Mining Better Samples for Contrastive Learning of Temporal Correspondence -Supplementary Materials-

Sangryul Jeon[1], Dongbo Min[2,*], Seungryong Kim[3], Kwanghoon Sohn[1,*]
[1]Yonsei University, [2]Ewha Womans University, [3]Korea University
{cheonjsr,khsohn}@yonsei.ac.kr
dbmin@ewha.ac.kr, seungryong_kim@korea.ac.kr

Here we describe more details on the implementation of our system in Sec. 1, and more ablation studies in Sec. 2. Also, additional qualitative results of our method are provided in a form of *video* on the validation set of DAVIS2017 [9], Youtube-VOS 2018 [11], VIP [12], and JHMDB dataset [4].

1. Implementation Details

Network architecture As summarized in Tab. 1, we adopt ResNet-18 [2] network architecture as our backbone, reducing the stride of convolutional layers to produce the increased spatial resolution of the output by a factor of four (*i.e.* downsampling factor of $1/8$). For instance, given the input image of 256×256 spatial resolution during training, the resulting feature maps have a size of 32×32.

Label propagation algorithm Due to the rapid progress in this research line, the label propagation algorithm of the state-of-the-art methods [10, 7, 6, 3] is not standardized. For a fair comparison, we simply follow the same label propagation algorithm of the best approach for each evaluation task; the algorithm of [3] for the evaluation on DAVIS2017 dataset [9], the one of [6] for Youtube-VOS 2018 [11] dataset, and the one of [10, 7] for JHMDB [4] and VIP dataset [12].

Specifically, in [10, 7], they propagate the given annotation at the first frame by utilizing additional spatial and temporal context in video. The spatial context is aggregated in nearest neighbor search scheme, considering top-k matching probabilities per each pixel. To provide temporal context, the predictions from the first frame to the last preceding n frames onto the target frame are considered, averaging all $n+1$ predictions to obtain the final propagated label. They set the hyperparameter of temporal context n to 1 for the evaluation on VIP dataset [12] and 7 for JHMDB dataset [4], and the one of spatial context k to 5 for both datasets.

*Co-corresponding author

Layer	Output	Configuration
input	$H \times W$	
conv1	$H/2 \times W/2$	$[7\times7, 64] \times 1$
maxpool	$H/4 \times W/4$	3×3
conv2	$H/4 \times W/4$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$
conv3	$H/8 \times W/8$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$
conv4	$H/8 \times W/8$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$
conv5	$H/8 \times W/8$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$

Table 1. Network architecture of our embedding networks. We modified ResNet-18 [2] architecture to increase the resolution of the output with the downsampling factor of $1/8$.

In [3], they additionally leverage temporal coherence constraint to restrict the set of context for each target pixel to a spatial neighborhood of the query pixel with radius r. Another difference is that they select top-k matching probabilities over all the set of temporal context instead of independently selecting the top-k from each frame. For the evaluation on DAVIS2017 dataset [9], they set the hyperparameters of $\{k, n, r\}$ to $\{10, 20, 12\}$.

Similarly, in [6], they utilized the spatial and temporal context for label propagation restricting the spatial attention to the local neighborhood. When compared to [3], the differences are three folds; 1) the predictions of all possible spatial context are aggregated in a form of softargmax [5], 2) the center of restricted local window is determined also via softargmax operator instead of the position of a query pixel, and 3) the temporal context n are set to 5 with slightly different frame indices such that $\{I^0, I^5, I^{t-5}, I^{t-3}, I^{t-1}\}$. To employ the label propagation algorithm of [6], we remove the maxpool layer of our network architecture to provide the same spatial resolution of the feature maps with respect to the one of [6].

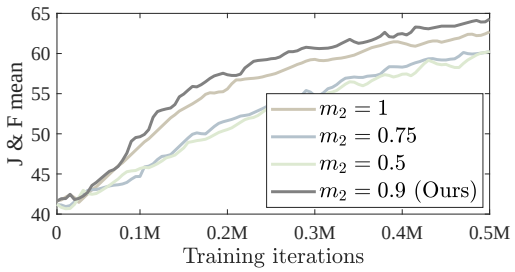

Figure 1. Convergence analysis of the performance with respect to various upper thresholds m_2. During the evaluations, the lower threshold m_1 is fixed to 0.

2. Ablation Study

We conduct additional series of ablation studies on the validation set of DAVIS-2017 dataset [9] to further examine the effects of our components.

Upper threshold m_2 To validate the effectiveness of our negative mining strategy, we additionaly examined the performance of our model varying with different upper thresholds m_2. During the evaluations, we fixed the lower threshold m_1 to 0. Fig. 1 shows that allowing very hard negative samples by setting upper threshold to 1 at the beginning of training results in lower accuracies. We can attribute this to the poor quality of representation in early training; using hard negatives can simply be too difficult for the current representation capability to discriminate.

Choice of confidence scores for planning curriculum We also report the performances with different choices of confidence scores for computing the variance in Tab. 2. We find that planning our curriculum based on the variance of C degrades the performance as imposing temporal coherence constraint to compute C may remove the information needed to evaluate the current representation power by discarding the confidence scores outside of the local window. The usage of Q and T yields roughly similar performances, but the sparsity of matrix T due to the optimal transport optimization allows us to reduce the computing time and memory.

Runtime analysis We measure the runtime of our components for a given pair of input images during training. The computation time required for sinkhorn algorithm [1] to solve optimal transport problem is 242 ms. We also report the runtime to compute the variances in Tab. 2. Note that using the sparsity of matrix T allows us to reduce the computing time and memory, compared to the ones when computing with matrix Q. The measurements are performed on an Intel Core i7-10700k CPU with two NVIDIA TITAN RTX GPUs.

var(Q)	var(T)	var(C)	$\mathcal{J}\&\mathcal{F}_{\text{mean}}$	Runtime (ms)
✓	-	-	70.0	38.1
-	-	✓	65.8	**16.5**
-	✓	-	**70.3**	17.3

Table 2. Ablation study on DAVIS-2017 validation set for different confidence scores to compute variance in Equ.(9) of the main paper. We also report the averaged runtime required to compute the variance.

References

[1] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 2

[2] 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. 1

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

[4] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 1

[5] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 1

[6] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2020. 1

[7] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *Advances in Neural Information Processing Systems*, pages 318–328, 2019. 1

[8] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[9] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 1, 2

[10] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 1

[11] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1

[12] 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