Spatial-then-Temporal Self-Supervised Learning for Video Correspondence —— CVPR 2023 Supplementary Material

Rui Li Dong Liu
University of Science and Technology of China, Hefei, China
liruid@mail.ustc.edu.cn, dongeliu@ustc.edu.cn

The supplementary material contains: 1) implementation details; 2) visualization of entropy-based selection; 3) more qualitative examples for video object segmentation.

1. Implementation Details

More training details. We include more training details for the second step of training in our spatial-then-temporal feature learning. When training on YouTube-VOS [8] with $\mathcal{L}_t$, we compute the local correlation map by setting the temperature τ to 1. Then the local correlation distillation loss $\mathcal{L}_{\text{lc}}$ is conducted between the local correlation maps computed at layer res_3 and res_4. During training, we find $\mathcal{L}^e_{\text{lc}}$ and $\mathcal{L}_{\text{gc}}$ have much smaller values than frame reconstruction loss $\mathcal{L}^p_{\text{rec}}$, which tends to have little impact on self-supervised training. Thus, we set the loss weight α/β to 1000/50 to ensure the scale consistency for each loss. The threshold T of entropy-based selection is set to 0.7.

Training details for Spatiotemporal. Besides, we also provide the training details for the training configuration of Spatiotemporal. To achieve the goal of simultaneously learning spatial and temporal features with both image and video data, we optimize the encoder ϕ with $\mathcal{L} = \mathcal{L}_{\text{nce}} + \gamma\mathcal{L}_{\text{t}}$, where the γ is set to 10 for the balance between spatial and temporal feature learning, and we set the learning rate to 0.001 with a cosine (half-period) learning rate schedule. We reduce the stride up to layer res_4 from 16 to 8 to increase the spatial resolution of feature maps by a factor of 2, and the final outputs of the encoder are further processed with global average pooling plus a MLP head to favor the image-level loss $\mathcal{L}_{nce}$. The training batch is constructed by sampling both the images/frames from ImageNet [3]/YouTube-VOS [8], sharing the same augmentations in [2]. More specifically, we ensure that each batch has the frames from the same video to favor reconstructive learning, and these frames will be also regarded as individual images to be utilized for contrastive learning.

Inference strategy of label propagation. All evaluation tasks can be considered as a label propagation process. Given only the label in the first frame, we apply a recurrent inference strategy here following prior works [1,4,7]. More specifically, we calculate the local correlation maps within a local range r between the target frame and the first frame ground truth labels as well as the predictions in the preceding t frames. Then we select the top-k most similar pixels to propagate to the current frame. For the encoder with a stride of 8, we set t/k to 20/10 for all tasks and r is 12, 5, and 16 for DAVIS [6], JHMDB [5], and VIP [9] tasks respectively.

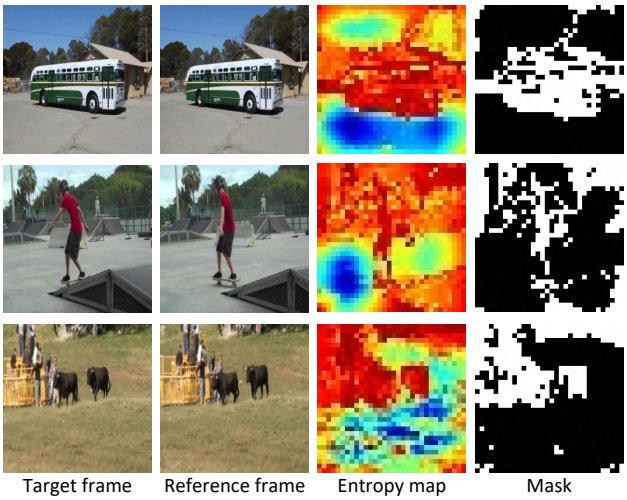

Figure 1. **Visualization of the entropy map.** We compute the entropy for each query in the target frame using Eq. (1). The mask with higher entropy is generated by setting a threshold T.

2. Qualitative examples for entropy-based selection

We give the visualization results of the entropy-based selection. After obtaining the local correlation map c between the target and reference frame, we calculate the entropy for each query i in the target frame:

$$\mathcal{H}(i) = \sum_j -\log c^{N-1}(i,j), i \in \{1,.,h^{N-1}w^{N-1}\}, j \in \mathcal{N}(i) \quad (1)$$

then we apply a min-max normalization on the entropy map. As shown in Figure 1, the entropy map has a higher response on moving objects involved in severe deformation

and occlusions, while the background tends to have lower entropy, which inspires us to filter out the regions with inadequate information for training.

3. More qualitative examples for video object segmentation

We enclose several representative videos on DAVIS-2017 [6] to verify the effectiveness of our method compared with state-of-the-art methods which provide the code and pre-trained models, e.g., CRW [4], VFS [7], and DUL [1]. We name it "demo.mp4" in our supplementary material. Without fine-tuning our pre-trained model on any additional dataset, we propagate the annotation of the first frame to the current frame. As observed in the enclosed video, the segmentation results produced by our method show clear improvements over state-of-the-art methods. The predictions of our method tend to have tight boundaries around the multiple objects even facing severe temporal discontinuity, e.g., appearance changes, large motion, and occlusions. These examples again verify the effectiveness of our two-step task and loss functions for learning better spatiotemporal representations.

References

[1] Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Dense unsupervised learning for video segmentation. In *NeurIPS*, pages 25308–25319, 2021. 1, 2

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[4] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. In *NeurIPS*, pages 19545–19560, 2020. 1, 2

[5] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013. 1

[6] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 2

[7] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *ICCV*, pages 10075–10085, 2021. 1, 2

[8] 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. 1

[9] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *ACM MM*, pages 1527–1535, 2018. 1