Temporal Transductive Inference for Few-Shot Video Object Segmentation

Mennatullah Siam
York University
Toronto, Canada
msiam@eecs.yorku.ca

Konstantinos G. Derpanis
York University
Toronto, Canada
kosta@eecs.yorku.ca

Richard P. Wildes
York University
Toronto, Canada
wildes@cse.yorku.ca

Abstract

Few-shot video object segmentation (FS-VOS), where the query images to be segmented belong to a video, recently has been introduced but is still under-explored. We propose a simple but effective temporal transductive inference (TTI) approach that uses the coherence across time in videos to improve the segmentation with a few-shot support set. We employ both temporally global and local constraints in videos. Global constraints focus on learning a consistent prototype on the sequence level, whereas local constraints focus on a coherent foreground/background region proportion within a local temporal window. Our model outperforms the state-of-the-art attention-based counterpart on few-shot Youtube-VIS by 2.8% in mean intersection over union (mIoU). Additionally, we propose a more realistic FS-VOS protocol that operates cross-domain. Our method outperforms the transductive inference baseline that uses static images with $\approx$1.3% improvement on two different benchmarks. These results demonstrate that our method provides a promising direction towards a label efficient approach of annotating video datasets with special applicability to rare classes that occur in different robotics settings such as autonomous driving. An online demo of our approach will be available at https://msiam.github.io/tti/.

1 Introduction

One of the major bottlenecks of deep learning methods in video segmentation [10] is the need for large-scale datasets that require laborious annotation effort and high cost. This shortcoming motivates our direction towards the few-shot video object segmentation task, which focuses on learning to segment a set of base classes with large-scale labelled datasets, while allowing generalization to novel classes with few labelled examples. Although few-shot object segmentation has been widely investigated in recent years [7, 1, 9, 12, 4], few-shot video object segmentation has been under explored [8, 2]. Transductive inference [1], which benefits from unlabelled query images, has shown a promising direction in few-shot segmentation, but has only been proposed for static images.

In this paper, we present the first attempt to perform temporal transductive inference, through using the unlabelled query frames for a video. Our model enforces temporal consistency globally on the sequence level and locally within a temporal window, unlike [1] that finetunes on static images separately without enforcing any constraints. Our approach outperforms the meta-learning based counter-part [2], and the single image transductive inference baseline [1] on two benchmarks.

We make the following contributions. (i) We propose a novel method for temporal transductive inference that enforces global and local temporal consistency to improve the segmentation accuracy with few labelled examples in 1-way setup (i.e. novel class against background). (ii) We propose a novel cross-domain few-shot video object segmentation setup, which explores a more realistic

Machine Learning for Autonomous Driving Workshop at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

scenario where both training data and few-shot support set are not necessarily drawn from the same data distribution as the query videos. It provides a more challenging benchmark as shown in Figure 1.

2 Temporal transductive inference (TTI)

We propose two main regularizers on finetuning that leverage temporal relations present in the query set. Our single image baseline [1] fixes the feature extraction weights and only learns linear classifier weights through initialization with support set prototypes [6, 9], followed by finetuning on the support set using cross entropy loss, $\mathcal{L}_{ce}$. Additionally, it minimizes the entropy, $\mathcal{L}_{\mathcal{H}}$, along with a KL-divergence loss, $\mathcal{L}_{KL}$, on the foreground/background(fg/bg) region proportion, for the query predictions to avoid degenerate solutions [1]. A naive approach for TTI uses one set of weights per video. However, the regularization from $\mathcal{L}_{KL}$ leads to degraded results when summed over all the query frames with different priors. Thus, we employ global and local regularizers, while maintaining separate prototypes per query frame.

Global temporal consistency. We first describe the global cue we use. For a video v with N_v frames, we compute on a video level global prototype, $\Omega_v^l = \frac{1}{N_v}\sum_{t=1}^{N_v}\omega_t^l$, where ω_t^l are the weights learned for frame t at finetuning iteration l. The we compute query signatures in a transductive manner using the predictions p^l from iteration l and query features $F^{(q)}$ as

$$z_{fg}^{l\ (q)} = \frac{\sum_{x,y} p_{fg}^{l\ (q)}(x,y)F^{(q)}(x,y)}{\sum_{x,y} p_{fg}^{l\ (q)}(x,y)}, z_{bg}^{l\ (q)} = \frac{\sum_{x,y} p_{bg}^{l\ (q)}(x,y)F^{(q)}(x,y)}{\sum_{x,y} p_{bg}^{l\ (q)}(x,y)}, \tag{1}$$

where foreground probability, $p_{fg}^l(x,y) = \sigma^l(x,y)$, with $\sigma^l = \textbf{sigmoid}(\tau(\langle F(x,y), \omega^l\rangle - b^l))$, is based on cosine similarity between weights, w^l, and query features, $F^{(q)}$, while background probability is simply $p_{bg}^l(x,y) = 1 - p_{fg}^l(x,y)$. Then a temporal regularization according to

$$\mathcal{L}_{global} = \frac{1}{N_v}\sum_{t=1}^{N_v}(1 - \langle\Omega_v^l, z_{fg}^{l\ (q)}\rangle) + \frac{1}{N_v}\sum_{t=1}^{N_v}\max(0, \langle\Omega_v^l, z_{bg}^{l\ (q)}\rangle)$$

is applied, where $\langle .,. \rangle$ denotes cosine similarity. The global loss leads to maximizing the similarity between the foreground query signature, $z_{fg}^{l\ (q)}$, and global prototype, Ω_v^l, while pushing it further away from background query signature, $z_{bg}^{l\ (q)}$. In every optimization step we recompute the global prototype and the query signatures. Thus, our method simultaneously regularizes the foreground/background (fg/bg) region proportion, while enforcing global consistency of the weights over all frames.

Local temporal coherence. Based on the assumption that objects undergo gradual scale changes temporally, we can expect the same for the fg/bg region proportion, i.e. it would change gradually within a local temporal window. Hence, we use the minimization of the rate of change for the fg/bg region proportion from consecutive frames to enforce local temporal coherence according to

$$\mathcal{L}_{local} = \sum_{t=1}^{N_v}\sum_{i=1}^{N_w}|P_t^l - P_{t+i}^l|, \tag{2}$$

where P_t^l is the fg/bg region proportion computed similar to [1] for frame t at optimization iteration l, and N_w is a local temporal window. We then use this loss, (2), as a local temporal regularizer to avoid erroneous predictions stemming from a wrongly estimated prior label marginal distribution. The final loss used for our proposed TTI is

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1\mathcal{L}_{\mathcal{H}} + \lambda_2\mathcal{L}_{KL} + \lambda_3(\mathcal{L}_{global} + \mathcal{L}_{local}), \tag{3}$$

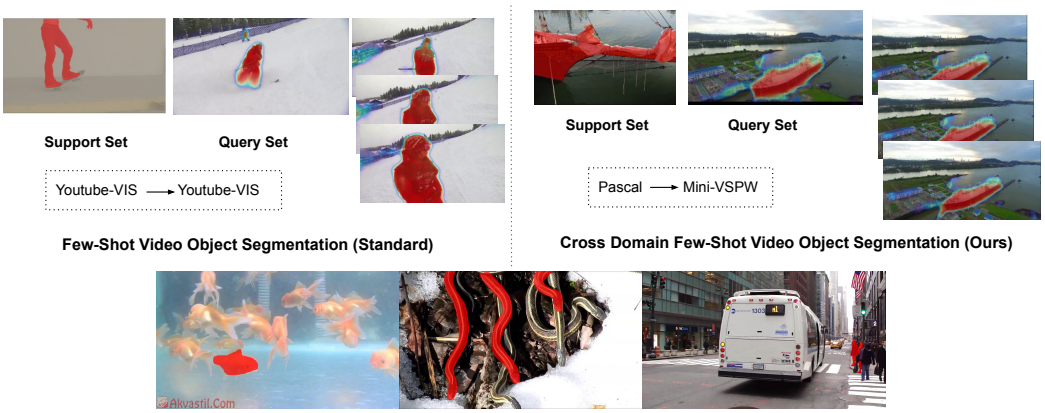


Figure 1: Shortcomings in YouTube-VIS. Top: No distribution shift between train and test dataset versus our proposed cross-domain setup. Bottom: non-exhaustive labels in the annotations.

with λ_i empirically selected weights. Finally, we found it useful after convergence to propagate an automatically selected key-frame prediction to the rest of the frames as a second stage finetuning. The key-frame is selected based on the maximum cosine similarity between query foreground signature and the global prototype.

3 Cross-domain FS-VOS benchmark

The previously proposed FS-VOS setup on YouTube-VIS [2] has two main shortcomings. First, YouTube-VIS is not exhaustively labelled, i.e. not all object occurrences in the sequence are labelled as shown in Figure 1. Second, it assumes that the training and test datasets are drawn from the same data distribution. This assumption does not resemble a realistic scenario where a domain shift occurs between the training and testing datasets. We focus on overcoming these shortcomings in our proposed setup. We evaluate cross-domain to simulate the domain shift that might occur between training and test data distribution in real-life applications. Thus, we propose the PASCAL-to-MiniVSPW setup, where PASCAL-VOC 2012 [3] is used as the training dataset, while Mini-VSPW is the test. This dataset has 128 sequences with 8890 frames exhausatively labelled. We manually map classes between PASCAL and VSPW and go through videos to filter out annotations not coinciding with PASCAL, since some of the VSPW classes can cause ambiguity such as "bottle or cup". Then, we follow the same Pascal-5^i splits [7] to have four splits dividing the novel classes.

4 Experimental results

Experimental protocol. We follow [2][1] and report the mean intersection over union (mIoU); we then propose to use the video consistency, VC_W, following [5], as a better metric on MiniVSPW, with small window, $W = 3$ due to the difficulty of our task. The evaluation is performed over five runs and reporting the average, in every run we sample 1000 tasks in the cross-domain setup following [1], while in YouTube-VIS we follow [2]. For the sake of fair comparison between our method and our baseline [1] on YouTube-VIS, we ensure running the two techniques using the same tasks to avoid differences from randomly sampled tasks. We follow the same architectural choices and hyperparameters as [1], and set $\lambda_3 = \lambda_2$, but choose ResNet-50 backbone for fair comparison with [2]. We train the base network on the base classes for the fold in a standard training with cross entropy with 100 epochs on PASCAL-VOC and YouTube-VIS. We use SGD with a learning rate of 2.5×10^{-3}, momentum of 0.9, weight decay of 1×10^{-4} and cosine learning rate decay. We also found it useful for Youtube-VIS to use an auxiliary dense contrastive loss [11] during the base training between temporally sampled and randomly augmented frames.

Comparison to state-of-the-art. Table 1 compares our approach with respect to the previous state of the art in FS-VOS. Our approach outperforms the recent state of the art meta-learning approach [2], which uses temporal information, by 2.8%. More importantly, we outperform the gain of using temporal information with respect to the single image baseline, where [2] outperforms the single image baselines by 0.9%, while we outperform our baseline, by 1.3%. This result confirms the benefit of using temporal information in the transductive inference scheme. In Table 2 we show the 1-shot

Method	mIoU							
	Query	Meta-Learning	1	2	3	4	Mean	Δ
PMMs [12]	Image	✓	32.9	61.1	56.8	55.9	51.7	-
PPNet [4]	Image	✓	45.5	63.8	60.4	58.9	57.1	-
DANet [2]	Video	✓	43.2	65.0	62.0	61.8	58.0	0.9
RePRI [1]	Image	✗	45.8	68.6	59.3	**64.2**	59.5	2.4
Naive Temporal RePRI	Video	✗	36.6	62.0	50.2	55.2	51.0	-
TTI (ours)	Video	✗	**48.2**	**69.0**	**62.8**	63.1	**60.8**	**3.7**

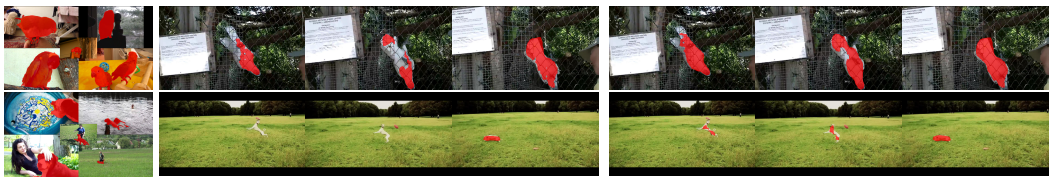
Table 1: Results on YouTube-VIS with ResNet-50 backbone 5-shot.

Method	mIoU					VC_3				
	1	2	3	4	Mean	1	2	3	4	Mean
RepRI [1]	34.8	47.5	32.0	24.2	34.6	15.9	**17.0**	19.7	14.1	16.7
TTI (ours)	**36.4**	**48.2**	**34.5**	**24.5**	**35.9**	**17.1**	16.4	**23.4**	**14.2**	**17.8**

Table 2: Pascal-to-MiniVSPW Results with ResNet-50 Backbone 1-shot.

Method	Youtube-VIS					Pascal-to-MiniVSPW				
	1	2	3	4	Mean	1	2	3	4	Mean
RePRI [1]	45.8	68.6	59.3	**64.2**	59.5	34.8	47.5	32.0	24.2	34.6
G	46.4	68.9	59.8	63.8	59.7	36.2	47.4	33.7	24.4	35.4
G + L	46.6	69.2	60.0	64.1	60.0	**36.6**	47.5	34.0	24.4	35.6
G + L + K	47.5	**69.5**	60.5	63.8	60.3	36.4	**48.2**	**34.5**	**24.5**	**35.9**
G + L + K + CL	**48.2**	69.0	**62.8**	63.1	**60.8**	-	-	-	-	-

Table 3: Ablation study, where G: Global, L: local, K: Keyframe propagation, CL: auxiliary contrastive loss between temporally sampled frames only applicable to training on Youtube-VIS.



(a) (b) (c)

Figure 2: Qualitative analysis on Youtube-VIS showing temporal stability of our approach w.r.t single image baseline. (a) 5-shot support set. (b) RePRI [1]. (c) TTI (ours).

results of our method with respect to the baseline on Pascal-to-MiniVSPW in a cross-domain setup. Our method outperforms the single image baseline in both mean intersection over union (mIoU) and video consistency, VC_3 metrics. Finally, we show qualitative results in Figure 2.

Ablation study. Table 3 shows the improvements from combining global and local constraints on both Youtube-VIS and Pascal-to-MiniVSPW. The different components we propose contribute to the final improvement of our model with respect to the baseline on both benchmarks, with the most significant gain from global consistency. Notably, our method attains ≈1.3% improvement with respect to the single image baseline without finetuning of the backbone features [2]. Thus, we achieve greater improvement by learning the linear classifier weights with the correct regularization running at 1.2 seconds per video, without the computational inefficiency of backbone finetuning.

5 Conclusion

We have presented a novel temporal transductive inference that uses global and local constraints to improve accuracy of few-shot video segmentation. These constraints are enforced as losses during learning, with the global addressing prototype consistency across a video, while the local addresses local region proportion consistency. Our approach outperforms recent state of the art methods on two different benchmarks.

References

[1] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13979–13988, 2021.

[2] Haoxin Chen, Hanjie Wu, Nanxuan Zhao, Sucheng Ren, and Shengfeng He. Delving deep into many-to-many attention for few-shot video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14040–14049, 2021.

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[4] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 142–158. Springer, 2020.

[5] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. VSPW: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4133–4143, 2021.

[6] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5822–5830, 2018.

[7] Amirreza Shaban, Zhen Bansal, Shrayand Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 167.1–167.13. BMVA Press, September 2017.

[8] Mennatullah Siam, Naren Doraiswamy, Boris N. Oreshkin, Hengshuai Yao, and Martin Jagersand. Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 860–867. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[9] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. AMP: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5249–5258, 2019.

[10] Wenguan Wang, Tianfei Zhou, Fatih Porikli, David Crandall, and Luc Van Gool. A survey on deep learning technique for video segmentation. *arXiv preprint arXiv:2107.01153*, 2021.

[11] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[12] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 763–778. Springer, 2020.