Supplementary to Rethinking Self-supervised Correspondence Learning: A Video Frame-level Similarity Perspective

Jiarui Xu Xiaolong Wang
UC San Diego

1. Implementation Details

Fine-grained Correspondence We apply recurrent inference strategy for fine-grained correspondence. To be more specific, we calculate the similarity between the current frame with the first frame ground truth labels as well the prediction results in the preceding m frames. Then the labels of top-k most similarly pixels are selected and propagated to the current frame. We only compute the similarity between features that are at most r pixels away from each other, i.e. *local* attention. The detailed hyperparamter setting for each dataset are listed in Table 1

	DAVIS	VIP	JHMDB
top-k	10	10	10
preceding frame m	20	8	4
propagation radius r	12,18	20	20

Table 1. Fine-grained Correspondence Inference Hyperparameter. On DAVIS, ResNet-18 and ResNet-50 models set $r = 12$ and $r = 18$ respectively.

Object-level Correspondence For the fair comparison, we use fine-tuning setting when comparing with previous approaches [1, 7]. Specifically, an additional 1×1 convolution is placed on top of the backbone to transform the frozen representation. Note that only this 1×1 convolution is learnable during fine-tuning. So such protocol could be considered as the linear evaluation. We fine-tune the the 1×1 convolution layer on the GOT-10K [2] dataset, which consists of $\sim$ 10,000 video clips and 1.4 million frames. Adam optimizer is adopted during fine-tuning. The learning rate is initialized to 0.001 and decays by 0.9 every epoch. There is no weight decay. The network is fine-tuned for 50 epochs. The batch size is 8 for all experiments. The inference hyperparamters are the same for without fine-tuning and with fine-tuning setting.

Evaluation Metrics The definitions of the metrics are as followed.

- $\mathcal{J}$ for video segmentation: It measures the region based segmentation similarity. Given an output segmentation M and the corresponding ground-truth mask G, $\mathcal{J}$ is defined as $\frac{M \cap G}{M \cup G}$.
- $\mathcal{F}$ for video segmentation: It evaluate the segmentation contour accuracy. Let P_c and R_c be the precision and recall between the contour points of M and G. $\mathcal{F}$ is defined as $\frac{2P_c R_c}{P_c + R_c}$.
- **Precision** for object tracking: Precision measures the percentage of frames where the (normalized) center error is less than a certain threshold, using the area-under-the-curve (AUC) evaluation.
- **Success** for object tracking: Success measures the percentage of frames where the IoU (Intersection over Union) is more than a certain threshold, using the AUC evaluation.

2. Visualization

Without fine-tuning on any additional dataset, the fine-grained correspondence are directly evaluated on the res_4 features of pre-trained ResNet. We visualize our correspondence on 3 downstream tasks and datasets in Figure 2,4,3, i.e. video object segmentation on DAVIS-2017 [5], human pose tracking on JHMDB [4], and human part tracking on VIP [8]. For DAVIS and VIP, there are usually more than one instances/parts. Our approach could output tight boundaries around the multiple target areas. For example, in the last row of Figure 3, the human parts could still be segmented when more people appears in the video. In human pose tracking, even though each joint is propagated individually, we could still estimate the pose accurately. We also compare our VFS with state-of-the-art method [3] in Figure 1. As last three rows illustrated, our VFS has less false positive object segmentation than [3]. It indicates that our VFS is more robust to distinguish similar pixels. Note that the inference hyperparamters for both methods are the same, the only difference is the pre-trained representation weight.

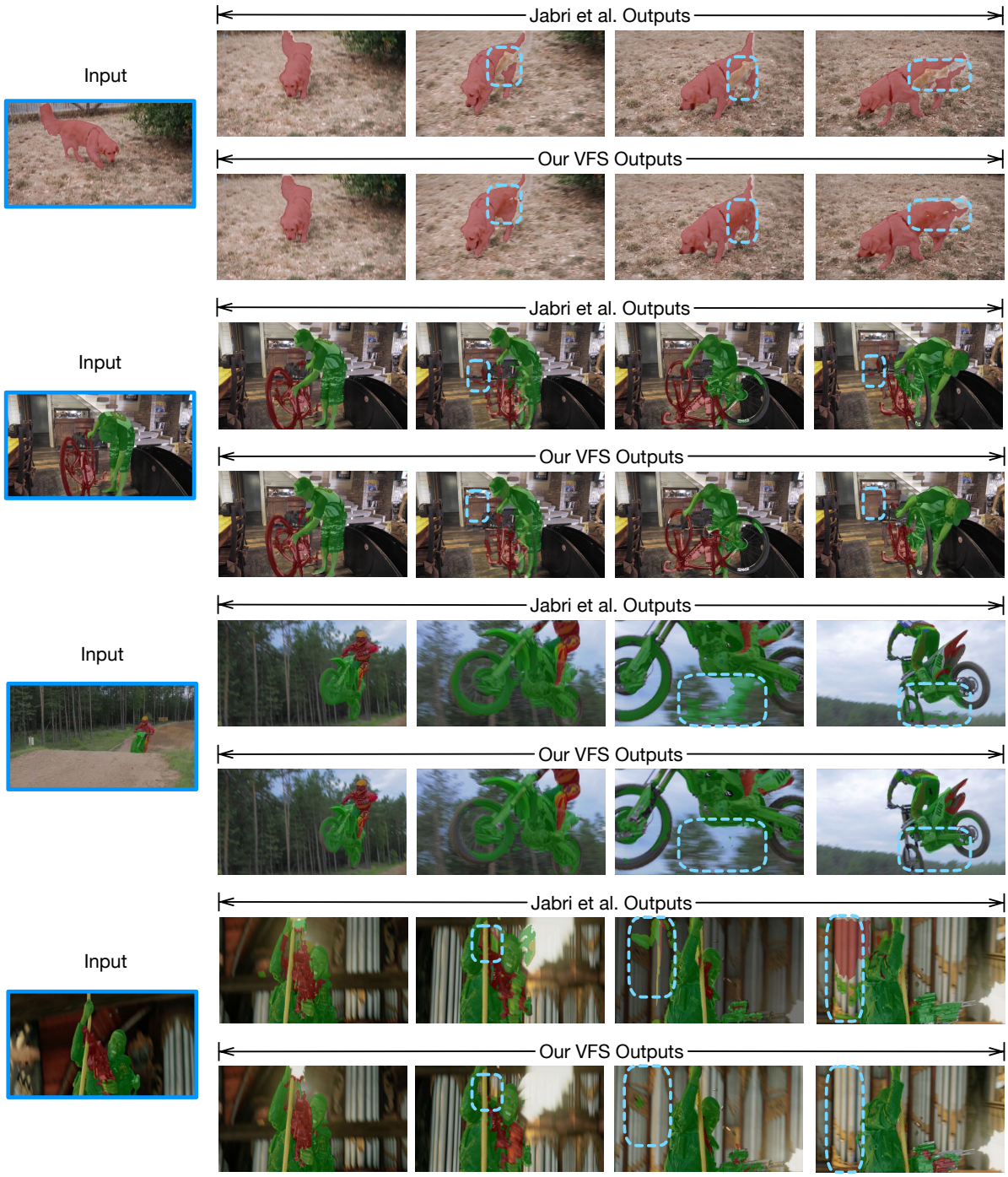

Figure 1. **Compare Fine-grained Correspondence on DAVIS**. Comparing with previous state-of-the-art Jabri et al. [3], our VFS could generate results of higher quality and with less false positives. Blue dash areas indicate failure cases in [3], where our approach could output plausible results. More comparison are provided in the project page.

We use fine-tuned res_5 features for object-level correspondence on OTB-100 visual object tracking [6]. The results are visualized in Figure 5. Our VFS could robustly track the target object even under difficult scenarios. For example, in the first row, there are multiple similar basketball players, and tracking target undergoes complicate object interaction as well as occlusion. Similarly for the deer in the third row, where the tracking target overtakes other similar deers. For the jumping person in the last row, the video suffers motion blur and large camera displacement.

We provide more visualization in our project page .

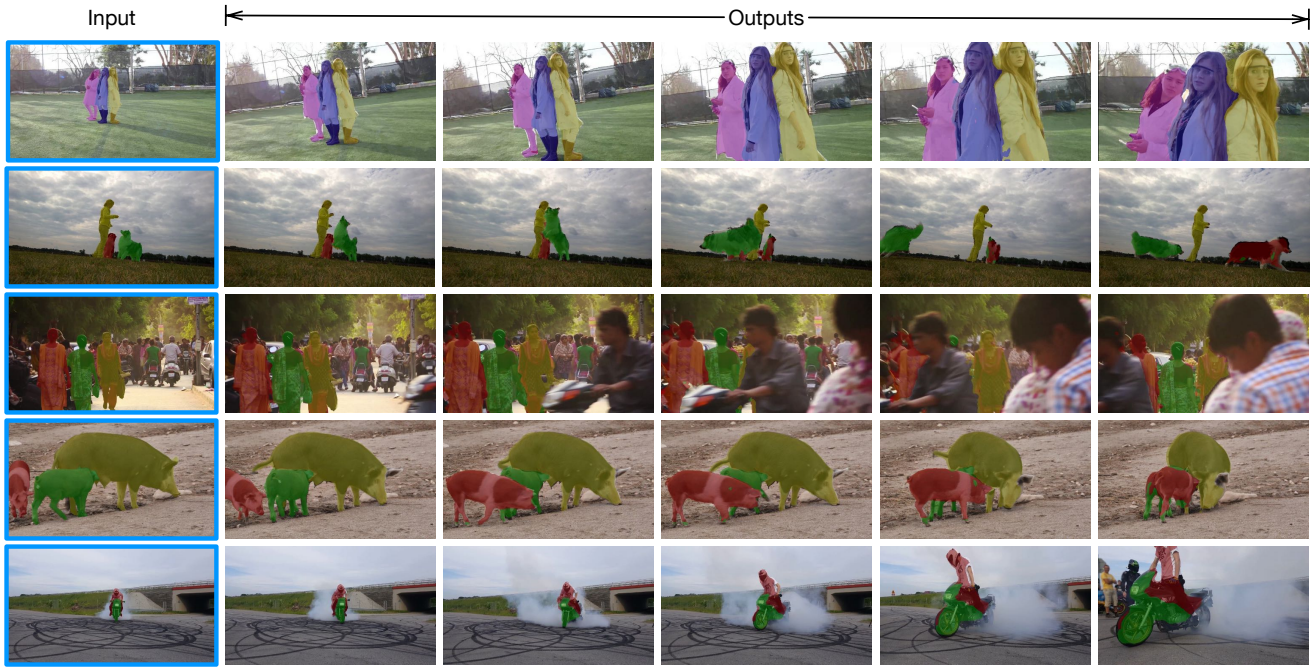

Figure 2. Qualitative Results for video object segmentation on DAVIS-2017 [5].

Figure 3. Qualitative Results for human part tracking on VIP [8].

References

[1] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos, 2020. 1

[2] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[3] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. In *Advances in Neural Information Processing Systems*, pages 19545–19560, 2020. 1, 2

[4] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid,

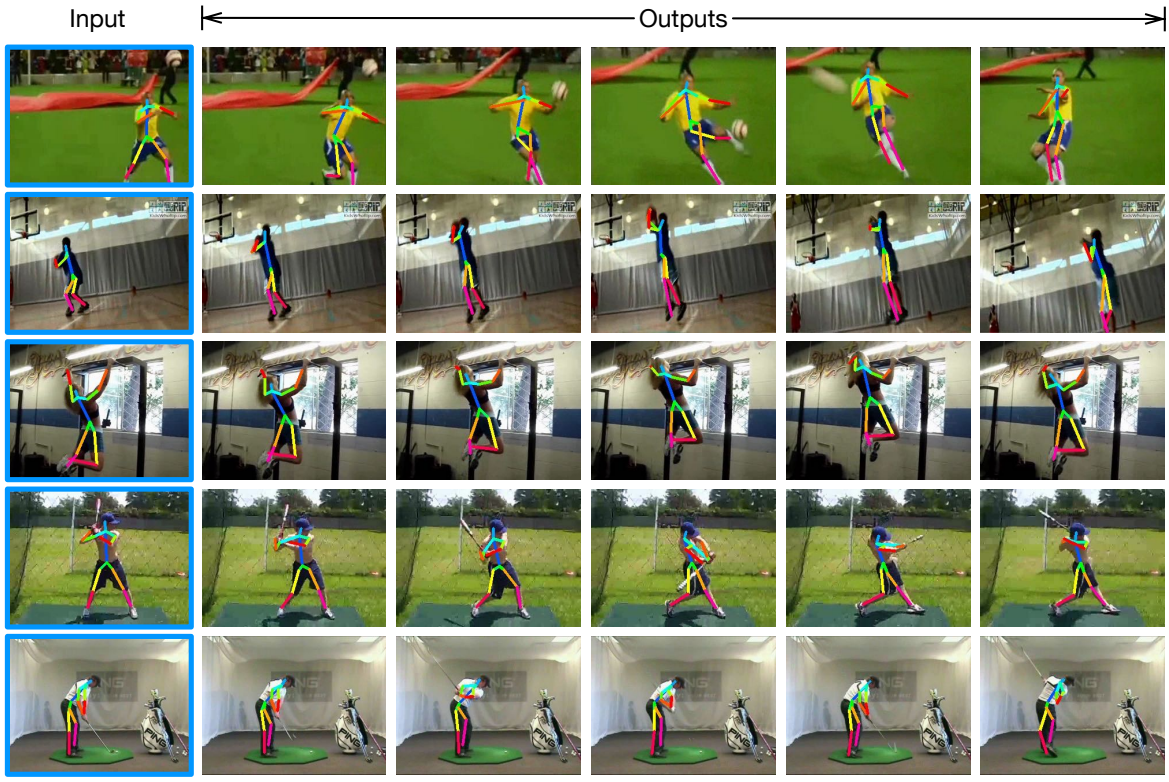

Figure 4. Qualitative Results for human pose tracking on JHMDB [4].

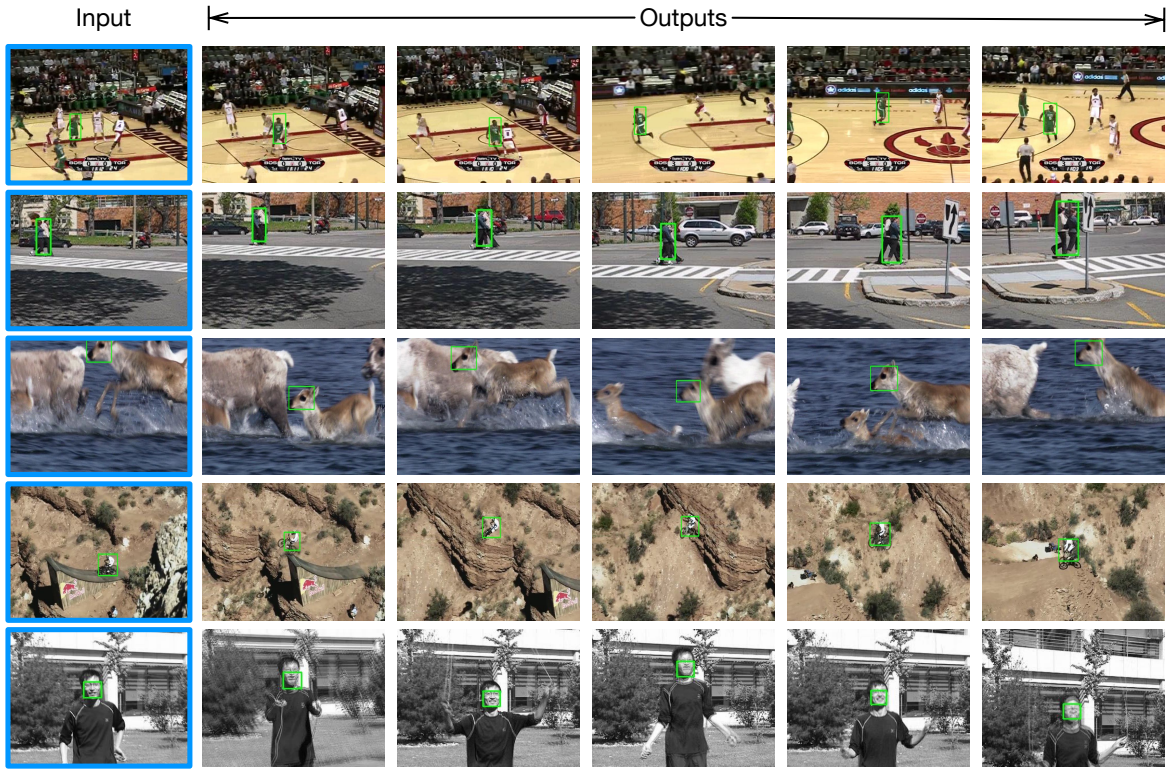

Figure 5. Qualitative Results for visual object tracking on OTB-100 [6].

and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 1, 4

[5] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 1, 3

[6] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *TPAMI*, 2015. 2, 4

[7] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. In *35th AAAI Conference on Artificial Intelligence*, 2021. 1

[8] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1527–1535, 2018. 1, 3