BriNet: Towards Bridging the Intra-class and Inter-class Gaps in One-Shot Segmentation

Xianghui Yang[1]
xianghui.yang@sydney.edu.au
Bairun Wang[2]
wangbairun@sensetime.com
Kaige Chen[2]
chenkaige@sensetime.com
Xinchi Zhou[1]
xinchi.zhou1@sydney.edu.au
Shuai Yi[2]
yishuai@sensetime.com
Wanli Ouyang[1]
wanli.ouyang@sydney.edu.au
Luping Zhou[1]
luping.zhou@sydney.edu.au

[1] The University of Sydney
Sydney, AU
[2] SenseTime, Inc.
Beijing, CN

Abstract

Few-shot segmentation focuses on the generalization of models to segment unseen object instances with limited training samples. Although tremendous improvements have been achieved, existing methods are still constrained by two factors. (1) The information interaction between query and support images is not adequate, leaving intra-class gap. (2) The object categories at the training and inference stages have no overlap, leaving the inter-class gap. Thus, we propose a framework, BriNet, to bridge these gaps. First, more information interactions are encouraged between the extracted features of the query and support images, *i.e.*, using an Information Exchange Module to emphasize the common objects. Furthermore, to precisely localize the query objects, we design a multi-path fine-grained strategy which is able to make better use of the support feature representations. Second, a new online refinement strategy is proposed to help the trained model adapt to unseen classes, achieved by switching the roles of the query and the support images at the inference stage. The effectiveness of our framework is demonstrated by experimental results, which outperforms other competitive methods and leads to a new state-of-the-art on both PASCAL VOC and MSCOCO dataset. This project can be found at https://github.com/Wi-sc/BriNet.

1 Introduction

The past decade has witnessed the fast development of deep learning in computer vision [1, 3, 4, 5, 6, 10, 11, 14, 18, 24, 40, 41]. Semantic segmentation is one of the fundamental tasks in computer vision which aims at predicting the pixel-wise label of images. Despite the success brought by deep neural networks, the training of deep segmentation models still relies on large-scale datasets, such as ImageNet [25], PASCAL VOC [8] and MSCOCO [19]. In some cases, large-scale datasets are hard to attain due to the image collection and annotation costs. Moreover, the segmentation performance decreases significantly when the trained model is applied to unseen classes of objects. To solve this problem, few-shot segmentation was proposed by Shaban *et al.* [28].

Few-shot segmentation studies how to segment the target objects in a query image given a few (even only one) support images containing the objects of the same class with ground-truth segmentation masks. Typically, few-shot segmentation models take three items as input, a support image, its segmentation mask, and a query image, at both the training (offline) and the testing (online) stages. Please note that the categories at the online stage have no intersections with those at the offline stage.

Impressive performance has been achieved as in follow-up works [7, 21, 32, 36, 37, 38]. However, we observe the two limitations. **First**, the interaction of query and support has not been fully exploited to handle the intra-class gap that comes from the variations of objects within the same class. Current interaction is usually unidirectional and only utilized after feature extraction, *i.e.*, using support image information to influence the query image attention. Besides, the support-query relationship is measured via the similarity between the averaged support features of the masked regions and the local features of the query images. But the single coarse correlation is insufficient to precisely localize the objects in query images. **Second** and more importantly, most works directly apply the trained models to segmenting unseen categories at the test stage, without considering the inter-class gap between the training and the inference object categories.

To address the above two gaps, we propose a framework named BriNet, which differs from former works in the following aspects. **First**, to narrow the intra-class gap between the support and query images, we introduce an Information Exchange Module (IEM) that learns the non-local *bi-directional* transforms between support and query images, since they contain objects of the same category. The joint learning of feature representations make the deep model focus on the similar parts, *i.e.* the target objects to segment. Besides, rather than globally pooling the whole object region in a support image, we partition the whole object into sub-regions and conduct local pooling in each region to capture more details of the object and this process is conducted in the Multi-scale Correlation Module (MCM). **Second**, to effectively handle the category gap between the training and inference stages, we propose an online refinement strategy to make the network adaptive and robust to unseen object classes. The roles of query and support images are exchanged to offset the lack of labels for query images and then the network is refined by minimizing the segmentation errors of the support images with ground-truth labels. This strategy provides an additional supervision signal which effectively alleviates performance drop caused by the category gap. Our strategy is versatile and able to work well with other few-shot segmentation methods for further improvements. Our proposed framework outperforms other competitive methods and leads to a new state-of-the-art on both PASCAL VOC and MSCOCO dataset. Fig. 1 shows an overview of our framework for one-shot segmentation.

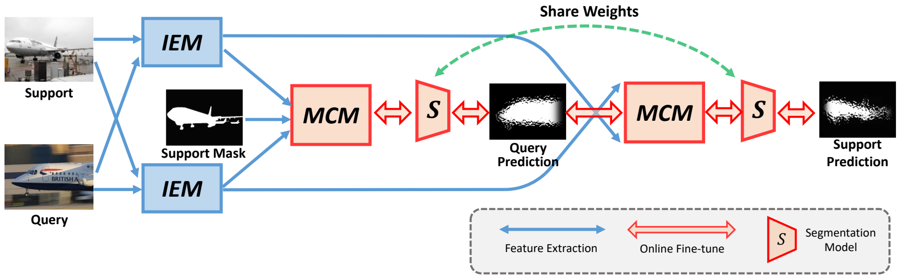

Figure 1: Overview of the proposed BriNet framework under one-shot segmentation scenario. The proposed model takes a query image and a support image with its segmentation mask as the input. The offline model consists of 2 novel modules, Information Exchange Module (IEM) and Multi-scale Correlation Module (MCM). At the online refinement stage, the roles of the query and the support images are switched, and the model is tuned to predict the segmentation mask of the support image that has the ground-truth, given the query image and the estimated query mask obtained from the initially trained model.

2 Related Work

Semantic segmentation. Semantic segmentation proceeds dense classification of each pixel in an image. Recent breakthroughs mainly benefit from deep CNNs [1, 3, 18, 24, 40]. The Dilated Convolution [4, 5], which is adopted in our work, enlarges the receptive field and boosts the segmentation performance. However, the training of deep-CNN-based segmentation models relies on large-scale datasets and once trained, the models cannot be deployed to unseen categories. Few-shot Semantic Segmentation is proposed to overcome the above issues.

Few-shot learning. Few-shot learning focuses on generalizing models to new classes with limited training samples. Few-shot classification, as a fundamental task, attracts lots of attention, including memory methods [20, 26], online refinement [9, 23], parameter prediction [2, 34], data augmentation with generative models [27, 35] and metric learning [17, 30, 31]. Our work is most related to online refinement. Inspired but different from former refinement strategy, we design a novel pseudo supervision subtly, which bridges the inter-class gap, specifically in the few-shot segmentation task.

Few-shot semantic segmentation. Few-shot semantic segmentation is firstly proposed by Shaban *et al.* [28]. A common paradigm employs a 2-branch architecture where the support branch generates the classification weights and the query branch produces the segmentation results, then followed by [7, 32, 36]. Among the following works, co-FCN [22] and SG-One [39] calculate the similarity of the query features and support features to establish the relationship between support and query images. Later on, CaNet [38] introduces an iterative refinement module to improve the prediction iteratively. Zhang *et al.* [37] model the support and query feature relation with local-local correlation, instead of the global-local one, by using attention graph network. Nguyen *et al.* [21] argue that there exists some unused information in test support images so that an online boosting mechanism is proposed, where support features are updated in the evaluation process. But they still ignore the information from the test query images. However, our framework utilize ignored query information to further bridge the gaps between both training and inference stages.

3 Task Description

Let $\mathcal{D}_{train} = \{(\mathbf{x}_*^{train}, \mathbf{m}_*^{train})\}$ be the training set and $\mathcal{D}_{test} = \{(\mathbf{x}_*^{test}, \mathbf{m}_*^{test})\}$ be the test set, where $\mathbf{x}_*$ and $\mathbf{m}_*$ denote the image set and the segmentation mask set, respectively, collected from either the query images (with the subscript q) or the support images (with the subscript s). Few-shot segmentation assumes that $\mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset$ and each pair of the query image $\mathbf{x}_q$ and its support images set $\{\mathbf{x}_s^i\}$ $(i = 1, \cdots, K)$ has at least one common object. Given the input triplets $(\mathbf{x}_q, \mathbf{x}_s^i, \mathbf{m}_s^i)$ sampled from $\mathcal{D}_{train}$, where $\mathbf{m}_s^i$ is the binary mask of $\mathbf{x}_s^i$, few-shot segmentation estimates the query mask $\hat{\mathbf{m}}_q$. For simplicity, in the following, we discuss our method under the scenario of $K = 1$ in Section 4.

4 Method

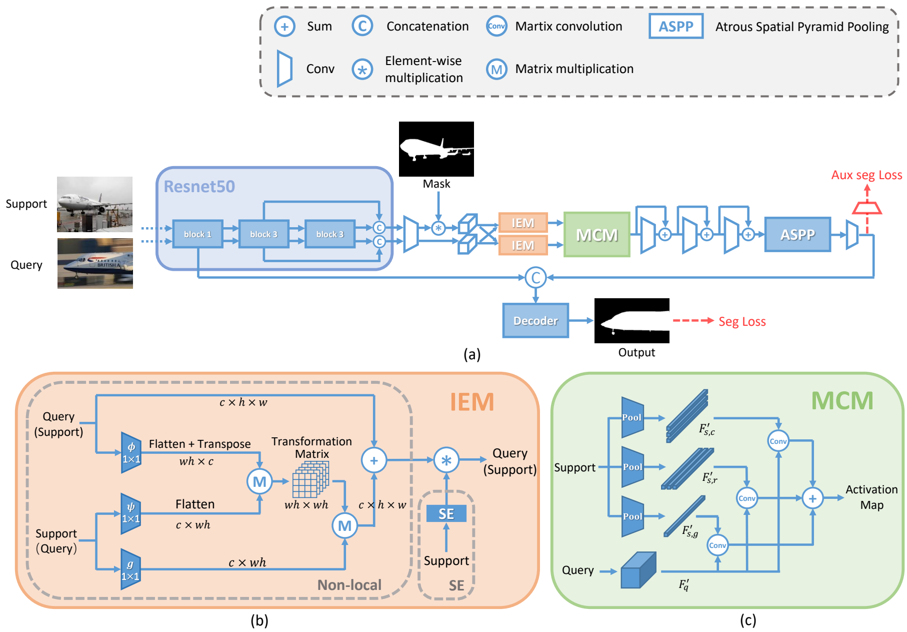

Figure 2: Network architecture of the proposed offline training model. (a) The overall architecture. (b) The detailed Information Exchange Module (IEM). (c) The detailed Multi-scale Correlation Module (MCM).

In this section, we present our proposed model, BriNet. It consists of an offline segmentation model and an online refinement algorithm. Specifically, for the **Offline** model, as shown in Fig. 2, an Information Exchange Module (IEM) and a Multi-scale Correlation Module (MCM) are proposed to enhance the similarity comparison and feature fusion, respectively. For the **Online** refinement, we propose a role-switching method to adapt the trained offline model to unseen categories, which will be illustrated in Sec. 4.2.

4.1 Offline Segmentation Model

Information Exchange Module. Given a pair of support and query images, their features are initially extracted by a common CNN model, such as ResNet. Following that, the information exchange modules are introduced to refine the features based on our belief that the common information contained in the support and the query images should be shared and co-weighted during feature extracting. The details of the IEM are given in Fig. 2 (b).

Specifically, before the initial feature maps F_s, $F_q \in \mathbb{R}^{c\times h\times w}$ extracted from $\mathbf{x}_s$, $\mathbf{x}_q \in \mathbb{R}^{3\times H\times W}$ by CNN could be embedded by an IEM, we apply the support mask $\mathbf{m}_s$ on the support feature map F_s to filter out unrelated background information. The resulting pure target object feature map can be written as $F_s^m = F_s \odot \mathbf{m_s}$, where $\odot$ denotes element-wise multiplication and $\mathbf{m}_s$ is down-sampled from $W\times H$ into the identical size $w\times h$ with F_s.

Our IEM takes F_s^m and F_q as inputs. It consists of Non-local block $\mathcal{T}$ [33] and Squeeze-and-Excitation (SE) block $\mathcal{C}$ [16], where Non-local block is used for information exchange and SE block aims to boosting channel information of target object. The Non-local block was also adopted by [15] for few-shot object detection but without applying foreground mask. IEM outputs the refined feature maps $F_q^{'}$ and $F_s^{'}$, formulated as Eq. 1,

$$F_q^{'} = \mathcal{T}(F_q, F_s^m) \odot \mathcal{C}(F_s^m), \quad F_s^{'} = \mathcal{T}(F_s^m, F_q) \odot \mathcal{C}(F_s^m). \tag{1}$$

Non-local block is originally proposed to capture long-range dependencies focusing on the regions of a single image input, where the non-local transformation is modeled by relations among local features. In contrast, in our approach, the Non-local block under support-query input setting is for inter-input transformation, denoted as Transformation Matrix in Fig. 2 (b), where all local features of one input (say, the support) are transformed into one local feature of the other input (say, the query). The support-query transformation is applied as Eq. 2 and Eq. 3,

$$\mathcal{T}(F_q, F_s^m)_i = \sum_j^{wh} \phi(F_q)_i{}^T \psi(F_s^m)_j \cdot g(F_s^m)_j + F_{q,i} \tag{2}$$

$$\mathcal{T}(F_s^m, F_q)_i = \sum_j^{wh} \phi(F_s^m)_i{}^T \psi(F_q)_j \cdot g(F_q)_j + F_{s,i}^m \tag{3}$$

where ϕ, ψ and g are 1×1 convolution kernels and i, j are the index of pixel. The SE block generates channel attention by Global Average/Max Pooling, followed by two sequential MLP layers.

Multi-scale Correlation Module. In contrast to previous coarse global mask pooling, our proposed method conduct region pooling in a more fine-grained manner, which achieves the balance between computation overhead and feature details. Fig. 2 (c) shows the details of MCM.

Specifically, given IEM ouputs $F_s^{'}, F_q^{'} \in \mathbb{R}^{c\times h\times w}$, apart from global average pooling, we apply 2 slide average pooling windows on feature map $F_s^{'} \in \mathbb{R}^{c\times w\times h}$ with size $s\times h, w\times s$ and strides s, s respectively, where s is the stride size. As a result, more fine-grained feature representations $F_{s,c}^{'} \in \mathbb{R}^{c\times \frac{w}{s}\times 1}, F_{s,r}^{'} \in \mathbb{R}^{c\times 1\times \frac{h}{s}}, F_{s,g}^{'} \in \mathbb{R}^{c\times 1\times 1}$ are obtained. After convoluting the query feature map $F_q^{'} \in \mathbb{R}^{c\times w\times h}$ with these 3 kernels respectively, the three activation maps are summed as Eq. 4,

$$F_{s-q} = F_{s,c}^{'} * F_q^{'} + F_{s,r}^{'} * F_q^{'} + F_{s,g}^{'} * F_q^{'} \tag{4}$$

where $*$ denotes convolution operation and F_{s-q} is the output of MCM.

Loss function. To boost performance, in addition to the final cross-entropy segmentation loss $\mathcal{L}_{seg}$, we introduce another auxiliary segmentation branch into our proposed architecture before the Decoder to shorten gradient propagation, as shown in Fig. 2. The auxiliary cross-entropy segmentation loss $\mathcal{L}_{aux-seg}$ is minimized together with the final segmentation loss $\mathcal{L}_{seg}$. Thus the overall loss function is

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{aux-seg}$$

4.2 Online Refinement

In order to make our model adaptive to the agnostic objects in the test stage, we propose to conduct online refinement. Our proposed online refinement algorithm takes advantage of the support-query pair available at the test stage and switches their roles to extract complementary information for the refinement iteratively.

According to the definition of few-shot segmentation, at test stage, the information about the agnostic categories in the support image x_s^{test}, as well as its corresponding mask m_s^{test}, and query image x_q^{test} are available, but the query mask m_q^{test} and other support-query pairs from $\mathcal{D}_{test}$ are unknown. Inspired by self-supervision, the core of our online refinement is to regard the query offline prediction $m_q^{\hat{t}est}$ as the pseudo (ground-truth) mask to assist the support image segmentation in return. This refinement could be conducted for a few rounds by switching the roles of the query and the support images iteratively.

Specifically, given a query image x_q^{test}, a support image x_s^{test} and a support ground-truth mask m_s^{test}, we do the online refinement in three steps. First, we feed x_q^{test}, x_s^{test} and m_s^{test} into the model obtained at the offline training to estimate the query mask $m_q^{\hat{t}est}$. Second, $m_q^{\hat{t}est}$ is treated as the pseudo ground-truth label for x_q^{test}, which constitutes the next input with x_q^{test}, x_s^{test} to predict the support mask $m_s^{\hat{t}est}$. Third, the segmentation model is constantly refined by minimizing the cross-entropy loss between the predicted support mask $m_s^{\hat{t}est}$ and the ground-truth support mask m_s^{test}. These 3 steps are repeated until the mean IoU between m_s^{test} and $m_s^{\hat{t}est}$ is higher than a threshold $t_i = t_0 * \frac{N-1}{N+i}$ in the i-th step or the maximum iteration time N has been reached. This algorithm is formulated as alternatively updating $\hat{m}_s$ and $\hat{m}_q$ according to Eq. 5,

$$\hat{m}_s = \mathcal{F}(x_s, x_q, \hat{m}_q), \quad \hat{m}_q = \mathcal{F}(x_q, x_s, m_s). \tag{5}$$

Here $\mathcal{F}$ indicates the embedding function of the model either trained offline or refined online in the last iteration.

5 Experiments

5.1 Datasets and Evaluation metric

We evaluate the performance of our model on two benchmark datasets commonly used for few-shot segmentation.

PASCAL-5$^{\mathbf{i}}$. This dataset was composed based on PASCAL VOC 2012 [8] and the extended SDS datasets [12]. Following the work in [28] and the conventional evaluation on this dataset, we adopt 4-fold cross validation that divides the 20 classes of PASCAL into four folds, three of which are used for training and the rest one for test. It is noted that the selection of support and query image pairs could influence the performance. Following [28], we randomly sample the support and query image pairs 1000 times from the test set for evaluation.

COCO-20$^{\mathbf{i}}$. Up to now, one-shot segmentation mainly takes PASCAL-5^i for evaluation. Only Zhang *et al*. [38] and Nguyen *et al*. [21] tested their methods on MSCOCO, and their dataset settings are even not the same. Zhang *et al*. [38] divide 80 classes into three parts, of which 40 classes are used for training, 20 classes for evaluation and the remaining 20 classes for test. Nguyen *et al*. [21] processed the dataset as the PASCAL-5^i, where the 80 classes are divided into 4 folds and each fold contains 20 classes, named COCO-20^i. We follow [21] to evaluate our model on COCO-20^i.

Evaluation metrics. To be consistent with the literature for comparison, class related foreground Intersection-over-Union (F-IoU) is adopted in this paper, which is computed as follows. First, the foreground intersection and union pixel numbers are summed according to classes; Second, the foreground Intersection-over-Union ratio is computed for each class; Third, the average IoU over all classes (mean IoU) are reported as the evaluation metric to reveal the overall performance.

5.2 Implementation details

We adopt ResNet50 [13] modified from Deeplab V3 [4] as our model backbone. In view of the task characteristic, we abandon ResBlock-4 and the later layers of ResNet50, which is consistent with the existing works in [37, 38]. The parameters of ResNet are initialized from the model pre-trained by ImageNet [25] and fixed during training.

Our model is trained using SGD for 400 epochs on Nvidia Titan V GPUs. We set the base learning rate to 2e-2 and reduce it to 2e-3 after 150 epochs. Momentum and weight decay of SGD are set to 0.9 and 5e-4, respectively. The input images have the size of 353×353. For data augmentation, we follow CaNet [38] to adopt random mirror, random rotation, random resize and random crop for both datasets. For online fine-tune, the iteration number is 20.

5.3 Results

We evaluate the performance of our proposed model with/without online refinement, and compare it with multiple state-of-the-art methods for few-shot segmentation. The results are reported in Tab. 1 and Tab. 2. It is worth mentioning that, for the two closely related methods CaNet [38] and PGNet [37], in addition to directly quoting the results from the original papers (for PASCAL-5^i), we also re-run the models and report the results, indicated as CaNet* and PGNet* in the tables. For CaNet* and PGNet*, the same support-query pairs are used for test as in our model. In this way, a strict comparison that further removes the difference in randomly sampling test pairs is conducted.

The results on COCO-20^i are given in Tab. 1. COCO-20^i is a very challenging dataset. As can be seen, on this task, even our offline model only has outperformed the four competitors in terms of the mean IoU. With the proposed online refinement, the performance of our model could be further boosted, making its advantage over the other methods in comparison more salient. The results on PASCAL-5^i are given in Tab. 2. This is a relatively easy task and all methods in comparison have better performance than what they do on COCO-20^i. The performance of our offline model is comparable to that of the second best method FW&B which builds on a more powerful backbone network ResNet-101. Compared with

Methods	Backbone	fold-0	fold-1	fold-2	fold-3	Mean
FW&B [21]	VGG16	18.35	16.72	19.59	25.43	20.2
FW&B [21]	ResNet101	16.98	17.98	20.96	28.85	21.19
PGNet*	ResNet50	32.24	30.51	31.61	29.73	31.02
CaNet*		**34.25**	34.44	30.87	**31.21**	32.69
Ours Offline	ResNet50	31.40	36.01	36.78	29.86	33.51
Ours Offline + Online		32.88	**36.20**	**37.44**	30.93	**34.36**

Table 1: Comparison with the state-of-the-art 1-shot segmentation performance on COCO-20^i. The symbol * indicates the model is re-run by ourselves.

Methods	Backbone	fold-0	fold-1	fold-2	fold-3	Mean
OSLSM [28]	VGG-16	33.6	55.3	40.9	33.5	40.8
co-FCN [22]		36.7	50.6	44.9	32.4	41.1
PL+SEG+PT [7]		-	-	-	-	42.7
AMP [29]		41.9	50.2	46.7	34.4	43.4
SG-One [39]		40.2	58.4	48.4	38.4	46.3
PANet [32]		42.3	58.0	51.1	41.2	48.1
CaNet [38]	ResNet50	52.5	65.9	51.3	51.9	55.4
PGNet [37]		56.0	66.9	50.6	50.4	56.0
CaNet*		51.11	66.09	50.06	52.57	54.96
PGNet*		53.63	65.70	48.54	49.28	54.29
FW&B [21]	ResNet-101	51.3	64.5	56.7	52.2	56.2
Ours Offline	ResNet-50	**56.85**	**67.52**	48.89	**53.23**	56.62
Ours Offline + Online		56.54	67.20	**51.56**	53.02	**57.08**

Table 2: Comparison with the state-of-the-art 1-shot segmentation performance on PASCAL-5^i. The symbol * indicates the model is re-run by ourselves.

CaNet* and PGNet* that use the same test pairs as ours, our offline model wins both of them with a large margin. Again, our online refinement could further improve our performance on this dataset consistently. Moreover, cross-referencing the results in Tab. 1 and Tab. 2, it seems that our online refinement contributes more to the performance improvement when the segmentation task is hard.

Fig. 3 shows six visual examples of segmentation results from our proposed BriNet and previous best models, CaNet and PGNet. Given the same query image, all of CaNet, PGNet and our BriNet are able to segment different classes with different support examples as guidance (the two rightmost columns in Fig. 3). However, our BriNet can generate more accurate and complete segmentation results compared with CaNet and PGNet, even when both of them totally fail (3rd column and 4th column). Our online refinement improves the model adaptation to agnostic object segmentation significantly (the last two rows in Fig. 3).

5.4 Ablation Study

To single out the contribution of each component proposed in our model, we conduct an ablation study in order to answer two questions: (i) How do IEM and MCM contribute to the performance of the offline model? (ii) Could our online refinement, as a general method, help other few-shot segmentation models improve the performance? 4-fold validation is used and the mean IoU values are reported.

IEM and MCM. To answer the first question, we compare our offline model with its

Model	COCO	PASCAL
BriNet w/o IEM	28.08	53.25
BriNet w/o MCM	30.64	53.49
BriNet	33.51	56.62

Table 3: Ablation study about IEM and MCM on COCO and PASCAL.

Model	COCO	PASCAL
CaNet*	32.69	54.96
CaNet* + Online	32.84	54.92
PGNet*	31.02	54.29
PGNet* + Online	31.89	54.71

Table 4: Ablation study about our online refinement.

variants that remove IEM and MCM, respectively. The results are given in Tab. 3. As seen, without either IEM or MCM, the performance of the offline model will significantly decrease on both PASCAL-5^i and COCO-20^i, showing the necessity of employing these two modules, as we argued before.

Online refinement. To answer the second question, we apply our online refinement to CaNet* and PGNet*, and the results are in Tab. 4. Significant improvement could be observed on the hard classes in COCO-20^i for both models. On PASCAL-5^i, although little effect is observed on CaNet*, our online refinement could help PGNet* to improve further. This experiment demonstrates the value of our online refinement method as a general strategy to improve few-shot segmentation.

Figure 3: Six visual examples (corresponding to six columns) from PASCAL-5^i under 1-shot segmentation. 1st row: the query images and support images (in the small window) with ground-truth segmentation. 2nd-5th rows: the query images and segmentation predicted by CaNet*, PGNet*, Offline and Offline + Online models, respectively. Our BriNet outperforms previous best frameworks and our online refinement algorithm improves the segmentation performance of offline model significantly. Best viewed in color.

6 Conclusions

In this paper we proposed BriNet, a novel framework for segmentation network with few-shot learning. Our model contributes the state-of-the-arts as follows. **First**, we introduce an information exchange module to boost the feature representations of the support and query images both. Besides, we represent the masked objects in the support image in a relatively more fine-grained way to better localize the objects in the query image. **Second**, we propose a new online refinement strategy to adapt the trained model to unseen test objects. Specifically, we tactically switch the roles of the query and the support images at the test stage and refine our model by minimizing the segmentation errors of the support images. In this way, we fully exploit the additional information in both the test query image and its supporters, which has not been well handled in the existing methods. The effectiveness of our model has been demonstrated in our experiment, which outperforms the state-of-the-arts methods by a margin.

References

[1] 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(12):2481–2495, Dec 2017. ISSN 1939-3539. doi: 10.1109/tpami.2016.2644615. URL http://dx.doi.org/10.1109/TPAMI.2016.2644615.

[2] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In Advances in neural information processing systems, pages 523–531, 2016.

[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, 2014.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4):834–848, Apr 2018. ISSN 2160-9292. doi: 10.1109/tpami.2017.2699184. URL http://dx.doi.org/10.1109/TPAMI.2017.2699184.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. Lecture Notes in Computer Science, page 833–851, 2018. ISSN 1611-3349. doi: 10.1007/978-3-030-01234-2_49. URL http://dx.doi.org/10.1007/978-3-030-01234-2_49.

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

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 88(2):303–338, June 2010.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

[10] Ross Girshick. Fast r-cnn. 2015 IEEE International Conference on Computer Vision (ICCV), Dec 2015. doi: 10.1109/iccv.2015.169. URL http://dx.doi.org/10.1109/ICCV.2015.169.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Jun 2014. doi: 10.1109/cvpr.2014.81. URL http://dx.doi.org/10.1109/CVPR.2014.81.

[12] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In 2011 International Conference on Computer Vision, pages 991–998, 2011.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/cvpr.2016.90.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. 2017 IEEE International Conference on Computer Vision (ICCV), Oct 2017. doi: 10.1109/iccv.2017.322. URL http://dx.doi.org/10.1109/ICCV.2017.322.

[15] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation, 2019.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. CoRR, abs/1709.01507, 2017. URL http://arxiv.org/abs/1709.01507.

[17] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, volume 2. Lille, 2015.

[18] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 2017. doi: 10.1109/cvpr.2017.549. URL http://dx.doi.org/10.1109/CVPR.2017.549.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. Lecture Notes in Computer Science, page 740–755, 2014. ISSN 1611-3349. doi: 10.1007/978-3-319-10602-1_48. URL http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[20] Tsendsuren Munkhdalai and Hong Yu. Meta networks, 2017.

[21] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. ArXiv, abs/1909.13140, 2019.

[22] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A. Efros, and Sergey Levine. Few-shot segmentation propagation with guided networks, 2018.

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

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, page 234–241, 2015. ISSN 1611-3349. doi: 10.1007/978-3-319-24574-4_28. URL http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, Apr 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL http://dx.doi.org/10.1007/s11263-015-0816-y.

[26] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In International conference on machine learning, pages 1842–1850, 2016.

[27] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Rogerio Feris, Abhishek Kumar, Raja Giryes, and Alex M. Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition, 2018.

[28] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. Procedings of the British Machine Vision Conference 2017, 2017. doi: 10.5244/c.31.167. URL http://dx.doi.org/10.5244/c.31.167.

[29] Mennatullah Siam, Boris Oreshkin, and Martin Jagersand. Adaptive masked proxies for few-shot segmentation, 2019.

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

[31] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 2018. doi: 10.1109/cvpr.2018.00131. URL http://dx.doi.org/10.1109/CVPR.2018.00131.

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

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 2018. doi: 10.1109/cvpr.2018.00813. URL http://dx.doi.org/10.1109/CVPR.2018.00813.

[34] Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In European Conference on Computer Vision, pages 616–634. Springer, 2016.

[35] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 2018. doi: 10.1109/cvpr.2018.00760. URL http://dx.doi.org/10.1109/CVPR.2018.00760.

[36] Yuwei Yang, Fanman Meng, Hongliang Li, Qingbo Wu, Xiaolong Xu, and Shuai Chen. A new local transformation module for few-shot segmentation. Lecture Notes in Computer Science, page 76–87, Dec 2019. ISSN 1611-3349.

doi: 10.1007/978-3-030-37734-2_7. URL http://dx.doi.org/10.1007/978-3-030-37734-2_7.

[37] Chenghui Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In ICCV 2019, 2019.

[38] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2019. doi: 10.1109/cvpr.2019.00536. URL http://dx.doi.org/10.1109/CVPR.2019.00536.

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

[40] 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), Jul 2017. doi: 10.1109/cvpr.2017.660. URL http://dx.doi.org/10.1109/CVPR.2017.660.

[41] Dongzhan Zhou, Xinchi Zhou, Hongwen Zhang, Shuai Yi, and Wanli Ouyang. Cheaper pre-training lunch: An efficient paradigm for object detection. arXiv preprint arXiv:2004.12178, 2020.