## Local All-Pair Correspondence for Point Tracking

Seokju Cho<sup>1</sup>, Jiahui Huang<sup>2</sup>, Jisu Nam<sup>1</sup>, Honggyu An<sup>1</sup>, Seungryong Kim<sup>1,†</sup>, and Joon-Young Lee<sup>2,†</sup>

<sup>1</sup> Korea University <sup>2</sup> Adobe Research

## A More Implementation Details

For generating the Panning-MOVi-E dataset [2], we randomly add 10-20 static objects and 5-10 dynamic objects to each scene. The dataset comprises 10,000 videos, including a validation set of 250. For the sinusoidal position encoding function [7]  $\sigma(\cdot)$ , we use a channel size of 20 along with the original unnormalized coordinate. This results in a total of 21 channels. For all qualitative comparisons, we use LocoTrack-B model with a resolution of  $384 \times 512$ .

Details of the evaluation benchmark. We evaluate the precision of the predicted tracks using the TAP-Vid benchmark [1]. This benchmark comprises both real-world video datasets and synthetic video datasets. **TAP-Vid-Kinetics** includes 1,189 real-world videos from the Kinetics [4] dataset. As the videos are collected from YouTube, they often contain edits such as scene cuts, text, fade-ins or -outs, or captions. **TAP-Vid-DAVIS** comprises real-world videos from the DAVIS [6] dataset. This dataset includes 30 videos featuring various concepts of objects with deformations. **TAP-Vid-RGB-Stacking** consists of 50 synthetic videos [5]. These videos feature a robot arm stacking geometric shapes against a monotonic background, with the camera remaining static. In addition to the TAP-Vid benchmark, we also evaluate our model on the **RoboTAP** dataset [8], which comprises 265 real-world videos of robot arm manipulation.

Table 1: Convolutional layer configurations for different model sizes.

Model	Channel Sizes	Kernel Size	Strides
Small	$(64, 128) \\ (64, 128, 128)$	(5, 2)	(4, 2)
Base		(3, 3, 2)	(2, 2, 2)

**Detailed architecture of local 4D correlation encoder.** We stack blocks of convolutional layers, where each block consists of a 2D convolution, group normalization [9], and ReLU activation. See Table 1 for details. For the small model, we use an intermediate channel size of (64, 128) for each block. For the base model, the intermediate channel sizes are (64, 128, 128) for each block. For every instance of group normalization, we set the group size to 16.

 $<sup>^{\</sup>dagger}\mathrm{Co}\text{-}\mathrm{corresponding}$  authors.

2 S. Cho et al.

**Details of correlation visualization.** For the correlation visualization in Fig. 3 of the main text, we train a linear layer to project the correlation embedding  $E_t^k$  into a local 2D correlation with a shape of 7×7. This local 2D correlation then undergoes a softargmax operation to predict the error relative to the ground truth. We begin with the pre-trained model and train the linear layer for 20,000 iterations. For clarity, we bilinearly upsample the 7×7 correlation to 256×256.

## **B** More Qualitative Comparison

We provide more qualitative comparisons to recent state-of-the-art methods [2,3] in Fig. 1 and Fig. 2. Our model establishes accurate correspondences in homogeneous areas and on deforming objects, demonstrating robust occlusion handling even under severe occlusion conditions.



Fig. 1: Additional qualitative comparison with state-of-the-art [2,3].

4 S. Cho et al.



Fig. 2: Additional qualitative comparison with state-of-the-art [2,3].

## References

- Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: Tap-vid: A benchmark for tracking any point in a video. Advances in Neural Information Processing Systems 35, 13610–13626 (2022) 1
- Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: Tapir: Tracking any point with per-frame initialization and temporal refinement. arXiv preprint arXiv:2306.08637 (2023) 1, 2, 3, 4
- Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. arXiv preprint arXiv:2307.07635 (2023) 2, 3, 4
- Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 1
- Lee, A.X., Devin, C.M., Zhou, Y., Lampe, T., Bousmalis, K., Springenberg, J.T., Byravan, A., Abdolmaleki, A., Gileadi, N., Khosid, D., et al.: Beyond pick-andplace: Tackling robotic stacking of diverse shapes. In: 5th Annual Conference on Robot Learning (2021) 1
- Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017) 1
- Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems 33, 7537–7547 (2020) 1
- Vecerik, M., Doersch, C., Yang, Y., Davchev, T., Aytar, Y., Zhou, G., Hadsell, R., Agapito, L., Scholz, J.: Robotap: Tracking arbitrary points for few-shot visual imitation. arXiv preprint arXiv:2308.15975 (2023) 1
- Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018) 1