gingold. Decomposing images into layers via rgb-space geometry. *ACM Trans. Graph.*, 36(1), November 2016.

[22] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.

[23] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, D. Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *ArXiv*, abs/1908.07919, 2019.

[24] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. *ArXiv*, abs/1908.06391, 2019.

[25] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. 10 2018.

[26] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2016.