Class Agnostic Object Segmentation using Few-Shot Weakly Supervised Guidance

Anonymous authors
Paper under double-blind review

Abstract

Significant progress has been made recently in developing few-shot object segmentation methods which is shown to be successful using pixel-level, scribbles or bounding boxes. This paper takes another approach, i.e., only requiring few image-level labelled data for guiding a class agnostic segmentation network. We propose a novel multi-modal interaction module that utilizes a co-attention mechanism using both visual and semantic representation. Our model using image-level labels achieves 4.8% improvement over previously proposed image-level few-shot object segmentation. It also outperforms state-of-the-art methods that use weak bounding box supervision on PASCAL-5^i. We further propose a novel setup, Temporal Object Segmentation for Few-shot Learning (TOSFL) for videos. TOSFL provides a novel benchmark for video segmentation, which can be used on a variety of public video data such as Youtube-VOS.

1 Introduction

Semantic segmentation is a vital task for various range of applications such as satellite imagery semantic segmentation Nivaggioli & Randrianarivo (2019); Chan et al. (2019), robotics Cordts et al. (2016); Ros et al. (2016) and medical image segmentation Ronneberger et al. (2015); Heimann & Meinzer (2009). However, annotating large-scale data with different class categories is cost inefficient especially for developing nations which have limited resources De-Arteaga et al. (2018). One way to overcome this problem is to focus on learning few-shot segmentation, where the existing literature has mainly relied on manually labelled segmentation masks. A few recent works (Rakelly et al., 2018; Zhang et al., 2019b; Wang et al., 2019) started to conduct experiments using weak annotations such as scribbles or bounding boxes. However, these weak forms of supervision involve more manual work compared to image level labels, which can be collected from text and images publicly available on the web. Developing nations can benefit from few-shot segmentation in applications such as satellite imagery segmentation, but it can also act as a mean to generally help emerging machine learning startups. In order for robotics and machine learning startups in developing nations to thrive with limited resources available, it is critical to consider the capabilities of publicly available web data for different tasks including semantic segmentation. This motivates our focus on extending semantic segmentation models to learn new classes with image-level supervision.

We propose a novel method for learning a class agnostic object segmentation guided by few image-level labels in a meta-learning framework. The class agnostic segmentation model can be meta-trained on large-scale semantic segmentation datasets on certain base classes then extended to new classes relying on few image-level labelled support set images similar to web data. Limited research has been conducted on using image-level supervision for few-shot segmentation (Raza et al., 2019) which lags significantly behind its strongly supervised counterpart. Our class agnostic object segmentation is comprised mainly of a multi-modal interaction module that combines the visual input with neural word embeddings. One variant of our method iteratively guides a bi-directional co-attention between the support and the query sets using both visual and neural word embedding inputs and outperforms (Raza et al., 2019) by 4.8%. Most work in few-shot segmentation considers the *static* setting where query and support images do not have temporal relations. However, in real world applications such as robotics, segmentation methods can benefit from temporal continuity and multiple viewpoints. It may be of tremendous benefits to utilize temporal knowledge existing in video sequences. Observations that pixels moving together mostly belong to the same object seem to be very common in videos, and it can be exploited to improve segmentation accuracy. We propose

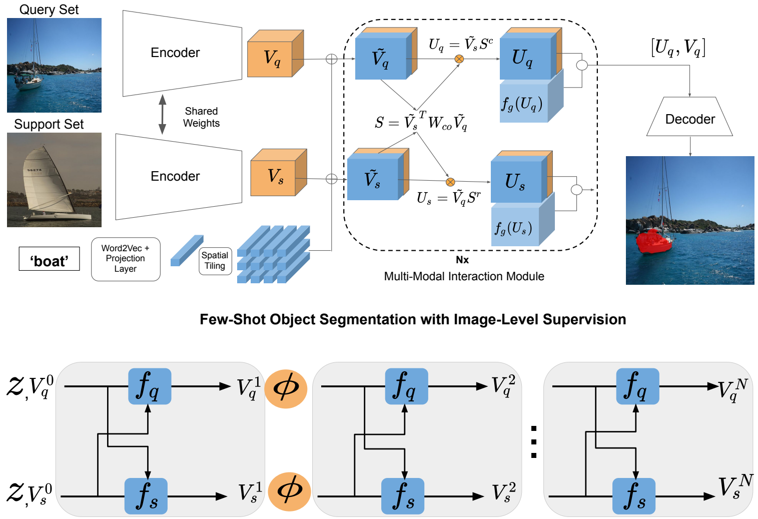

Figure 1: Architecture of Few-Shot Object segmentation model with co-attention and overview of the stacked co-attention. The $\oplus$ operator denotes concatenation, $\circ$ denotes element-wise multiplication. Only the decoder and multi-modal interaction module parameters are learned, while the encoder is pretrained on ImageNet.

a novel setup, temporal object segmentation with few-shot learning (TOSFL) to benefit from such assumptions.

2 Proposed Method

The human perception system is inherently multi-modal. Inspired from this and to leverage the learning of new concepts we propose a multi-modal interaction module that embeds semantic conditioning in the visual processing scheme as shown in Fig. 1. The overall model consists of: (1) Encoder. (2) Multi-modal Interaction module. (3) Segmentation Decoder. The multi-modal interaction module is described in detail in this section while the encoder and decoder modules are explained in Section 4.1. We follow a 1-way k-shot setting similar to (Shaban et al., 2017).

2.1 Multi-Modal Interaction Module

One approach for few-shot segmentation is to condition the model on visual features of the support set. However, without a mechanism to relate pixels from support to query and computing the affinities between them it is hard to capture the spatial relationships among both support and query. In order to leverage the interaction between support and query images we choose to use a co-attention module that computes a pixel-level affinity matrix. Nonetheless, even with coattention one of the main challenges in dealing with the image-level annotation in few-shot segmentation is that quite often both support and query images may contain a few salient common objects from different classes. Inferring a good prototype for the object of interest from multi-object support images without relying on pixel-level masks or even bounding boxes becomes particularly challenging. Yet, it is exactly in this situation, that we can expect the semantic word embeddings to be useful at helping to disambiguate the object relationships across support and query images.

Below we discuss the technical details behind the implementation of this idea. Initially, in a k-shot setting, a base network is used to extract features from i^{th} support set image I_s^i and from the query image I_q, which we denote as $V_s \in R^{W \times H \times C}$ and $V_q \in R^{W \times H \times C}$. Here H and W denote the height and width of feature maps, respectively, while C denotes the number of feature channels. Furthermore, a projection layer is used on the semantic word embeddings to construct $z \in R^d$ ($d = 256$). It is then spatially tiled and concatenated with the visual features resulting in flattened matrix representation $\tilde{V_q} \in R^{C \times WH}$ and $\tilde{V_s} \in R^{C \times WH}$. An affinity matrix S is computed to capture the similarity between them via a fully connected layer $W_{co} \in R^{C \times C}$ learning the correlation between feature channels:

$$S = \tilde{V_s}^T W_{co} \tilde{V_q}.$$

The affinity matrix $S \in R^{WH \times WH}$ relates each pixel in $\tilde{V_q}$ and $\tilde{V_s}$. A softmax operation is performed on S row-wise and column-wise depending on the desired direction of relation:

$$S^c = \text{softmax}(S), \quad S^r = \text{softmax}(S^T)$$

For example, column $S^c_{*,j}$ contains the relevance of the j^{th} spatial location in V_q with respect to all spatial locations of V_s, where $j = 1, ..., WH$. The normalized affinity matrix is used to compute attention summaries U_q and U_s:

$$U_q = \tilde{V_s} S^c, \quad U_s = \tilde{V_q} S^r.$$

The attention summaries are further reshaped such that $U_q, U_s \in R^{W \times H \times C}$ and gated using a gating function f_g with learnable weights W_g and bias b_g:

$$f_g(U_q) = \sigma(W_g * U_q + b_g),$$
$$U_q = f_g(U_q) \circ U_q.$$

Here the $\circ$ operator denotes element-wise multiplication. The gating function restrains the output to the interval $[0, 1]$ using a sigmoid activation function σ in order to mask the attention summaries. The gated attention summaries U_q are concatenated with the original visual features V_q to construct the final output from the attention module to the decoder.

We propose to stack the multi-modal interaction module to learn an improved representation. Stacking allows for multiple iterations between the support and the query images for an iterative refinement as shown in Fig. 1. The co-attention module has two streams f_q, f_s that are responsible for processing the query image and the support set images respectively. The inputs to the co-attention module, V_q^i and V_s^i, represent the visual features at iteration i for query image and support image respectively. In the first iteration, V_q^0 and V_s^0 are the output visual features from the encoder. Each multi-modal interaction then follows the recursion $\forall i = 0, .., N-1$:

$$V_q^{i+1} = \phi(V_q^i + f_q(V_q^i, V_s^i, z))$$

The nonlinear projection ϕ is performed on the output from each iteration, which is composed of a 1x1 convolutional layer followed by a ReLU activation function. We use residual connections in order to improve the gradient flow and prevent vanishing gradients. The support set features $V_s^i, \forall i = 0, .., N-1$ are computed similarly.

3 Temporal Object Segmentation with Few-shot Learning Setup

We propose a novel few-shot video object segmentation (VOS) task. In this task, the image-level label for the support frame is provided to guide object segmentation in the query frames. Both instance-level and category-level setup are proposed. In the instance-level setup the query and support images are temporally related. In the category-level setup the support set and query sets are sampled from different sequences for the same object category. Even in the category-level the query set with multiple query images can be temporally related. This provides a potential research direction for ensuring the temporal stability of the learned representation that is used to segment multiple query images. The task is designed as a binary segmentation problem following Shaban et al. (2017) and the categories are split into multiple folds, consistent with existing few-shot segmentation tasks defined on Pascal-5^i and MS-COCO. This design ensures that the proposed task assesses the ability of few-shot video object segmentation algorithms to generalize over unseen classes. We utilize Youtube-VOS dataset training data which has 65 classes, and we split them into 5 folds. Each fold has 13 classes that are used as novel classes, while the rest are used in the meta-training phase. In the instance-level mode a randomly sampled class Y^s and sequence $V = \{I_1, I_2, ..., I_N\}$ are used to construct the support set $S_p = \{(I_1, Y_1^s)\}$ and query images I_i. For each query image a ground-truth binary segmentation mask M_Y^s is constructed by labelling all the instances belonging to Y^s as foreground. Accordingly, the same image can have multiple binary segmentation masks depending on the sampled Y^s. During the category-level mode different sequences $V^s = \{I_1^s, I_2^s, ..., I_N^s\}$ and $V^q = \{I_1^q, I_2^q, ..., I_N^q\}$ for the same class Y^s are sampled. Then random frames $\{I_i^s\}_{i=0}^k$ sampled from V^s and $\{I_i^q\}_{i=0}^l$ similarly are used to construct the support and query sets respectively.

Table 1: Quantitative results for 1-way, 1-shot segmentation on the PASCAL-5^i dataset showing mean-Iou and binary-IoU. P: use pixel-wise segmentation masks for supervision. IL: use Image-Level labels. BB: use bounding boxes. Red: validation scheme following Zhang et al. (2019b). Blue: validation scheme following Wang et al. (2019)

		1-shot		5-shot
Method	Type	mIoU	bIoU	mIoU
FG-BG	P	-	55.1	-
OSLSM (Shaban et al., 2017)	P	40.8	-	43.9
CoFCN (Rakelly et al., 2018)	P	41.1	60.1	41.4
PLSeg (Dong & Xing, 2018)	P	-	61.2	-
AMP (Siam et al., 2019)	P	43.4	62.2	46.9
PANet (Wang et al., 2019)	P	48.1	66.5	55.7
CANet (Zhang et al., 2019b)	P	55.4	66.2	57.1
PGNet (Zhang et al., 2019a)	P	56.0	69.9	58.5
CANet (Zhang et al., 2019b)	BB	**52.0**	-	-
PANet (Wang et al., 2019)	BB	**45.1**	-	**52.8**
(Raza et al., 2019)	IL	-	**58.7**	-
Ours(V+S)-1	IL	**53.5**	**65.6**	-
Ours(V+S)-2	IL	**50.5** ±0.7	**64.1** ±0.4	**51.7** ±0.07

4 Experiments

4.1 Experimental Setup

Network Details: We utilize a ResNet-50 (He et al., 2016) encoder pre-trained on ImageNet (Deng et al., 2009) to extract visual features. The segmentation decoder is comprised of an iterative optimization module (IOM) (Zhang et al., 2019b) and an atrous spatial pyramid pooling (ASPP) (Chen et al., 2017a;b).

Meta-Learning Setup: We sample 12,000 tasks during the meta-training stage. Since there exist already large-scale semantic segmentation datasets so their usage to learn base classes and a better representation to generalize to new classes is deemed acceptable. In order to evaluate test performance, we average accuracy over 5000 tasks with support and query sets sampled from the meta-test dataset D_{test} belonging to classes L_{test}. We perform 5 training runs with different random generator seeds and report the average of the 5 runs and the 95% confidence interval.

Training and Evaluation Details: PASCAL-5^i splits PASCAL-VOC 20 classes into 4 folds each having 5 classes. The mean IoU Shaban et al. (2017) and binary IoU Rakelly et al. (2018) are the two metrics used for the evaluation process. We have noticed some deviation in the validation schemes used in previous works. Zhang et al. (2019b) follow a procedure where the validation is performed on the test classes to save the best model, whereas Wang et al. (2019) rather train for a fixed number of iterations. We choose the more challenging approach in (Wang et al., 2019). During the meta-training, we freeze ResNet-50 encoder weights while learning both the multi-modal interaction module and the decoder. We train all models using momentum SGD with learning rate 0.01 that is reduced by 0.1 at epoch 35, 40 and 45, momentum 0.9, and weight decay of 5x10^{-4}.

4.2 Comparison to the state-of-the-art

We compare the result of our best variant, *i.e*: Stacked Co-Attention (V+S) against the other state of the art methods for 1-way 1-shot and 5-shot segmentation on PASCAL-5^i in Table 1. We report the results for different validation schemes. Ours(V+S)-1 follows (Zhang et al., 2019b) and Ours(V+S)-2 follows (Wang et al., 2019). Without the utilization of segmentation mask or even sparse annotations, our method with the least supervision of image level labels performs (53.5%) close to the current state of the art strongly supervised methods (56.0%) in 1-shot case and outperforms the ones that use bounding box annotations. It improves over the previously proposed image-level supervised method with a significant margin (4.8%). For the k-shot extension of our method we perform average of the attention summaries on the k-shot support samples.

Table 2: Ablation Study for different components with 1 run on Pascal-5^i and Youtube-VOS. V: visual, S: semantic. SCoAtt: Stack Co-Attention. Cond: Concatenation based conditioning.

(a) Pascal-5^i

Method	mIoU
V-Cond	42.7
V-CoAtt	44.6
V+S-Cond	50.1
V+S-CoAtt	50.2
V+S-SCoAtt	**51.0**

(b) Youtube-VOS

Method	mIoU
V+S-Cond	42.3
V+S-SCoAtt	**43.7**

Table 3: Ablation study with 5 runs on Pascal-5^i and Youtube-VOS for different variants of our method.

(a) Pascal-5^i

Method	1-shot	5-shot
V-CoAtt	44.4 ± 0.3	49.1 ± 0.3
S-Cond	**51.2** ± 0.6	51.4 ± 0.3
V+S-SCoAtt	50.5 ± 0.7	**51.7** ± 0.07

(b) Youtube-VOS

Level	Method	Mean-IoU
Instance	V-CoAtt	38.0 ± 0.7
	S-Cond	41.7 ± 0.7
	V+S-SCoAtt	**43.8** ± 0.5
Category	V-CoAtt	36.1
	S-Cond	**37.7**
	V+S-SCoAtt	37.6

4.3 Ablation Study

We perform an ablation study to evaluate different components of our method. Table 2 show results for 1 run and compares using a simple conditioning on the support set features through concatenation with the query visual features against performing co-attention between support and query feature maps. It shows clearly the benefit from performing co-attention. Nonetheless, visual features solely is not capable to disambiguate between different common objects and the visual with semantic embeddings even with simple concatenation shows an improvement. Further stacking the co-attention module proves to improve the results as well specifically on Youtube-VOS.

Table 3 shows the results for 5 runs on the three variants we proposed on both datasets. It shows that using the visual features only (V), lags behind utilizing word embeddings (S). Semantic representation helps to resolve the ambiguity and improves the result. Going from 1 to 5 shots, the (V) method improves, because multiple shots are likely to repeatedly contain the object of interest and the associated ambiguity decreases, but still it lags behind both variants supported by semantic input. Interestingly, our results show that the baseline of conditioning on semantic representation is a very competitive variant, but it is not able to benefit from multiple shots in the support set unlike (V+S). It also the (V+S) joint visual and semantic processing on Youtube-VOS provides significant gain in the instance-level setup.It is worth noting that unlike the conventional video object segmentation setups, the proposed video object segmentation task poses the problem as a binary segmentation task conditioned on the image-level label. Both support and query frames can have multiple salient objects appearing in them, however the algorithm has to segment only one of them corresponding to the image-level label provided in the support frame.

5 Conclusion

In this paper we proposed a multi-modal interaction module that relates the support set image and query image using both visual and word embeddings. The main takeaways from the experiments are that: (i) few-shot segmentation significantly benefits from utilizing word embeddings and (ii) it is viable to perform high quality few-shot segmentation using stacked joint visual semantic processing with weak image-level labels. Our proposed TOSFL setup can provide a potential research direction in understanding the temporal coherence of the representation learned for novel categories.

REFERENCES

Lyndon Chan, Mahdi S Hosseini, and Konstantinos N Plataniotis. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *arXiv preprint arXiv:1912.11186*, 2019.

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, 2017a.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017b.

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

Maria De-Arteaga, William Herlands, Daniel B. Neill, and Artur Dubrawski. Machine learning for the developing world. *ACM Trans. Manage. Inf. Syst.*, 9(2):9:1–9:14, August 2018. ISSN 2158-656X. doi: 10.1145/3210548. URL `http://doi.acm.org/10.1145/3210548`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Nanqing Dong and Eric P. Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, pp. 4, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis*, 13(4):543–563, 2009.

Adrien Nivaggioli and Hicham Randrianarivo. Weakly supervised semantic segmentation of satellite images. In *2019 Joint Urban Remote Sensing Event (JURSE)*, pp. 1–4. IEEE, 2019.

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

Hasnain Raza, Mahdyar Ravanbakhsh, Tassilo Klein, and Moin Nabi. Weakly supervised one shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

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

Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5249–5258, 2019.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9197–9206, 2019.

Chi 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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9587–9595, 2019a.

Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5217–5226, 2019b.