Supplementary Materials: Polygonal Point Set Tracking

Gunhee Nam[1*] Miran Heo[2] Seoung Wug Oh[3] Joon-Young Lee[3] Seon Joo Kim[2]

[1]Lunit Inc. [2]Yonsei University [3]Adobe Research

This supplementary material provides details of the proposed evaluation dataset, named as PoST, synthetic data, and more quantitative and qualitative results. In Section 1, we describe additional statistic details of PoST and visualize sample sequences. We introduced the method to synthesize data for data augmentation in the main paper. In Section 2, some samples of the synthetic data are provided. In Section 3 and Section 4, we provide more detailed quantitative and qualitative results and compare with other methods.

1. Details of PoST

For better analysis, we report statistical details of PoST in Table 1. The table provides six properties for each sequence: the number of ground-truth points, input size, sequence length, motion (MO), scale change (SC) and occluded points (OC).

We measure camera and object motion (MO) by computing the distance between corresponding points as follows:

$$\text{MO} = \frac{1}{TN} \sum_{t=1}^{T-1} \sum_{i=0}^{N-1} ||\mathbf{P}_t^i - \mathbf{P}_{t-1}^i||_2, \tag{1}$$

where $\mathbf{P}_t^i$ denotes the i^{th} $(0 \leq i < N)$ normalized ground-truth point of the t^{th} $(0 \leq t < T)$ frame and $||\cdot||_2$ refers to the Euclidean distance between points. Scale change (SC) is one of the challenging scenarios in the point tracking. We measure the changes by computing temporal changes of distance between adjacent points as follows:

$$\text{SC} = \frac{1}{TN} \sum_{t=1}^{T-1} \sum_{i=0}^{N-1} \left| ||\mathbf{P}_t^i - \mathbf{P}_t^{i-1}||_2 - ||\mathbf{P}_{t-1}^i - \mathbf{P}_{t-1}^{i-1}||_2 \right|. \tag{2}$$

Since we skip annotating the ambiguous corresponding points due to self-occlusion, we also report the average number of these points in each sequence (OC).

Some examples of PoST are shown in Figure 1. We sample seven random points and each corresponding point is represented by a unique index of a different color. The green outline in the first frame indicates the ground-truth contour of the target object provided in the dataset. The missing points in some frames imply the ambiguous points due to self-occlusion.

2. Samples of Synthesizing Data

Data augmentation by synthesizing image segmentation data from [2] is one of the important features to enhance the point tracking performance in our framework. For clarification of the synthetic data, some samples are shown in Figure 2. Note that the black areas are the result of random transformation to simulate the movement of the scene and the objects. Ground-truth corresponding point samples are marked in the same way as done in Figure 1, and the white-colored dense points are also added. Supervision for the point correspondence is provided by the synthetic data. Since we transform the original point set along with the image, the point correspondence is maintained over the entire frame.

3. Quantitative Result Analysis

In Figure 3, we evaluate our method and the other competitors [3, 7, 5, 6] using spatial and temporal accuracy (SA, TA) with various thresholds. Following Equation (9) in the main paper, the spatial accuracy directly measures the distance between the ground-truth and the predicted point while the temporal accuracy implies the shape tracking performance. For both metrics, our method achieves the best performance compared to the other competitors for all thresholds. Specifically, for the spatial accuracy, our method outperforms the other methods with significant margins even using a high threshold. Also, the difference in saturation tendency of the two metrics indicates that our approach exceeds the other methods, especially in terms of the point correspondence.

*Work mostly done during a M.S. student at Yonsei University

	bear	blackswan	boy	car-roundabout	car-shadow	cheetah	cup	drop	fish	freeway
# of points	10	18	18	15	12	21	12	6	10	6
Input size	640x400	1920x1080	384x540	1920x1080	1920x1080	1280x720	600x534	1280x718	720x480	762x506
Length	151	41	21	71	31	181	371	21	81	31
MO	0.049	0.035	0.037	0.084	0.051	0.048	0.068	0.138	0.060	0.049
SC	0.014	0.003	0.011	0.011	0.014	0.008	0.006	0.005	0.014	0.006
OC	0.000	0.000	0.000	0.875	0.000	1.368	0.684	0.000	0.000	0.000

	giraffe	helicopter	hiphop	labcoat	minion	monkey	monkey-head	monkey-horse	penguin	pig
# of points	11	17	11	16	17	10	13	20	11	8
Input size	658x484	1920x1080	960x540	3840x2160	540x575	480x270	960x540	960x540	384x212	695x480
Length	191	41	81	41	31	31	31	31	41	281
MO	0.026	0.157	0.159	0.249	0.029	0.060	0.026	0.039	0.021	0.114
SC	0.005	0.011	0.018	0.093	0.006	0.023	0.003	0.004	0.007	0.011
OC	0.100	1.000	0.778	4.600	0.000	0.000	0.500	0.000	0.000	0.517

	plane	pot	skater	slackline	soldier	station	sunset	tower	toy	worm
# of points	35	7	35	16	23	16	3	22	16	11
Input size	1488x914	960x540	720x480	1920x816	528x224	656x492	960x540	1280x720	960x540	480x264
Length	31	241	81	51	31	371	31	21	351	81
MO	0.036	0.037	0.235	0.567	0.339	0.238	0.017	0.002	0.029	0.045
SC	0.006	0.007	0.012	0.033	0.019	0.007	0.029	0.001	0.003	0.012
OC	0.000	0.360	11.111	5.667	2.500	2.526	0.000	0.000	0.167	0.000

Table 1: The statistical details of PoST. We report six properties for each sequence: the number of ground-truth points, input size, sequence length (*i.e*. the number of frames), motion (MO) including both of camera and object motion, scale change (SC) and self-occluded points per frame (OC).

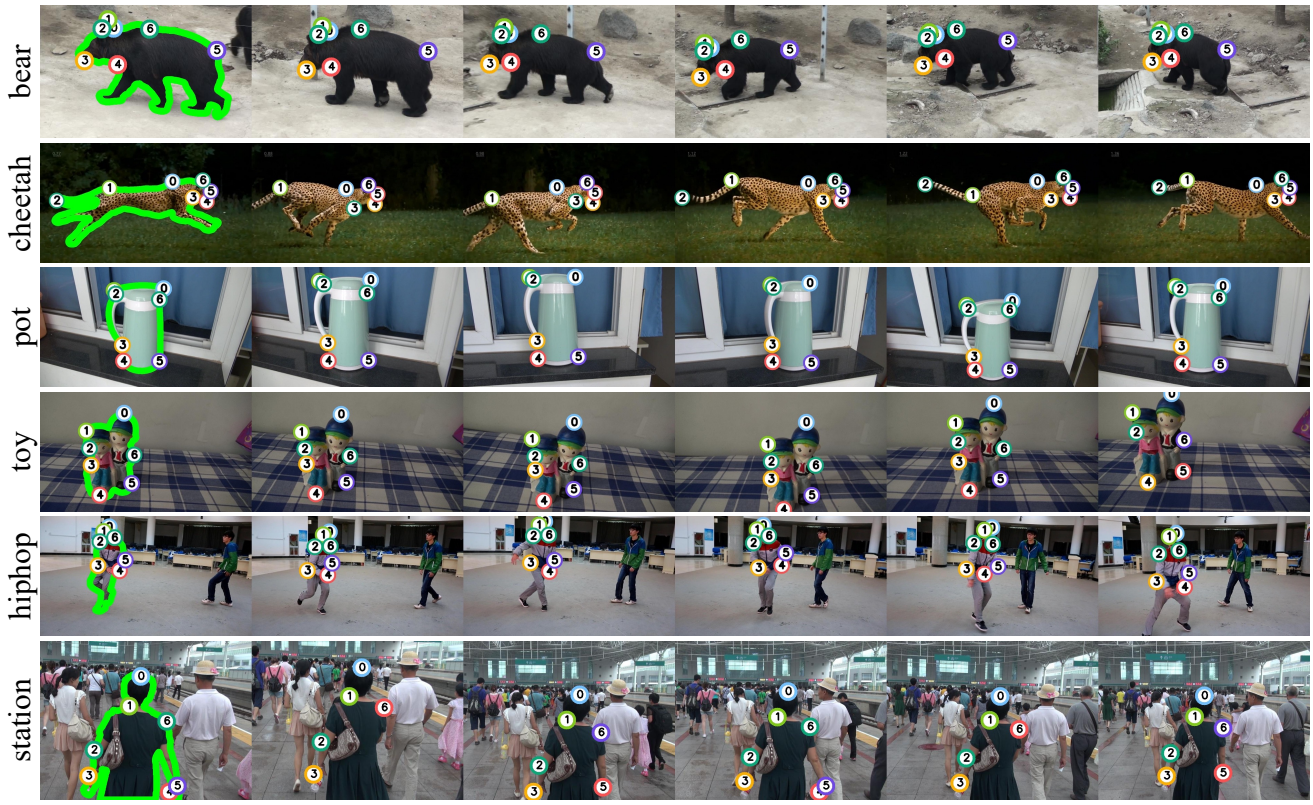


Figure 1: Examples of PoST evaluation dataset. Randomly sampled seven points are marked with the same colored corresponding points as the same indices.

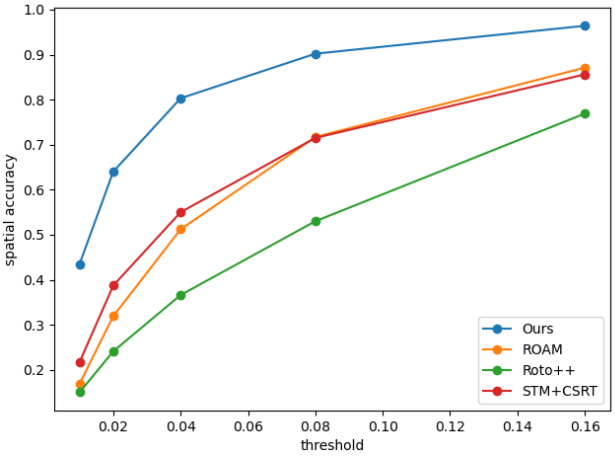
Figure 2: Example of synthetic dataset.

4. Qualitative Result Comparison

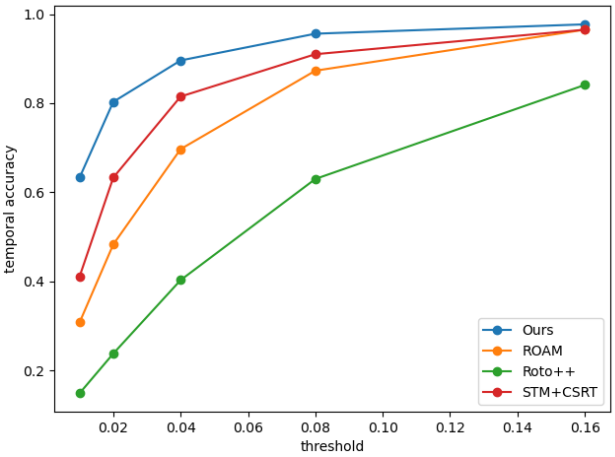
Figure 4 provides some qualitative results of our method and the other methods for comparison. We sample the sequences from the datasets of PoST (pot), DAVIS2016 [1] (blackswan, camel) and CPC [4] (plane). Visualization of the results is done in the same way as Figure 2. While Roto++ [3] and ROAM [7] suffer drift problem and show inferior performance in terms of point correspondence, our method estimates the point set location more stably. Note that we use the first and the last frames as the key frame for Roto++ (2kf).

5. Cross-domain Evaluation

Because our method can track points containing low-level features, robustness in cross-domain prediction is expected. To prove this hypothesis, we test our model trained as described in the main paper on cell tracking sequences published in [8]. Cell tracking images belong to a different domain from the training data, in that the images are grayscale and the edge of the target deforms unstably. As Figure 5 shows, while a state-of-the-art video object segmentation method [6] fails to predict the desired object, our method tracks cells successfully. We only report the early four frames for [6] since the mask prediction is failed following frames. For our method, we visualize the results on the following frames separately in the third partition.



(a) Spatial Accuracy (SA).



(b) Temporal Accuracy (TA).

Figure 3: Plots of each metrics on different thresholds. Our method outperforms other competitors over entire thresholds.

References

[1] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018.

[2] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[3] Wenbin Li, Fabio Viola, Jonathan Starck, Gabriel J Brostow, and Neill DF Campbell. Roto++ accelerating professional rotoscoping using shape manifolds. *ACM Transactions on Graphics (TOG)*, 2016.

[4] Yao Lu, Xue Bai, Linda Shapiro, and Jue Wang. Coherent parametric contours for interactive video object segmentation.

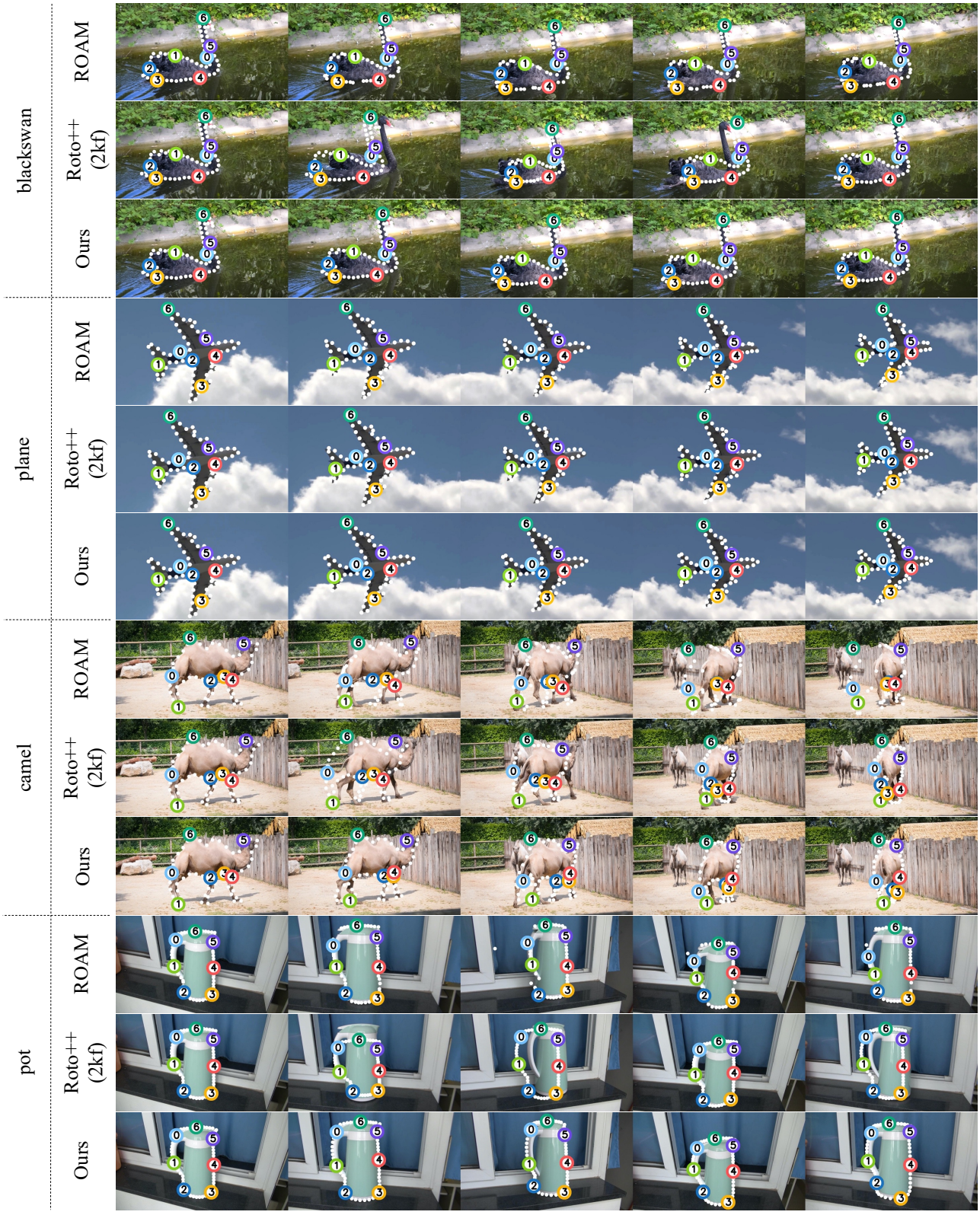


Figure 4: Qualitative results of our method, Roto++ and ROAM on various datasets. While the other methods do not take the point correspondence into account, our method estimates the correspondence more precisely than the others.

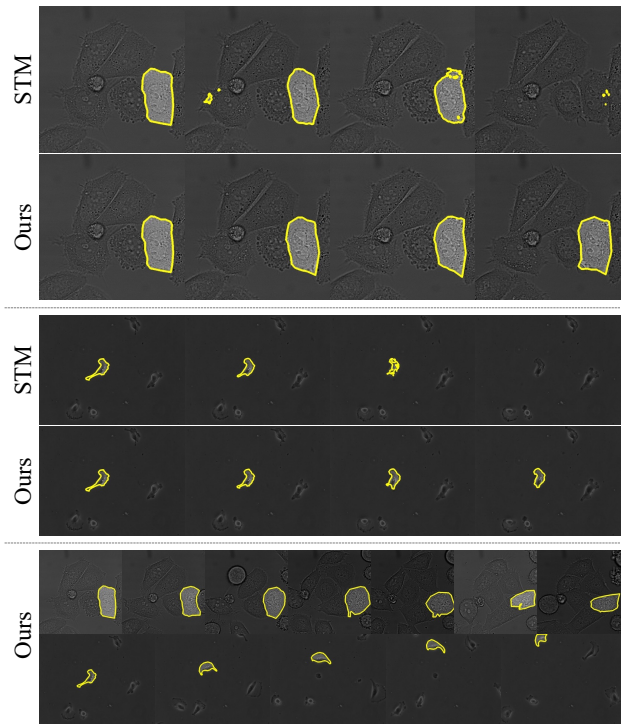


Figure 5: Cell tracking for cross-domain evaluation. Comparison with a state-of-the-art method of video object segmentation is shown in the first two partitions separated by the dash lines. We visualize our results over following frames in the last partition.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[5] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[7] Juan-Manuel Perez-Rua, Ondrej Miksik, Tomas Crivelli, Patrick Bouthemy, Philip HS Torr, and Patrick Perez. Roam: A rich object appearance model with application to rotoscoping. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[8] Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. An objective comparison of cell-tracking algorithms. *Nature methods*, 2017.