Supplemental Material: Locality-Aware Inter-and Intra-Video Reconstruction for Self-Supervised Correspondence Learning

Liulei Li[1,6*], Tianfei Zhou[2], Wenguan Wang[3†], Lu Yang[4], Jianwu Li[1] , Yi Yang[5]

[1] Beijing Institute of Technology [2] ETH Zurich [3] ReLER, AAII, University of Technology Sydney

[4] Beijing University of Posts and Telecommunications [5] CCAI, Zhejiang University [6] Baidu Research

https://github.com/0liliulei/LIIR

In this document, we first present more ablative experimental results for the spatial compactness prior (§1), long-term dependence (§2) and the number of feature point sampling (§3). Then, we show visual effects of the essential components (§4) for more in-depth analysis. We further offer additional qualitative video propagation results in §5. Finally, we discuss the limitations (§6) and potential societal impacts (§7) of LIIR.

1. Analysis of Spatial Compactness Prior

We use M 2D Gaussian distributions to approximate the affinity matrix A between two frames. Table 1 provides a detailed analysis of the hyper-parameter M. The 1st row corresponds to a baseline model that disregards the spatial compactness prior in both training and inference phases. We see from the table that **1)** when $M = 1$ in training, the model performs worse than the baseline model, as such a rigorous matching constraint easily leads to overconfident pre-dictions; **2)** when M (in training) becomes larger, the performance greatly improves; **3)** the models trained with $M = 2$ shows consistently better performance than those trained with $M = 3$; and **4)** in the inference stage, $M = 2$ always leads to the best performance. Accordingly, we set M to 2 in both training and inference stages.

#	training stage	inference stage	DAVIS $\mathcal{J}\&\mathcal{F}_m \uparrow$	VIP mIoU ↑
1	–	–	69.8	39.6
2	$M = 1$	–	69.4	39.3
3	$M = 1$	$M = 1$	69.0	39.1
4	$M = 1$	$M = 2$	69.6	39.4
5	$M = 1$	$M = 3$	69.6	39.4
6	$M = 2$	–	71.5	40.8
7	$M = 2$	$M = 1$	71.2	40.4
8	$M = 2$	$M = 2$	**72.1**	**41.2**
9	$M = 2$	$M = 3$	71.9	**41.2**
10	$M = 3$	–	71.3	40.7
11	$M = 3$	$M = 1$	71.1	40.4
12	$M = 3$	$M = 2$	71.7	40.9
13	$M = 3$	$M = 3$	71.6	40.8

Table 1. **Detailed analysis** of M in spatial compactness prior on DAVIS$_{17}$ [4] `val` and VIP [6] `val`. "–": without using spatial compactness prior. See §1 for details.

2. Analysis of Long-Term Dependence

LIIR leverages multiple reference frames during testing as in [2]. We analysis the impact of long-term dependence in Table 2. As seen, our method is more robust and shows smaller drop *wrt* [2]: 4.7% *vs* 6.8%.

Methods	Reference Frame	$\mathcal{J}\&\mathcal{F}_m$
Ours	~~I_0~~, ~~I_5~~, I_{t-5}, I_{t-3}, I_{t-1}	72.1 → 67.4 (-4.7)
MAST[2]	~~I_0~~, ~~I_5~~, I_{t-5}, I_{t-3}, I_{t-1}	65.5 → 58.7 (-6.8)

Table 2. Analysis of long-term dependence of reference frames on DAVIS$_{17}$ [4] `val`. See §2 for details..

3. Analysis of Feature Point Sampling

We give the ablative study on the number of feature points sampled during inter-video reconstruction in Table 3. It can be seen that, with 1440 frames, better results can be achieved if we sample more points per frame (70.4→71.4→72.1), supporting our claim about instance separation. With the same number of total sampled points, we gain better performance if we consider more frames (70.9→71.7→72.1). It is reasonable as more frames can provide much rich/challenging context. With our limited GPU capacity, we choose to sample 1440 frames and four feature points per frame, so as to maximize the performance. But we can speculate that, if with enough GPU capacity, sampling more features points from more videos will further improve the performance.

4. Visualization of Ablation Study

Fig. 1 depicts visual effect of each essential component in LIIR. Starting from the baseline model (b), we progres-

*Work done during an internship at Baidu Research.

†Corresponding author: *Wenguan Wang*.

#Frames	1440	1440	1440	960	480
#Feature Points Per-frame	4	2	1	6	12
$\mathcal{J}\&\mathcal{F}_m$	**72.1**	71.4 (-0.7)	70.4 (-1.7)	71.7 (-0.4)	70.9 (-1.2)

Table 3. Ablative study on the number of feature points sampled during inter-video reconstruction on DAVIS$_{17}$ [4] `val` (see §3).

sively add position encoding (c), inter-video reconstruction (d) and spatial compactness (e). As seen, with explicit positional encoding, our model is able to heavily suppress background regions (*e.g.*, shelves in the first row). The inter-video reconstruction enables more accurate discrimination between targets and semantically similar distractors (*e.g.*, "motorcycle" in the third row). Last, incorporating the spatial compactness prior facilitates more precise correspondences, leading to high-quality final segmentation results.

5. Additional Qualitative Results

We provide additional video propagation results on four datasets, including DAVIS$_{17}$ [4] `val` in Fig. 2, Youtube-VOS[5] `val` in Fig. 3, VIP[6] `val` in Fig. 4 and JHMDB[1] `val` in Fig. 5. We observe that even training with no annotations, LIIR is able to produce highly exquisite results.

6. Limitation

Although LIIR demonstrates remarkable performance and high generalizability in correspondence matching, we still see a large performance gap between LIIR and current top-leading supervised models (*e.g.*, STM [3] in VOS). However, as a self-supervised method, LIIR can be easily scaled to leverage any available collection of video data for training. This could lead to more accurate correspondence learning from massive unlabeled data instead of using small-scale datasets only (*e.g.*, DAVIS, YouTube-VOS). Apart from that, inter-video reconstruction spares massive space to bank the negative samples, this coerces the memory size of GPUs and extends the training time. Fortunately, the performance of LIIR has improved by leaps and bounds to compensate for it, for instance, up to **2.9%** on DAVIS$_{17}$ [4]. Further more, note that the inter-video reconstruction is only applied during training, thus it does not introduce extra computational cost at inference. In our future work, we will explore towards the above direction to narrow the performance gap between supervised methods, and find the way to reduce the memory space taken up by negative videos and speed up training simultaneously.

7. Broader Impact

The method, LIIR, described in this paper can potentially be harnessed to improve accuracy in any application of computer vision where establishing accurate temporal correspondence is crucial. Some applications, like patient monitoring in hospitals, elderly care, online meeting in the era of pandemic, are clearly beneficial to society. It can also contribute to commercial affairs such as autonomous driving, augmented reality and movie production. In principle, there is no ethical problem with the design purpose of LIIR.

References

[1] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, 2013. 2, 5

[2] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *CVPR*, 2020. 1

[3] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 2

[4] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 1, 2, 3

[5] 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 *ECCV*, 2018. 2, 4

[6] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *ACMMM*, 2018. 1, 2, 4

Figure 1. **Visual effects of essential components** in LIIR. (a) ground truth, (b) baseline, (c) b + position encoding, (d) c + inter-video reconstruction, and (e) d + spatial compactness. See §4 for details.

Figure 2. **More visualization results for video object segmentation** on DAVIS_{17} [4] `val`.

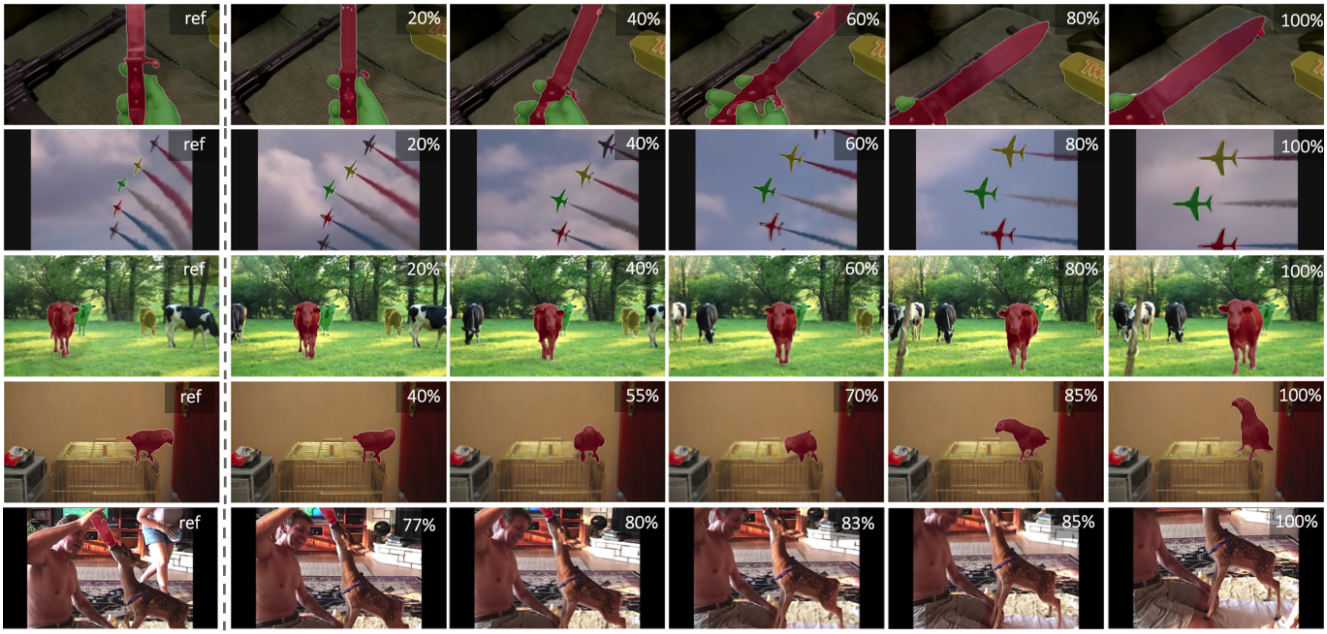

Figure 3. **More visualization results for video object segmentation** on Youtube-VOS [5] `val`.

Figure 4. **More visualization results for body part propagation** on VIP [6] `val`.

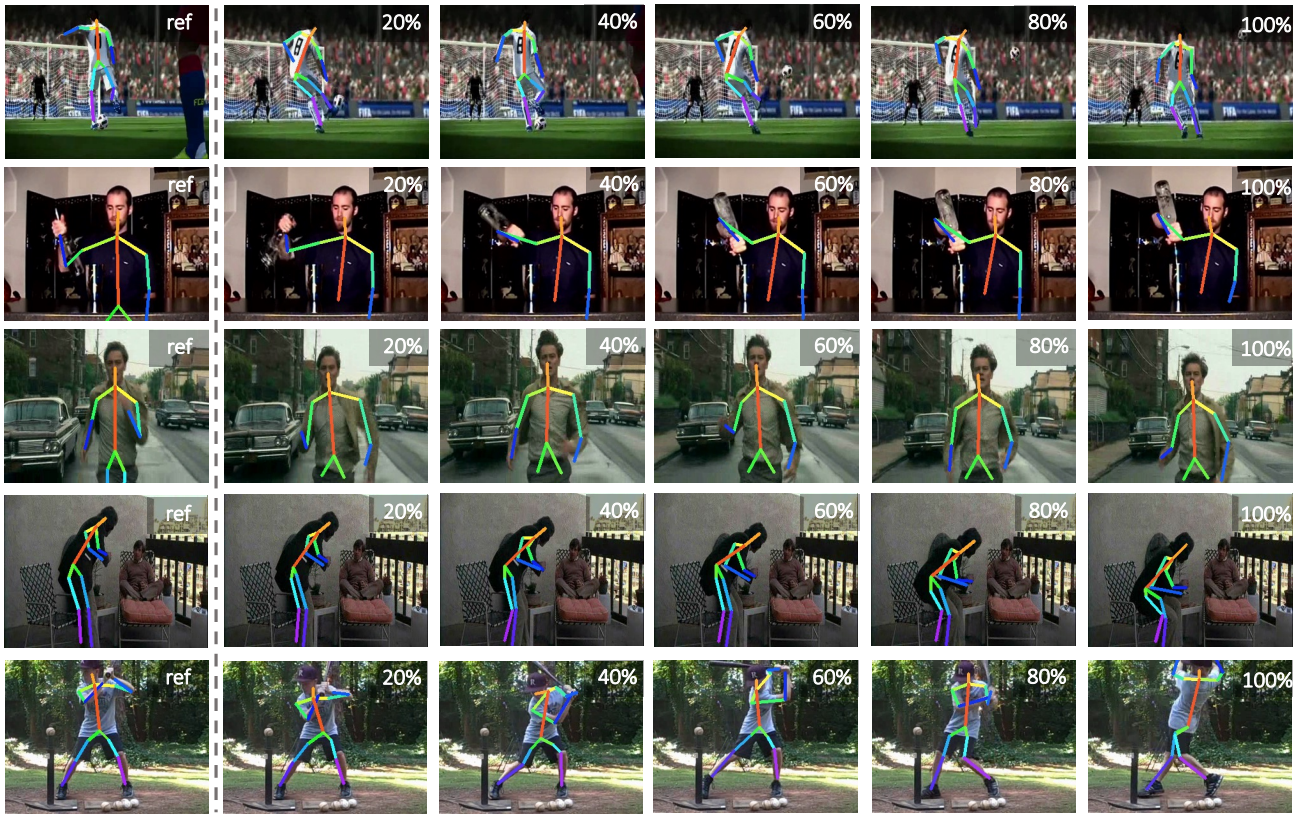

Figure 5. **More visualization results for keypoint propagation** on JHMDB [1] `val`.