

On the Texture Bias for Few-Shot CNN Segmentation

Reza Azad¹, Abdur R Fayjie^{1,2}, Claude Kauffman², Ismail Ben Ayed¹, Marco Pedersoli¹, and Jose Dolz¹

¹ ETS Montreal, Montreal, Canada

² CRCHUM Montreal, Montreal, Canada

reza.azad.1@ens.etsmtl.ca

Abstract. Despite the initial belief that Convolutional Neural Networks (CNNs) are driven by shapes to perform visual recognition tasks, recent evidence suggests that texture bias in CNNs provides higher performing and more robust models. This contrasts with the perceptual bias in the human visual cortex, which has a stronger preference towards shape components. Perceptual differences may explain why CNNs achieve human-level performance when large labeled datasets are available, but their performance significantly degrades in low-labeled data scenarios, such as few-shot semantic segmentation. To remove the texture bias in the context of few-shot learning, we propose a novel architecture that integrates a set of Difference of Gaussians (DoG) to attenuate high-frequency local components in the feature space. This produces a set of modified feature maps, whose high-frequency components are diminished at different standard deviation values of the Gaussian distribution in the spatial domain. As this results in multiple feature maps for a single image, we employ a bi-directional convolutional long-short-term-memory to efficiently merge the multi scale-space representations. We perform extensive experiments on two well-known few-shot segmentation benchmarks –Pascal i5 and FSS-1000– and demonstrate that our method outperforms significantly state-of-the-art approaches.

1 Introduction

Deep models, and particularly convolutional neural networks (CNNs), have shown an impressive performance in many visual recognition tasks, including semantic segmentation [1]. However, their extreme hunger for labeled training data strongly limits their scalability to new classes and reduces their applicability to rare categories. Few-shot learning [2,3] has appeared as an appealing alternative to train deep models in a low-labeled data scenario. In this setting, the model is trained to accommodate for novel categories with only a handful of labeled images, typically known as *support* images. In few-shot segmentation approaches, the learned knowledge from the *support* images is typically fed into a parametric module to guide the segmentation of the unseen images, referred to as *queries*.

Recent works have demonstrated that the CNN bias towards recognizing textures rather than shapes introduces several benefits under the standard learning

paradigm [4,5], which contrasts with the inductive bias found in the human visual cortex, that is driven by shapes [6]. This does not represent a problem when training and testing classes are drawn from the same distribution in large-labeled datasets. Nevertheless, in low-labeled data regime, the difference on perceptual biases poses difficulties to CNNs to mimic human performance, particularly if there exists a distributional shift between training and testing classes, such as in the few-shot learning scenario [7].

Thus, we argue that attenuating high-frequency local components in the feature space yields to a better generalization under distributional shift in the context of few-shot semantic segmentation. To achieve this, we propose a novel architecture (Figure 1) for few-shot semantic segmentation. Particularly, the proposed model integrates a set of Difference of Gaussians (DOGs) on the feature representations. At each scale-space of the DOGs, the original high-frequency signals are attenuated differently, according to the standard deviation values, σ , employed to model the Gaussian distribution in the spatial domain, which results in multiple versions of the feature maps for a single image. Then, following the standard literature on few-shot segmentation, we generate class representative prototypes from the learned representations, with the difference that in our setting we have multiple prototypes per image, i.e., one at each scale-space of the DOG. Thus, for each query image, our model produces an ensemble of segmentations, each one associated with a prototype. To generate the final prediction, we cast the problem into a sequential segmentation problem, where each segmentation on the ensemble represents a time-point. Then, to efficiently fuse temporal, i.e., multiple segmentation masks, and spatial features we resort to a Bi-directional convolutional long-short-term memory (Bi-ConvLSTM) [8], which bidirectionally encourages information exchange between LSTM units. Furthermore, for K-shot setting, the proposed approach learns a parametric fusion of the different support images, by jointly analyzing their contribution.

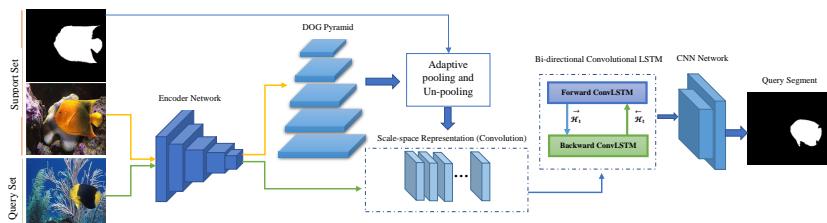


Fig. 1. Overview of the proposed method (DoG-LSTM) for few-shot segmentation. It first applies a pyramid of Difference of Gaussians (DoG) on the learned *support* features to attenuate high-frequency local components on the feature space. To perform segmentation on a *query* image, the multiple scale-space *support* representations are combined with the *query* features, and later fed as input to a bi-directional convLSTM. The convLSTM merges the information from multiple representations and generates the final *query* segmentation map.

Our contributions can be summarized as follows: (1) we propose to reduce the texture bias in CNNs in the task of few-shot segmentation by attenuating high-frequency local components on the feature space, (2) to merge the multiple segmentations produced at different scale-space representations we reformulate the problem as a sequential segmentation task and employ a bi-directional convLSTM to efficiently fuse all the information, and (3) we report new state-of-the-art results on few-shot segmentation across several public benchmarks. Furthermore, we provide bounding box annotations for the FSS-1000 dataset [9], with the goal of fostering future research on weakly supervised few-shot segmentation³.

2 Related Work

Few-shot segmentation. Pioneer works on few-shot semantic segmentation [10,11,12] incorporated two independent branches: a conditional branch that generates the prototypes (e.g., embedding) from the *support* set, and a segmentation branch, which takes the learned prototypes and the *query* image as input and produces the segmentation masks. More recently, researchers have unified these dual-branch architectures into a single branch network which can derive better guidance features with the addition of a masked average pooling layer [13,14,15,16]. Zhang et al. [13] integrate a similarity guidance module that recalibrates the query feature map based on a similarity score between the representative prototype and each spatial location on the query features. In [14], authors present an approach to generate the weights of the final segmentation layer for the novel classes via imprinting. Wang et al. [15] propose a novel prototype alignment regularization between support and query images which provides better generalization. Particularly, support-to-query and query-to-support few-shot segmentation is performed in order to align the prototypes between both sets. Similarly, Nguyen et al. [16] integrated a regularization that estimate feature relevance by encouraging jointly high-feature activations on the foreground and low-feature activations on the background. In other recent works [17,18], deep attention has been exploited to learn attention weights between support and query images for further label propagation.

LSTM-based semantic segmentation. Long Short Term Memory (LSTM) [19] has been widely studied as a particular recurrent neural network structure to model long-range dependencies. Tasks such as semantic segmentation are typically addressed using models based on convolutions, where the learned intermediate representations of the input images preserve the spatial information, important for precise object segmentation. Furthermore, compared to models based on fully connected layers, these require a lower amount of learnable parameters. Thus, since LSTM models input-to-state and state-to-state with full connections this type of structure is not suitable to tackle spatio-temporal data. To address this limitation, convLSTM was presented in [20], where a convolution operator is integrated in the state-to-state and input-to-state transitions.

³ <https://github.com/rezazad68/fewshot-segmentation>

In the context of image segmentation, several works have exploited this idea. For example, in the case of 3-dimensional data, e.g. videos or medical imaging, convLSTMs are integrated to encode the spatial-temporal relationships between frames or slices [21,22,23,24]. If only 2D images are available instead, an alternative is to leverage convLSTM for multi-level feature fusion [25,26]. Li et al. [25] employed convLSTM units to progressively refine the segmentation masks from high-level to low-level features. In [26], features derived from the skip connections in the encoding path of UNet [27] were non-linearly fused with their corresponding features in the decoding path by employing a bi-directional convLSTM, instead of a simple concatenation. In a concurrent work, Hu et al. [17] employ a ConvLSTM to merge multiple segmentations in a k -shot scenario ($k > 1$), where each segmentation is generated from a different support image. This is different from our work, where our goal is to fuse the segmentations from a single support image ($k = 1$) derived from the multiple scale-space representations. Furthermore, we use a bidirectional ConvLSTM to foster the exchange of information between the forward and backward path of each recurrent module.

3 Methodology

3.1 Problem Formulation

Following the standard notation and set-up in few-shot semantic segmentation, we define three datasets: a training set $D_{train} = \{(X_i^t, Y_i^t)\}_{i=1}^{N_{train}}$, a support set $D_{support} = \{(X_i^s, Y_i^s)\}_{i=1}^{N_{support}}$, and a test set $D_{test} = \{(X_i^q)\}_{i=1}^{N_{test}}$. In this setting, $X_i \in \mathbb{R}^{H \times W \times 3}$ denotes an RGB image, with H and W being the height and the width of the image, respectively, and $Y_i \in \{0, 1\}^{H \times W}$ is its corresponding pixel-level mask. Furthermore, each set contains N images. The classes, denoted as $c \in C$, are shared among the support and test set, and are disjoint with the training set, i.e., $\{C_{train}\} \cap \{C_{support}\} = \emptyset$.

The purpose of few-shot learning is to train a neural network $f_\theta(\cdot)$ on the training set D_{train} to have the ability to segment a novel class $c \notin C_{train}$ on the test set D_{test} based on k references from the support set $D_{support}$. To reproduce this mechanism during the training process, the network is trained on D_{train} following the episodic paradigm [28]. Specifically, assuming a c -way k -shot learning task, each episode is generated by sampling: (1) a support training set $D_{train}^S = \{(X_s^t, Y_s^t(c))\}_{s=1}^k \subset D_{train}$ for each class c , where $Y_s^t(c)$ is the binary mask for the class c corresponding to the image X_s^t and (2) a query set $D_{train}^Q = \{X_q^t, Y_q^t(c)\} \subset D_{train}$, where X_q^t is the query image and $Y_q^t(c)$ its corresponding binary mask for the class c . The input of the model is composed of the support training set and the query image, $f_\theta(D_{train}^S, X_q^t)$, which are employed to estimate the segmentation mask for the class c in the query image, $\hat{Y}_q^t(c)$. Then, the parameters of the neural network θ are optimized by employing an objective

function between $Y_q^t(c)$ and $\hat{Y}_q^t(c)$ ⁴. During the testing phase, the model $f_\theta(\cdot)$ is evaluated on the test set D_{test} given k images from the support set $D_{support}$.

3.2 Removing Texture Bias

Recent findings suggest that perceptual bias on CNNs do not correlate with those in the human visual cortex [4], which may limit the performance of these models in low-labeled data scenarios [7]. Inspired by this, we propose to reduce the texture bias of CNNs in the context of few-shot segmentation. To achieve this, we integrate a set of Difference of Gaussians (DoGs) into the learned feature space to attenuate high-frequency local components, i.e., texture. Specifically, we first use a CNN to encode the input images into the latent space, resulting in $F_s \in \mathbb{R}^{W' \times H' \times M}$ and $F_q \in \mathbb{R}^{W' \times H' \times M}$ for the support and query samples. The variables W' , H' and M represent the width, height and feature dimensionality on the latent space, respectively. To encode the high-frequency information during training, we apply a DoG on each channel $m \in M$ of the feature map from the support samples F_s , which can be formulated as:

$$G_s = \Gamma_{\sigma_1, \sigma_2}(F_s) = (F_s^m * \frac{1}{2\pi\sigma_2^2} \exp^{-\frac{x^2+y^2}{2\sigma_2^2}}) - (F_s^m * \frac{1}{2\pi\sigma_1^2} \exp^{-\frac{x^2+y^2}{2\sigma_1^2}}), \forall m \in M \quad (1)$$

where σ_1 and σ_2 are ($\sigma_2 > \sigma_1$) are the variance of the Gaussian filters, x and y represent the spatial position in the encoded feature space and $*$ denotes the convolution operator. To encode different frequency information we apply a pyramid of DoGs with increasing σ values, similar to [29]. This results in L level representations ($L = 4$) for each support sample (See Fig. 2), where the novel feature maps at each level ($l \in L$) can be denoted as $G_s^l \in \mathbb{R}^{W' \times H' \times M}$.

Support images can contain cluttered background, as well as multiple object categories. Thus, we need to find a representative embedding f_s that corresponds exclusively to the target class. Since we have L feature representations, each of them encoding different high-frequency local components, we generate L prototypes per class. To obtain the class prototypes, the novel encoded feature maps at each scale G_s^l are averaged over the known foreground regions in the support mask $Y_s(c)$. Thus, at each level we can estimate f_s^l as:

$$f_s^l = \frac{1}{|\tilde{Y}_s(c)|} \sum_{i=1}^{W'H'} G_s^l \tilde{Y}_s(c) \quad (2)$$

where the support mask $Y_s(c)$ is down-sampled to $\tilde{Y}_s(c) \in \{0, 1\}^{H' \times W'}$ to match the spatial resolution of the feature maps G_s^l and $|\tilde{Y}_s(c)| = \sum_i \tilde{Y}_{s,i}(c)$ is the number of foreground locations in $\tilde{Y}_s(c)$. Then, each prototype is unpooled to the same spatial resolution as the query features F_q and the upsampled prototypes

⁴ Typically the standard cross-entropy loss function is employed in the few-shot segmentation literature.

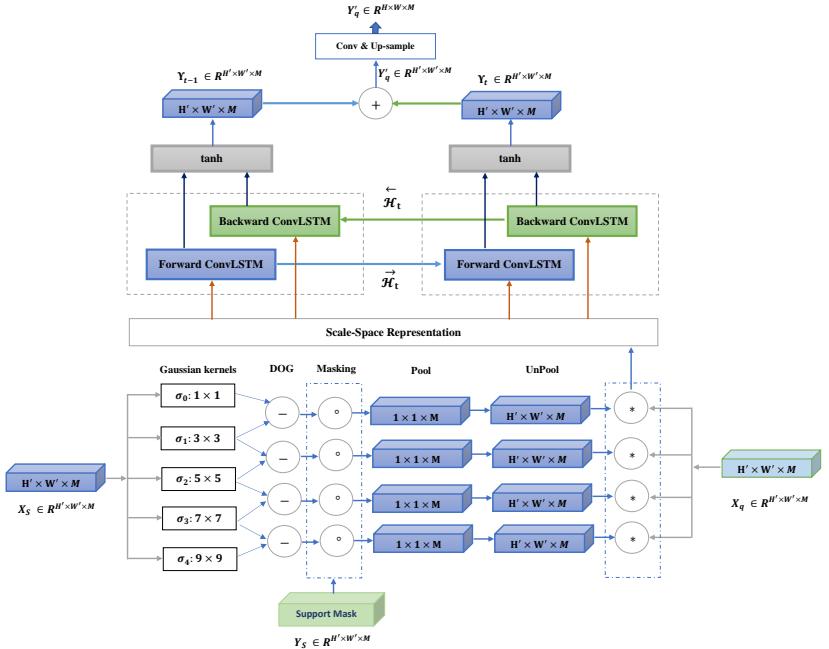


Fig. 2. The scale-space encoding block in the proposed method.

are convolved with F_q . We then define the scale-space representation (SSR), which will serve as input signal of the BConvLSTM. This representation can be formulated as a convolution operation between the class representative feature maps at each scale-space and the feature maps derived from the query image:

$$SSR = \{BN(\psi_s^l * F_q)\}, \forall l \in L \quad (3)$$

where ψ_s^l are the upsampled prototypes f_s^l , and BN denotes a batch normalization layer.

3.3 Encoding Scale-Space Representation

Fusion of the query features F_q with the multi-scale class representations from the support features ψ_s^l produces L joint feature maps, one at each scale-space representation. While logical or average operations may be a straightforward solution to obtain a unique representation, they fail to exploit the inner relationship between sequential scale-space representations. To efficiently solve this, we reformulate the problem as a sequential task, and integrate a bidirectional convolutional long short term memory (BConvLSTM) [23] on the output of the CNN architecture (Fig. 2). Even though LSTM have been proposed to deal with sequential problems, this sequential processing strategy may fail to explicitly encode the spatial correlation, since they use full connections in input-to-state and

state-to-state transitions. To overcome this limitation, ConvLSTM was proposed in [20], which leverages convolution operations into input-to-state and state-to-state transitions instead. Specifically, three gating functions are calculated in the ConvLSTM, which are defined as:

$$i_t = \sigma(\mathbf{W}_{xi} * \mathcal{X}_t + \mathbf{W}_{hi} * \mathbf{H}_{t-1} + \mathbf{W}_{ci} \circ \mathbf{C}_{t-1} + b_i) \quad (4)$$

$$f_t = \sigma(\mathbf{W}_{xf} * \mathcal{X}_t + \mathbf{W}_{hf} * \mathbf{H}_{t-1} + \mathbf{W}_{cf} \circ \mathbf{C}_{t-1} + b_f) \quad (5)$$

$$o_t = \sigma(\mathbf{W}_{xo} * \mathcal{X}_t + \mathbf{W}_{hf} * \mathbf{H}_{t-1} + \mathbf{W}_{co} \circ \mathbf{C}_{t-1} + b_0) \quad (6)$$

where \mathcal{X}_t and \mathbf{H}_t denote the input (i.e., *SSR* in eq. (3)) and hidden state at time t , respectively, and b is used to represent the bias term in each state. Similarly, \mathbf{W}_x , \mathbf{W}_h and \mathbf{W}_c represent the set of learnable parameters. Last, 'o' denotes the Hadamard product. The LSTM module generates a new proposal for the cell state by looking at the previous \mathbf{H} and current \mathcal{X} , resulting in:

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_{xc} * \mathcal{X}_t + \mathbf{W}_{hc} * \mathbf{H}_{t-1} + b_c) \quad (7)$$

Now we linearly combine the newly generated proposal $\tilde{\mathbf{C}}_t$ with the previous state \mathbf{C}_{t-1} to generate the final cell state in the recurrent model:

$$\mathbf{C}_t = f_t \circ \mathbf{C}_{t-1} + i_t \circ \tilde{\mathbf{C}}_t \quad (8)$$

Finally, the new hidden state \mathbf{H} can be estimated as:

$$\mathbf{H}_t = o_t \circ \tanh(\mathbf{C}_t) \quad (9)$$

Inspired by [23], we employ in this work a bidirectional convLSTM (BConvLSTM) [23] to encode the different scale-space representations (SSR) at the output of the convolutional network (Fig. 2). The bidirectional modules with forward and backward paths allow to strength the spatio-temporal information exchanges between the two sides, facilitating the memorization of both past and future sequences. This contrasts with the standard convLSTM, where only the dependencies on the forward direction are employed for the predictions. Thus, the output prediction for a query image X^q is given at the output of the BConvLSTM, which is defined as:

$$\hat{Y}^q = \tanh(\mathbf{W}_y^{\vec{\mathbf{H}}} * \vec{\mathbf{H}} + \mathbf{W}_y^{\overleftarrow{\mathbf{H}}} * \overleftarrow{\mathbf{H}} + b) \quad (10)$$

where $\vec{\mathbf{H}}$ and $\overleftarrow{\mathbf{H}}$ represent the hidden states of the forward and backward convLSTM units, respectively, and b is the bias term. Last, the output of the BConvLSTM is passed through a series of convolutions, followed by upsampling and batch normalization layers to produce the final segmentation masks in the original input image resolution.

3.4 k-shot Segmentation

To fuse information from several support images in the k-shot scenario ($k > 1$), most previous works estimate the class prototype ψ by simply taking the average of the representation vectors among k samples (non-parametric approach) [10,13,14]. However, this strategy assumes that each k sample has equal importance, and thus fails to provide a robust category representation when dealing with noisy or corrupted samples. To deal with this limitation, we propose to use a non-linear parametric method to further improve the model performance on the k -shot setting. The key idea is to generate the representation between the query and each k support samples and then to apply BCConvLSTM on these representations to get the final representation in a non-linear parametric fashion. Moving k -shot setting inside the scale-space representation gives the BCConvLSTM more freedom to generate better representations using various samples.

3.5 Weakly-supervised Few-shot Segmentation

Few-shot learning has aroused as an appealing strategy to alleviate the need of large labeled datasets to train deep neural networks. To push this idea further, we explore the performance of the proposed method when other forms different than full-supervision, i.e., full pixel-level masks, are available. Particularly, bounding box annotations are investigated, which are less time-consuming to obtain than exhaustive segmentation masks. In this context, we relax the support mask by considering all the area inside the bounding box as the foreground. We show in the experiments (Section 4.3) that, compared to pixel-level annotations, our model achieves very competitive results by employing sparse support annotations.

4 Experiments

In this section, we first present the datasets employed to evaluate our method and the experimental setting in our experiments. We then report the results compared to state-of-the-art segmentation approaches in one-shot and five-shot scenarios, demonstrating the benefits of our method.

4.1 Datasets

We perform extensive evaluations on two few-shot semantic segmentation benchmark datasets, i.e., PASCAL-5ⁱ and FSS-1000, whose details are given below.

PASCAL-5ⁱ. PASCAL-5ⁱ [10] is the most popular few-shot segmentation benchmark, which inherits from the well-known PASCAL dataset [30]. The images in PASCAL-5ⁱ are split into 4 folds, each having 5 classes, with 3 folds used for training and 1 for evaluation. Following the standard procedure in [10,16], we employ 1000 support-query pairs randomly sampled in the test split for each class at test time. More details on PASCAL-5ⁱ are provided in [10].

FSS-1000. A limitation of PASCAL-5ⁱ is that it contains relatively few distinct tasks, i.e., 20 excluding background and unknown categories. FSS-1000 dataset [9] alleviates this issue by introducing a more realistic dataset for few-shot semantic segmentation, which emphasizes the number of object classes rather than the number of images. Indeed, FSS-1000 contains a total of 1000 classes, where only 10 images and their corresponding ground truth for each category are provided. Out of the 1000 classes, 240 are dedicated to the test task and the remaining for training. The FSS-1000 dataset [9] only provides pixel-level annotations. Thus, to investigate the effect of using weak annotations in this dataset we generated bounding box annotations. Each bounding box is obtained from one randomly chosen instance mask in each support image. The generated bounding box annotations are provided with the code employed in the experiments.

4.2 Experimental Set-up

Network details. We employ VGG [31] pre-trained on ImageNet as the backbone feature extractor. The motivation behind this choice is to be able to compare our work with most existing methods in the few-shot segmentation literature. The proposed model is trained end-to-end by using Adam [32]. The initial learning rate is set to 10^{-4} and reduced by 10^{-1} at every 10K iterations. The model is trained for 50K episodes with a batch size of 5.

Evaluation protocol. To evaluate the performance of the few-shot segmentation models, there exist small differences in the literature. In [11], authors ignore image classes and estimate the mean of foreground intersection-over-union (IoU) and background IoU over all the test images. In other works [10, 18], the per-class foreground IoU is measured. Then, the average IoU over all classes (mIoU) is employed to report the final results. As pointed out in [18], the mIoU is a better metric in the context of few-shot semantic segmentation for several reasons. First, if a given image contains objects which are very small, the model may completely fail to segment those objects. Nevertheless, the background IoU can still be very high, which misleads information about the real performance of the model. And second, the foreground IoU is more suitable for binary segmentation problems, such as object segmentation in videos or foreground vs background extraction, while our purpose is on semantic segmentation.

Implementation details. The work is carried out using one NVidia Titan X GPU. The code is written in Keras with tensorflow as backend and it is publicly available at: <https://github.com/rezazad68/fewshot-segmentation>.

4.3 Results

Comparison with state-of-the-art. The comparison of the proposed model with state-of-the-art methods in the FSS-1000 and PASCAL-5ⁱ datasets is reported in Tables 1 and 2, respectively. Results in Table 1 show that the proposed model outperforms the state-of-the-art methods in both 1-shot and 5-shot

settings. Particularly, in the 1-shot task, our method achieves a significant improvement of 5.5% over the second best performing model. In the case of 5-shot learning, we found that fusing the segmentations from the different supports in a non-parametric way brings nearly 1% of improvement with respect to the 1-shot setting. Nevertheless, combining the 5 support segmentations in a parametric fashion, i.e., with BConvLSTM, increases the mIoU by 2.5%.

Table 1. Results of 1-way 1-shot and 1-way 5-shot segmentation on the FSS-1000 data set employing the mean Intersection Over Union (mIoU) metric. Our methods are shadowed in light gray. Best results are highlighted in bold.

Method	mIoU
	1-shot
OSLSM [10]	70.29%
co-FCN [12]	71.94%
FSS-1000 [9]	73.47%
FOMAML [33]	75.19%
Baseline	74.19%
Baseline + DoG	78.71%
Baseline + DoG + BConvLSTM	80.83%
	5-shot
OSLSM [10]	73.02%
co-FCN [12]	74.27%
FSS-1000 [9]	80.12%
FOMAML+ regularization [33]	80.60%
FOMAML+ regularization+UHO [33]	82.19%
Baseline + DoG + BConvLSTM (non parametric fusion)	81.65%
Baseline + DoG + BConvLSTM (parametric fusion)	83.36%

We now report in Table 2 an extensive evaluation of all previous works on the most common benchmark in few-shot semantic segmentation, i.e., PASCAL-5ⁱ. To make a fair comparison under different feature extractor backbones, we split the table into two groups. The *top* group shows the approaches that rely on VGG-16 as backbone architecture, whereas the methods in the *bottom* resort to ResNet to extract features. From the reported values, we can observe that the proposed approach clearly outperforms all previously known methods, under the same backbone and in both 1- and 5-shot scenarios. Specifically, compared to the second best performing approach (i.e., [16]), our method achieves nearly 6% and 5% of improvement in 1- and 5-shot, respectively. Furthermore, our approach consistently achieves the best performance in all except one fold among the 1- and 5-shot scenarios. Another interesting observation is that even comparing to methods based on ResNet, our approach achieves very competitive performance, being only surpassed by the very recent work in [17] (+3.2% in 1-shot and +1.6% in 5-shot). Nevertheless, as observed in several recent works, such as

[16], a deeper network shows a clearer tendency to increase the results, which may explain this gap in performance between our method and the approach in [17]. It is important to note that some methods include additional techniques to improve the final segmentations. For example, [34] employ dense-CRF as a post-processing step and [18] integrate an additional module that iteratively refines the query segmentation results. These quantitative results demonstrate the strong learning and generalization capabilities of the proposed model, particularly in the extreme case of 1-shot.

We want to bring to the reader’s attention that the recent work in [35] is not included in the current evaluation. The motivation behind this is that authors employed a different evaluation protocol for their method, which is based on the background IoU metric, slightly different from the metric used in this work. Nevertheless, the results reported in [35] show that their method achieves comparable results to OSLSM [10] and co-FCN [12]. This suggests that this approach may potentially underperform compared to our model.

Table 2. Results of 1-way 1-shot and 1-way 5-shot segmentation on PASCAL-5ⁱ data set employing the mean Intersection-Over-Union (mIoU) metric. Best results for each backbone architecture are highlighted in bold. We employ ∇ to denote the difference between 1- and 5-shot settings.

Method	1-shot					5-shot					∇
	fold ¹	fold ²	fold ³	fold ⁴	Mean	fold ¹	fold ²	fold ³	fold ⁴	Mean	
Backbone (VGG 16)											
FSS-1000 [9]	–	–	–	–	–	37.4	60.9	46.6	42.2	56.8	–
OSLSM [10]	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9	3.1
co-FCN [12]	36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4	0.3
SG-One [13]	40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1	0.8
AMP [14]	41.9	50.2	46.7	34.7	43.4	41.8	55.5	50.3	39.9	46.9	3.5
PANet [15]	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7	7.6
Feat Weight [16]	47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1	3.2
Meta-Seg [36]	42.2	59.6	48.1	44.4	48.6	43.1	62.5	49.9	45.3	50.2	1.6
MDL [37]	39.7	58.3	46.7	36.3	45.3	40.6	58.5	47.7	36.6	45.9	0.6
OS _{Adv} [38]	46.9	59.2	49.3	43.4	49.7	47.2	58.8	48.8	47.4	50.6	0.9
Proposed	56.2	66.0	56.1	53.8	58.0	57.5	70.6	56.6	57.7	60.6	2.6
Backbone (ResNet)											
Feat Weight [16] ‡	51.3	64.5	56.7	52.2	56.2	54.9	67.4	62.2	55.3	59.9	3.7
AMCG [17] ‡	–	–	–	–	–	61.2	–	–	–	62.2	1.0
CANet [18] †	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1	1.7
LTM [34] †	52.8	69.6	53.2	52.3	57.0	57.9	69.9	56.9	57.5	60.6	3.6
PGNet [39] †	56.0	66.9	50.6	50.4	56.0	57.7	68.7	52.9	54.6	58.5	2.5

*Employed architectures: † ResNet50, ‡ ResNet101

Qualitative results. We depict the visual results of the proposed method in Fig. 3 and Fig. 4 for Pascal5ⁱ and FSS-1000 datasets, respectively. These figures show the support image-mask pair and the segmentation generated by our method for multiple query images, as well as their corresponding ground truths for several categories. Without any post-processing step, the proposed model provides satisfying segmentation results on unseen classes with only one annotated support image. It is noteworthy to highlight that the same support image can be employed to segment multiple query images presenting high appearance variability. For example, our model can successfully segment cats (first row of Fig. 3) when only fractions of the target are shown, such as the head (first column) or even a partial head (third column). Looking at other categories, e.g., bike or table in Fig. 3 or bat in Fig. 4, we observe that the proposed method can also handle objects viewed from a different perspective or presenting different shapes. This illustrates that our model has a strong ability to successfully generalize to unseen classes from only a handful of labeled examples.

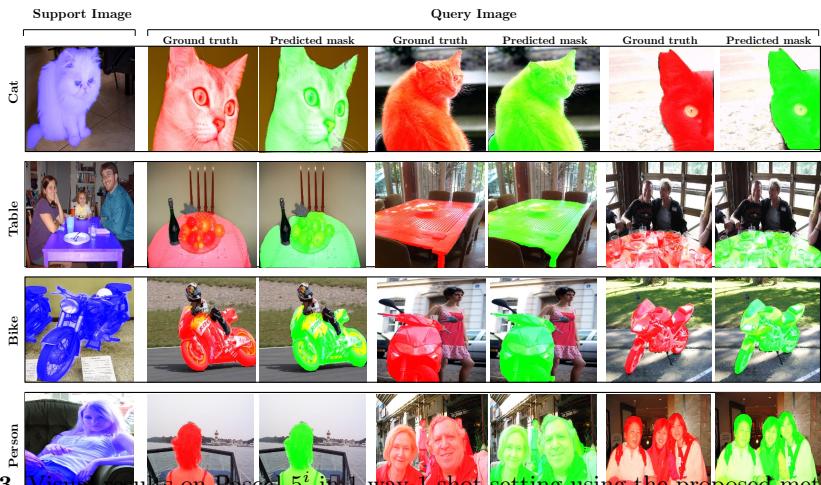


Fig. 3. Visual results on Pascal5ⁱ in 1-way 1-shot setting using the proposed method. The support set, as well as predictions on several query images with corresponding ground truths are shown.

Model complexity. The functionality of the proposed method in the demand of computational resources is also investigated in this work. Table 3 shows the model complexity of several methods, as well as their segmentation results on Pascal5ⁱ for 1-shot. In this table, we include the models that either report their number of parameters or provide reproducible code. We observe that the proposed method is ranked among the lightest methods, while achieving the second

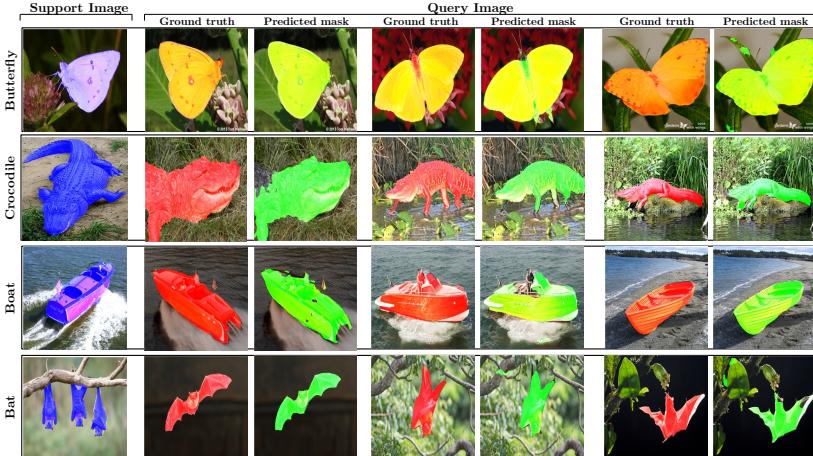


Fig. 4. Visual results on FSS-1000 class dataset in 1-way 1-shot setting using the proposed method. The support set, as well as predictions on several query images with corresponding ground truths are shown.

best segmentation performance. Compared to similar methods, in terms of complexity (e.g., co-FCN [12], SG-One [13], AMP [14] and PANet [15]), our model brings between 10 and 17% gain on improvement. On the other hand, methods achieving similar results (e.g., AMCG [17]) need $\times 4$ more parameters.

Table 3. Parameter complexity in different approaches and their performance (mIoU) on 1-shot segmentation on PASCAL-5ⁱ. Methods are ordered based on complexity.

Method	1-shot mIoU	#params(M)
OSLSM [10]	40.8	276.7
Meta-Seg [36]	48.6	268.5
AMCG with ConvLSTM [17]	61.2	90.8
AMCG [17]	57.3	89.5
AMP [14]	43.4	34.7
co-FCN [12]	41.1	34.2
Proposed Method	58.0	22.7
SG-One [13]	46.3	19.0
PANet [15]	48.1	14.7

Weakly supervised performance. We further evaluate the proposed model with weaker forms of annotations, e.g., bounding boxes. As reported in Table 4, our method achieves comparable performance to full supervision when bounding

boxes are available in the support set of novel categories. Furthermore, compared to the very recent PANet architecture [15] our model brings 10% of performance gain in the context of weak supervision. This suggests that our model is able to deal efficiently with noise introduced by bounding box annotations, which ultimately results in more representative class prototypes that approach those obtained by pixel-level annotations.

Table 4. Full supervision vs weak-supervision performance in the 1-shot setting.

Method	mIoU	
	FSS-1000	PASCAL
Proposed (Pixel annotations)	80.83%	58.0%
Proposed (Bounding box annotations)	78.23%	56.4%
PANet [15] (Bounding box annotations)	-	45.1%

Table 5. Effect of combining different level feature maps in the encoder network. Best result is highlighted in bold.

Block 3	Block 4	Block 5	mIoU
✓			76.33
	✓		78.31
		✓	79.47
✓	✓		78.05
✓		✓	80.57
	✓	✓	79.48
✓	✓	✓	80.83

Ablation study on multi-scale fusion features. Similarly to [18], we investigated the effect of employing different levels of features, or a combination of those. Particularly, we investigated the three last blocks of VGG-16. In our case, *block5* gives the best performance when a single block is used. If multiple blocks are used instead, we observed that combining the three blocks provides the best performance, even though the contribution of the *block4* is marginal compared to the fused features from *block3* and *block5* (+0.26%). The low performance of shallower layers alone can be explained by the fact that they exploit lower-level cues, which are insufficient to properly find object regions. By integrating these with higher-level features, which correspond to object categories, our model can efficiently identify class-agnostic regions on new images. Furthermore, fusion of features at several levels of abstraction can help to handle larger scale object

variations. Thus, the final multi-scale model employed in our experiments corresponds to the architecture combining the three last feature blocks.

5 Conclusions

We have presented a novel segmentation network that addresses the problem of texture bias on CNNs in a few-shot learning scenario. Particularly, the proposed model presents two novel contributions. First, we integrated a pyramid of Difference of Gaussians to attenuate high-frequency local components in the feature space. Second, to merge information at multiple scale-space representations we reformulated the problem as a sequential task and resorted to bi-directional convolutional LSTMs. For evaluation purposes, we have compared the proposed method to prior work, and performed ablations on important elements of our model on two public benchmarks: FSS-1000 and Pascal5ⁱ. Results demonstrated that the proposed model significantly outperforms the prior methods and achieves a new state-of-the-art performance on few-shot segmentation, while maintaining a lightweight structure.

References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
2. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR. (2017)
3. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org (2017) 1126–1135
4. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: ICLR. (2019)
5. Brendel, W., Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. In: ICLR. (2019)
6. Landau, B., Smith, L.B., Jones, S.S.: The importance of shape in early lexical learning. *Cognitive development* **3**(3) (1988) 299–321
7. Ringer, S., Williams, W., Ash, T., Francis, R., MacLeod, D.: Texture bias of CNNs limits few-shot classification performance. arXiv preprint arXiv:1910.08519 (2019)
8. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper ConvLSTM for video salient object detection. In: The European Conference on Computer Vision (ECCV). (2018)
9. Wei, T., Li, X., Chen, Y.P., Tai, Y.W., Tang, C.K.: FSS-1000: A 1000-class dataset for few-shot segmentation. arXiv preprint arXiv:1907.12347 (2019)
10. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. In: BMVC. (2018)
11. Dong, N., Xing, E.: Few-shot semantic segmentation with prototype learning. In: BMVC. Volume 3. (2018)
12. Rakelly, K., Shelhamer, E., Darrell, T., Efros, A., Levine, S.: Conditional networks for few-shot semantic segmentation. In: ICLR Workshop. (2018)
13. Zhang, X., Wei, Y., Yang, Y., Huang, T.: SG-one: Similarity guidance network for one-shot semantic segmentation. arXiv preprint arXiv:1810.09091 (2018)
14. Siam, M., Oreshkin, B.N., Jagersand, M.: AMP: Adaptive masked proxies for few-shot segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 5249–5258
15. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 9197–9206
16. Nguyen, K., Todorovic, S.: Feature weighting and boosting for few-shot segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 622–631
17. Hu, T., Yang, P., Zhang, C., Yu, G., Mu, Y., Snoek, C.G.: Attention-based multi-context guiding for few-shot semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8441–8448
18. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: 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. (2019) 5217–5226
19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8) (1997) 1735–1780

20. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. (2015) 802–810
21. Chen, J., Yang, L., Zhang, Y., Alber, M., Chen, D.Z. In: Advances in neural information processing systems. (2016) 3036–3044
22. Valipour, S., Siam, M., Jagersand, M., Ray, N.: Recurrent fully convolutional networks for video segmentation. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE (2017) 29–36
23. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convLSTM for video salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 715–731
24. Zhang, D., Icke, I., Dogdas, B., Parimal, S., Sampath, S., Forbes, J., Bagchi, A., Chin, C.L., Chen, A.: A multi-level convolutional LSTM model for the segmentation of left ventricle myocardium in infarcted porcine cine mr images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE (2018) 470–473
25. Li, R., Li, K., Kuo, Y.C., Shu, M., Qi, X., Shen, X., Jia, J.: Referring image segmentation via recurrent refinement networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5745–5753
26. Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S.: Bi-directional convlstm u-net with densley connected convolutions. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2019) 0–0
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
28. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in neural information processing systems. (2016) 3630–3638
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2) (2004) 91–110
30. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2) (2010) 303–338
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
33. Hendryx, S.M., Leach, A.B., Hein, P.D., Morrison, C.T.: Meta-learning initializations for image segmentation. arXiv preprint arXiv:1912.06290 (2019)
34. Yang, Y., Meng, F., Li, H., Wu, Q., Xu, X., Chen, S.: A new local transformation module for few-shot segmentation. In: International Conference on Multimedia Modeling, Springer (2020) 76–87
35. Tian, P., Wu, Z., Qi, L., Wang, L., Shi, Y., Gao, Y.: Differentiable meta-learning model for few-shot semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. (2020)
36. Cao, Z., Zhang, T., Diao, W., Zhang, Y., Lyu, X., Fu, K., Sun, X.: Meta-seg: A generalized meta-learning framework for multi-class few-shot semantic segmentation. IEEE Access **7** (2019) 166109–166121
37. Dong, Z., Zhang, R., Shao, X., Zhou, H.: Multi-scale discriminative location-aware network for few-shot semantic segmentation. In: 2019 IEEE 43rd Annual Computer

- Software and Applications Conference (COMPSAC). Volume 2., IEEE (2019) 42–47
- 38. Yang, G., Niu, D., Zhang, C., Zhao, X.: Recognizing novel patterns via adversarial learning for one-shot semantic segmentation. *Information Sciences* (2020)
 - 39. Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q., Yao, R.: Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision*. (2019) 9587–9595

Supplementary Material



Fig. 1. Additional visual results on the FSS-1000 class dataset in 1-way 1-shot setting using the proposed method. The support set, as well as predictions on several query images with corresponding ground truths are shown.

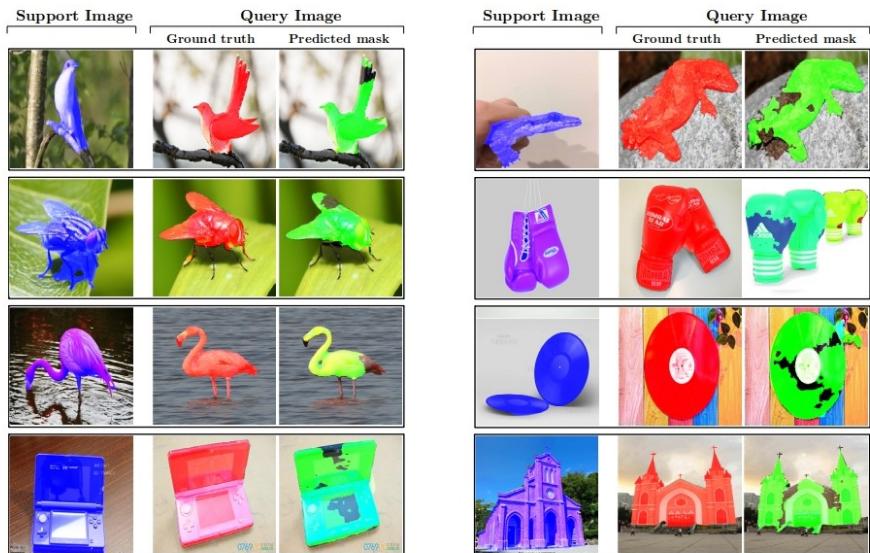


Fig. 2. Visual examples of *bad* segmentation results on the FSS-1000 class dataset in 1-way 1-shot setting using the proposed method. The support set, as well as predictions on several query images with corresponding ground truths are shown.



Fig. 3. Visual examples of segmentation results on the FSS-1000 class dataset in 1-way 1-shot setting using the proposed method *with bounding box annotations*. The support set (i.e., image and its corresponding bounding box annotation), as well as predictions on several query images with corresponding ground truths are shown.

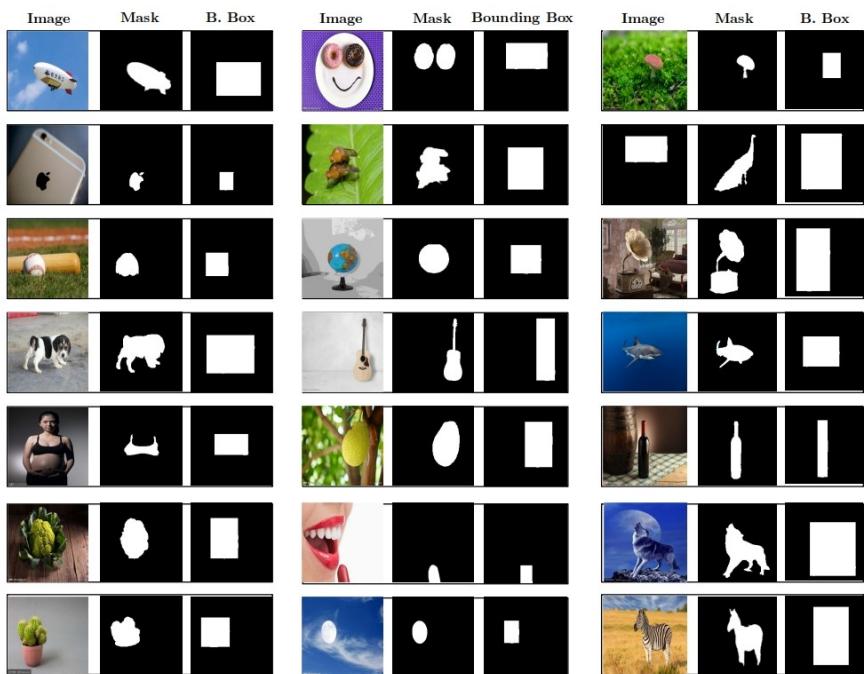


Fig. 4. Examples of *bounding box annotations* generated on the FSS-1000 class dataset.