Weakly Supervised Few-shot Object Segmentation using Co-Attention with Visual and Semantic Embeddings

Mennatullah Siam[1,4*], **Naren Doraiswamy**[2*], **Boris N. Oreshkin**[3*], **Hengshuai Yao**[4]
and **Martin Jagersand**[1]
[1] University of Alberta
[2] Indian Institute of Science
[3] Element AI
[4] HiSilicon, Huawei Research
{mennatul, mj7}@ualberta.ca, narend@iisc.ac.in, boris@elementai.com, hengshuai.yao@huawei.com

Abstract

Significant progress has been made recently in developing few-shot object segmentation methods. Learning is shown to be successful in few-shot segmentation settings, using pixel-level, scribbles and bounding box supervision. This paper takes another approach, i.e., only requiring image-level label for few-shot object segmentation. We propose a novel multi-modal interaction module for few-shot object segmentation that utilizes a co-attention mechanism using both visual and word embedding. 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. Our results show that few-shot segmentation benefits from utilizing word embeddings, and that we are able to perform few-shot segmentation using stacked joint visual semantic processing with weak image-level labels. We further propose a novel setup, Temporal Object Segmentation for Few-shot Learning (TOSFL) for videos. TOSFL can be used on a variety of public video data such as Youtube-VOS, as demonstrated in both instance-level and category-level TOSFL experiments.

1 Introduction

Existing literature in few-shot object segmentation 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. Limited research has been conducted on using image-level supervision for few-shot segmentation [Raza *et al.*, 2019].

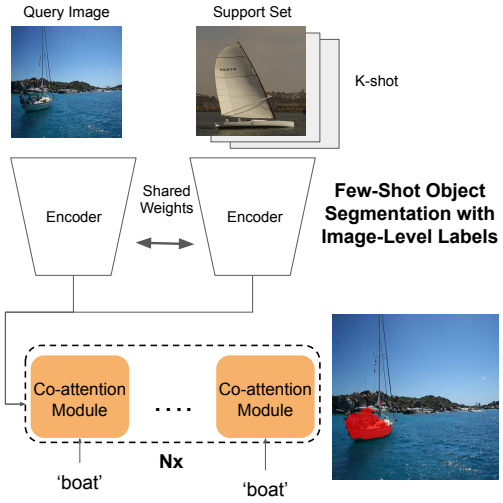


Figure 1: Overview of stacked co-attention to relate the support set and query image using image-level labels. Nx: Co-attention stacked N times. "K-shot" refers to using K support images.

On the other hand, deep semantic segmentation networks are very successful when trained and tested on relatively large-scale manually labelled datasets such as PASCAL-VOC [Everingham *et al.*, 2015] and MS-COCO [Lin *et al.*, 2014]. However, the number of object categories they cover is still limited despite the significant sizes of the data used. The limited number of annotated objects with pixel-wise labels included in existing datasets restricts the applicability of deep learning in inherently open-set domains such as robotics [Dehghan *et al.*, 2019; Pirk *et al.*, 2019; Cordts *et al.*, 2016].

In this paper, we propose a multi-modal interaction module to bootstrap the efficiency of weakly supervised few-shot object segmentation by combining the visual input with neural word embeddings. Our method iteratively guides a bi-directional co-attention between the support and the query sets using both visual and neural word embedding inputs, using only image-level supervision as shown in Fig. 1. It outperforms [Raza *et al.*, 2019] by 4.8% and improves over meth-

*equally contributing

ods that use bounding box supervision [Zhang *et al.*, 2019b; Wang *et al.*, 2019]. We use the term 'weakly supervised few-shot' to denote that our method during the inference phase only utilizes few-shot image-level labelled data to guide the class agnostic segmentation. Additionally we propose a novel setup, temporal object segmentation with few-shot learning (TOSFL). The TOSFL setup for video object segmentation generalizes to novel object classes as can be seen in our experiments on Youtube-VOS dataset [Xu *et al.*, 2018]. TOSFL only requires image-level labels for the first frames (support images). The query frames are either the consecutive frames (instance-level) or sampled from another sequence having the same object category (category-level). The TOSFL setup is interesting as it is closer to the nature of learning novel objects by human than the strongly supervised static segmentation setup. Our setup relies on the image-level label for the support image to segment different parts from the query image conditioned on the word embeddings of this image-level label. In order to ensure the evaluation for the few-shot method is not biased to a certain category, it is best to split into multiple folds and evaluate on different ones similar to [Shaban *et al.*, 2017].

1.1 Contributions

- We propose a novel few-shot object segmentation algorithm based on a multi-modal interaction module trained using image-level supervision. It relies on a multi-stage attention mechanism and uses both visual and semantic representations.
- We propose a novel weakly supervised few-shot video object segmentation setup. It complements the existing few-shot object segmentation benchmarks by considering a practically important use case not covered by previous datasets. Video sequences are provided instead of static images which can simplify the few-shot learning problem.
- We conduct a comparative study of different architectures proposed in this paper to solve few-shot object segmentation with image-level supervision. Our method compares favourably against the state-of-the-art methods relying on pixel-level supervision and outperforms the most recent methods using weak annotations [Raza *et al.*, 2019; Wang *et al.*, 2019; Zhang *et al.*, 2019b].

2 Related Work

2.1 Few-shot Object Segmentation

[Shaban *et al.*, 2017] proposed the first few-shot segmentation method using a second branch to predict the final segmentation layer parameters. [Rakelly *et al.*, 2018] proposed a guidance network for few-shot segmentation where the guidance branch receives the support set image-label pairs. [Dong and Xing, 2018] utilized the second branch to learn prototypes. [Zhang *et al.*, 2019b] proposed a few-shot segmentation method based on a dense comparison module with a siamese-like architecture that uses masked average pooling to extract features on the support set, and an iterative optimization module to refine the predictions. [Siam *et al.*, 2019] proposed a method to perform few-shot segmentation using adaptive masked proxies to directly predict the parameters of the novel classes. [Zhang *et al.*, 2019a] in a more recent work proposed a pyramid graph network which learns attention weights between the support and query sets for further label propagation. [Wang *et al.*, 2019] proposed prototype alignment by performing both support-to-query and query-to-support few-shot segmentation using prototypes.

The previous literature focused mainly on using strongly labelled pixel-level segmentation masks for the few examples in the support set. It is labour intensive and impractical to provide such annotations for every single novel class, especially in certain robotics applications that require to learn online. A few recent works experimented with weaker annotations based on scribbles and/or bounding boxes [Rakelly *et al.*, 2018; Zhang *et al.*, 2019b; Wang *et al.*, 2019]. In our opinion, the most promising direction to solve the problem of intense supervision requirements in the few-shot segmentation task, is to use publicly available web data with image-level labels. [Raza *et al.*, 2019] made a first step in this direction by proposing a weakly supervised method that uses image-level labels. However, the method lags significantly behind other approaches that use strongly labelled data.

2.2 Attention Mechanisms

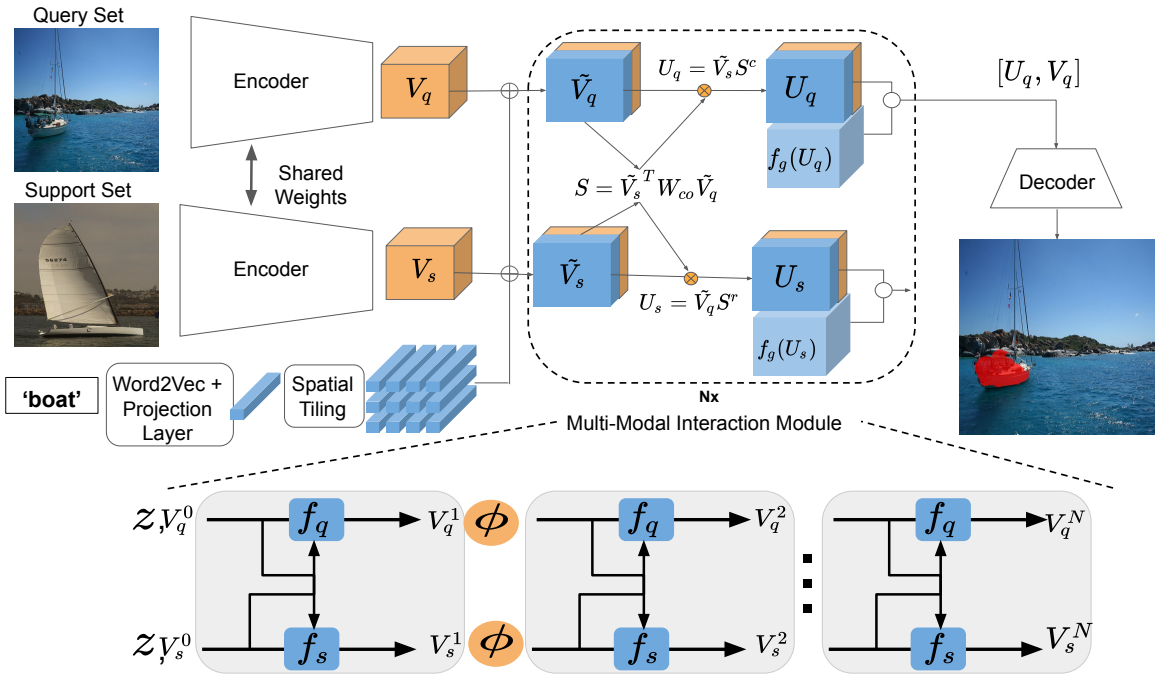
Attention was initially proposed for neural machine translation models [Bahdanau *et al.*, 2014]. Several approaches were proposed for utilizing attention. [Yang *et al.*, 2016] proposed a stacked attention network which learns attention maps sequentially on different levels. [Lu *et al.*, 2016] proposed co-attention to solve a visual question and answering task by alternately shifting attention between visual and question representations. [Lu *et al.*, 2019] used co-attention in video object segmentation between frames sampled from a video sequence. [Hsieh *et al.*, 2019] rely on attention mechanism to perform one-shot object detection. However, they mainly use it to attend to the query image since the given bounding box provides them with the region of interest in the support set image. To the best of our knowledge, this work is the first one to explore the bidirectional attention between support and query sets as a mechanism for solving the few-shot image segmentation task with image-level supervision.

3 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. 2. 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 5.1. We follow a 1-way k-shot setting similar to [Shaban *et al.*, 2017].

3.1 Multi-Modal Interaction Module

One of the main challenges in dealing with the image-level annotation in few-shot segmentation is that quite often both



Few-Shot Object Segmentation with Image-Level Supervision

Figure 2: 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.

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 cues 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 depicted in Fig. 2. 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 representations $\tilde{V}_q \in R^{C \times WH}$ 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:

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

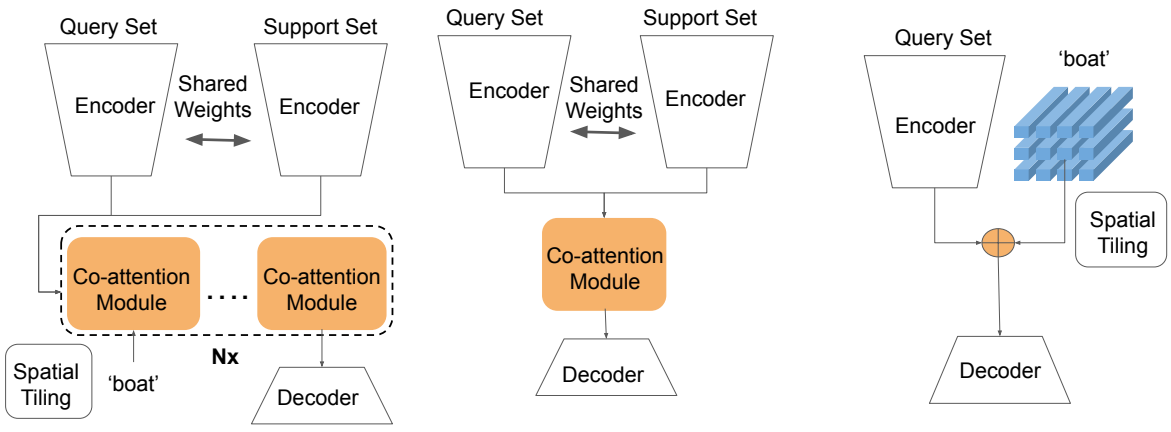
For example, column $S^c_{*,j}$ contains the relevance of the j^{th} spatial location in V_q with respect to all spatial locations of V_s, where $j = 1, ..., WH$. The normalized affinity matrix is used to compute attention summaries U_q and U_s:

$$U_q = \tilde{V}_s S^c, \quad U_s = \tilde{V}_q S^r.$$

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.



(a) V+S (b) V (c) S

Figure 3: Different variants for image-level labelled few-shot object segmentation. V+S: Stacked Co-Attention with Visual and Semantic representations. V: Co-Attention with Visual features only. S: Conditioning on semantic representation only from word embeddings.

3.2 Stacked Gated Co-Attention

We propose to stack the multi-modal interaction module described in Section 3.1 to learn an improved representation. Stacking allows for multiple iterations between the support and the query images. 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.

4 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.

5 Experiments

In this section we demonstrate results of experiments conducted on the PASCAL-5^i dataset [Shaban *et al.*, 2017] compared to state of the art methods in section 5.2. Not only do we set strong baselines for image level labelled few shot segmentation and outperform previously proposed work [Raza *et al.*, 2019], but we also perform close to the state of the art conventional few shot segmentation methods that use detailed pixel-wise segmentation masks. We then demonstrate the results for the different variants of our approach depicted in Fig. 3 and experiment with the proposed TOSFL setup in section 5.3.

5.1 Experimental Setup

We utilize a ResNet-50 [He *et al.*, 2016] encoder pre-trained on ImageNet [Deng *et al.*, 2009] to extract visual features.

		1-shot						5-shot				
Method	Type	1	2	3	4	mIoU	bIoU	1	2	3	4	mIoU
[Shaban *et al.*, 2017]	P	33.6	55.3	40.9	33.5	40.8	-	35.9	58.1	42.7	39.1	43.9
[Rakelly *et al.*, 2018]	P	36.7	50.6	44.9	32.4	41.1	60.1	37.5	50.0	44.1	33.9	41.4
[Dong and Xing, 2018]	P	-	-	-	-	-	61.2	-	-	-	-	-
[Siam *et al.*, 2019]	P	41.9	50.2	46.7	34.7	43.4	62.2	41.8	55.5	50.3	39.9	46.9
[Wang *et al.*, 2019]	P	42.3	58.0	51.1	41.2	48.1	66.5	51.8	64.6	59.8	46.5	55.7
[Zhang *et al.*, 2019b]	P	52.5	65.9	51.3	51.9	55.4	66.2	55.5	67.8	51.9	53.2	57.1
[Zhang *et al.*, 2019a]	P	56.0	66.9	50.6	50.4	56.0	69.9	57.7	68.7	52.9	54.6	58.5
[Zhang *et al.*, 2019b]	BB	-	-	-	-	**52.0**	-	-	-	-	-	-
[Wang *et al.*, 2019]	BB	-	-	-	-	**45.1**	-	-	-	-	-	**52.8**
[Raza *et al.*, 2019]	IL	-	-	-	-	-	**58.7**	-	-	-	-	-
Ours(V+S)-1	IL	49.5	65.5	50.0	49.2	**53.5**	**65.6**	-	-	-	-	-
Ours(V+S)-2	IL	42.5	64.8	48.1	46.5	**50.5**	**64.1**	45.9	65.7	48.6	46.6	**51.7**
						±0.7	±0.4					±0.07

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

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; Chen *et al.*, 2017b]. The IOM module takes the output feature maps from the multi-modal interaction module and the previously predicted probability map in a residual form.

We sample 12,000 tasks during the meta-training stage. 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.

PASCAL-5^i splits PASCAL-VOC 20 classes into 4 folds each having 5 classes. The mean IoU and binary IoU are the two metrics used for the evaluation process. The mIoU computes the intersection over union for all 5 classes within the fold and averages them neglecting the background. Whereas the bIoU metric proposed by [Rakelly *et al.*, 2018] computes the mean of foreground and background IoU in a class agnostic manner. 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] do not perform validation and 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 and momentum 0.9. L2 regularization with a factor of $5\text{x}10^{-4}$ is used to avoid over-fitting. Batch size of 4 and input resolution of 321×321 are used during training with random horizontal flipping and random centered cropping for the support set. An input resolution of 500×500 is used for the meta-testing phase similar to [Shaban *et al.*, 2017]. In each fold the model is meta-trained for a maximum number of 50 epochs on the classes outside the test fold on pascal-5^i, and 20 epochs on both MS-COCO and Youtube-VOS.

Method	Type	1-shot	5-shot
[Wang *et al.*, 2019]	P	20.9	29.7
Ours-(V+S)	IL	15.0	15.6

Table 2: Quantitative Results on MS-COCO Few-shot 1-way.

5.2 Comparison to the State-of-the-art

We compare the result of our best variant (see Fig. 3), *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 during the meta-training on the k-shot samples from the support set. Table 2 demonstrates results on MS-COCO [Lin *et al.*, 2014] compared to the state of the art method using pixel-wise segmentation masks for the support set.

5.3 Ablation Study

We perform an ablation study to evaluate different variants of our method depicted in Fig. 3. Table 3 shows the results on the three variants we proposed on PASCAL-5^i. It clearly shows that using the visual features only (V-method), lags 5% behind utilizing word embeddings in the 1-shot case. This is mainly due to two reasons having multiple common objects between the support set and the query image and tendency

(a) 'bicycle' (b) 'bottle' (c) 'bird'

(d) 'bicycle' (e) 'bird' (f) 'boat'

Figure 4: Qualitative evaluation on PASCAL-5^i 1-way 1-shot. The support set and prediction on the query image are shown in pairs.

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

Table 3: Ablation Study on 4 folds of Pascal-5^i for few-shot segmentation for different variants showing mean-IoU. V: visual, S: semantic. V+S: both features.

Method	1	2	3	4	5	Mean-IoU
V	40.8	34.0	44.4	35.0	35.5	38.0 ± 0.7
S	42.7	40.8	48.7	38.8	37.6	41.7 ± 0.7
V+S	**46.1**	**42.0**	**50.7**	**41.2**	**39.2**	**43.8** ± 0.5

Table 4: Ablation Study on 5 folds on Youtube-VOS Instance-level TOSFL. V: visual, S: semantic. V+S: both features.

to segment base classes used in meta-training. Semantic representation obviously helps to resolve the ambiguity and improves the result significantly as shown in Figure 5. 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: in the 1-shot case it even outperforms the (V+S) variant. However, the bottleneck in using the simple scheme to integrate semantic representation depicted in Fig. 3c is that it is not able to benefit from multiple shots in the support set. The (V+S)-method in the 5-shot case improves over the 1-shot case by 1.2% on average over the 5 runs, which confirms its ability to effectively utilize more abundant visual features in the 5-shot case. One reason could explain the strong performance of the (S) variant. In the case of a single shot, the word embedding pretrained on a massive text database may provide a more reliable guidance signal than a single image containing multiple objects that does not necessarily have visual features close to the object in the query image.

(a) Label 'Bike' (b) Prediction (V) (c) Prediction (V+S)

Figure 5: Visual Comparison between the predictions from two variants of our method.

Table 4 shows the results on our proposed novel video object segmentation instance-level task, comparing variants of the proposed approach. As previously, the baseline V-method based on co-attention module with no word embeddings, lags behind both S- and (V+S)-methods. 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. According to our observations, this multi-object situation occurs in this task much more frequently than *e.g.* in the case of Pascal-5^i. Additionally, not only the target, but all the nuisance objects present in the video sequence will relate via different viewpoints or deformations. We demonstrate in Table 4 that the (V+S)-method's joint visual and semantic processing in such scenario clearly provides significant gain.

6 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. We proposed to meta-learn a stacked co-attention module that guides the segmentation of the query based on the support set features and vice versa. The two main takeaways from the experiments are that

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**

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

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

Table 6: Ablation Study for different components with 1 run on Youtube-VOS. V: visual, S: semantic. SCoAtt: Stack Co-Attention. Cond: Concatenation based conditioning.

(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.

A Additional Results

In this section we present additional qualitative and quantitative results on pascal-5^i and Youtube-VOS datasets. We show more qualitative results on pascal-5^i that motivate the benefit of using both visual and semantic features in Fig. 7. It shows in first two rows that when both the query and support set have multiple common objects semantic features help to disambiguate this situation. It also shows in the last two rows that with visual features only there is a higher tendency to segment partly pixels belonging to classes that were provided in the meta-training stage as well. In this case semantic features again help to disambiguate this situation.

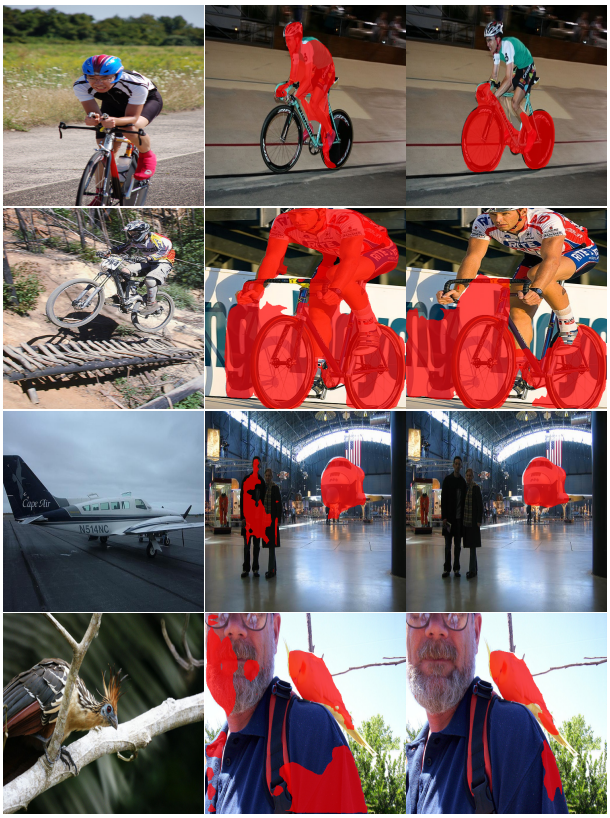
We further show another ablation study to evaluate different components in Tables **??**. We compare using a simple conditioning on the support set features through concatenation with the query visual features against performing co-attention between support and query feature maps. It shows clearly the benefit from performing co-attention. Nonetheless, visual features solely is not capable to disambiguate the above mentioned situations, while the visual with semantic features even with simple concatenation shows an improvement. Further combining semantic features with stacked co-attention shows further gain on both pascal-5^i and Youtube-VOS. In Table 7 we compare our variants in category-level TOSFL setup, in which the support set is sampled from a different sequence than the query set. It shows similar conclusions to Pascal-5^i results.

Finally we visualize the output gated attention maps in Fig. 6 to show the intermediate output from performing co-attention with both visual and semantic features. It demonstrates that our multi-modal interaction module in the first row successfully attends to the pixels belonging to the novel class.

Method	mIoU
V-CoAtt	36.1
S-Cond	**37.7**
V+S-SCoAtt	37.6

Table 7: Ablation study for different variants on the Category Level TOSFL 1-shot on Youtube-VOS.

Figure 6: Visualizing Output from Gated Co-Attention on PASCAL-5^i query images. (a) Support Set Image. (b) Query Set Image. (c) Gated Attention Map Output.



(a) Support Set (b) Ours(V) (c) Ours (V+S)

Figure 7: Qualitative analysis on fold 0 pascal-5^i between our method (V+S) and ours (V) that can not disambiguate multiple common objects and is biased toward base classes used in training.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Chen *et al.*, 2017a] 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, 2017.

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

[Cordts *et al.*, 2016] 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*, pages 3213–3223, 2016.

[Dehghan *et al.*, 2019] Masood Dehghan, Zichen Zhang, Mennatullah Siam, Jun Jin, Laura Petrich, and Martin Jagersand. Online object and task learning via human robot interaction. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2132–2138. IEEE, 2019.

[Deng *et al.*, 2009] 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*, pages 248–255. Ieee, 2009.

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

[Everingham *et al.*, 2015] Mark Everingham, S.M. Ali Eslami, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[He *et al.*, 2016] 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*, pages 770–778, 2016.

[Hsieh *et al.*, 2019] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. In *Advances in Neural Information Processing Systems*, pages 2721–2730, 2019.

[Lin *et al.*, 2014] 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. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[Lu *et al.*, 2019] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3623–3632, 2019.

[Pirk *et al.*, 2019] Sören Pirk, Mohi Khansari, Yunfei Bai, Corey Lynch, and Pierre Sermanet. Online object representations with contrastive learning. *arXiv preprint arXiv:1906.04312*, 2019.

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

[Raza *et al.*, 2019] 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*, pages 0–0, 2019.

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

[Siam *et al.*, 2019] 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*, pages 5249–5258, 2019.

[Wang *et al.*, 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*, pages 9197–9206, 2019.

[Xu *et al.*, 2018] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[Zhang *et al.*, 2019a] 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*, pages 9587–9595, 2019.

[Zhang *et al.*, 2019b] 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*, pages 5217–5226, 2019.