Supplementary Material: Self-Supervised Multi-Object Tracking with Path Consistency

Zijia Lu, Bing Shuai, Yanbei Chen, Zhenlin Xu, Davide Modolo
AWS AI Labs
lu.zij@northeastern.edu, {bshuai,yanbec,xzhenlin,dmodolo}@amazon.com

In the supplementary material, we include the details of our model structure (Section 1), simplification of PCL for easy computation (Section 2), implementation details (Section 3), additional ablation study (Section 4) and our approach for selecting start, end frames for PCL (Section 5).

1. Model Structure

As introduced in Section 3.1 of the paper, our model generates object embeddings, h, to compute the object matching probability, p. In this section, we present the details of our model structure.

Model Input. Recall that the input to our model at training and inference is a clip of T continuous frames, along with objects detected by an off-the-shell detector. Take frame t as an example. It contains $N_t - 1$ detected objects and one special $null$ object, ϕ. For the real, detected objects, let $\mathbf{B}^t \in \mathbb{R}^{N_t-1,5}$ denote the four coordinates (top, right, left, bottom) and confidence scores of their bounding boxes and $\mathbf{I}^t \in \mathbb{R}^{N_t-1,H,W,3}$ denote the cropped image patches inside the bounding boxes. For the $null$ object, we learn a fixed embedding vector, as will be explained below. Thus, $\{\mathbf{B}^t, \mathbf{I}^t\}$ are the inputs to our model.

Model Structure. Our model computes the embeddings for all objects. First, it computes features for the detected objects based on their visual and spatial information. It encodes the visual information via convolution layers,

$$\mathbf{V}^t \in \mathbb{R}^{N_t-1,D} = \text{convolution}(\mathbf{I}^t), \tag{1}$$

where $\mathbf{V}^t$ is obtained visual embedding and D is the embedding dimension size. Next, it concatenates $\mathbf{V}^t$ with the spatial information $\mathbf{B}^t$ to obtain the joint embedding, $\mathbf{F}^t \in \mathbf{R}^{N_t-1,D+5} = \text{concat}(\mathbf{V}^t, \mathbf{B}^t)$. Lastly, we also include a learned embedding of the null object, f_ϕ, to obtain the feature matrix of all objects,

$$\tilde{\mathbf{F}}^t \in \mathbb{R}^{N_t,D+5} = \text{concat}(\mathbf{F}^t, f_\phi). \tag{2}$$

As $\tilde{\mathbf{F}}^t$ is computed for each object individually, we further refine the embeddings by considering the context of other objects in the same frame via self-attention layers,

$$\mathbf{H}^t \in \mathbb{R}^{N_t,D} = \text{self-attention}(\tilde{\mathbf{F}}^t), \tag{3}$$

where $\mathbf{H}^t$ is the final embeddings of all objects in frame t. Then h_i^t (the i-th row of $\mathbf{H}^t$) is the embedding for the object o_i^t in the frame and used to compute matching probability $p(o_i^t \to o_j^r)$ in Eq(1) of the paper. Note that, we only consider matching from real, non-$null$ object to $null$ object, as searching for the matches of $null$ object is ambiguous.

Overall, our object embedding encodes the visual, spatial information of an object and considers the other objects in the same frame. While one can also incorporate the objects in adjacent frames with temporal cross-attention, it increases learning difficulty and cannot converge well given the small scale of existing tracking datasets. Therefore, we do not include it in our model.

2. Simplified Formula of PCL

In Eq(6) of the paper, we introduce our path consistency loss $\mathcal{L}_{\text{PC}}(o_i^{t_s}, t_e)$, which contains two terms: KL divergence and probability entropy. Here we show that the loss can be simplified for easy computation,

$$\begin{aligned}\mathcal{L}_{\text{PC}}\left(o_i^{t_s}, t_e\right) &= \frac{1}{|\Pi|}\sum_{\pi} KL(q_\pi^{t_e}||\hat{q}) + H(q_\pi^{t_e}) \\ &= H(\hat{q}) - \frac{1}{|\Pi|}\sum H(q_\pi^{t_e}) + \frac{1}{|\Pi|}\sum H(q_\pi^{t_e}) \\ &= H(\hat{q}),\end{aligned} \tag{4}$$

which is simply the entropy of the averaged association probability distribution, $\hat{q}$.

3. Implementation Details

Model. We set the length of input video clip as $T = 48$. The image dimension is $H = W = 64$ and feature dimension $D = 64$. Specifically, we maintain the aspect ratios of the image patches, resize their longest sides to 64 and add

Spatial	Visual	IDF1	HOTA	MOTA	IDsw
✓		62.9	57.3	62.8	1329
	✓	66.8	60.2	63.7	293
✓	✓	**68.9**	**60.9**	**63.7**	**257**

Table 1. Effect of Input Modality.

padding to the other sides. Our model uses 6 convolution layers (each with a kernel size of 3x3 and spatial stride of 2) and 2 self-attention layers (each with 8 attention heads). Our convolution layers have the same structure as the CNN in UNS.

Training. In addition to our PCL and regularization losses, we follow the practices in [4–6] that improve appearance models via detection techniques. Formally, our method creates two different views of a input video clip via data augmentation (random flip, shift, etc.) and requires the matching probability p to be consistency across the views. It improves convergence speed and avoids trivial solutions.

Inference. Similar to all prior works[4–10], we found it is beneficial to combine our learned model with a motion tracker (SORT [1]). Thus, we match new objects to tracklets based on the average of our tracklet-object similarity (see Eq(2) of paper) and IoU score from the motion tracker. For computation efficiency, we only maintain the latest $M = 4$ object instances in each tracklet. A unmatched tracklet is kept in a buffer for 30 frames to handle occlusion.

4. Ablation Study on Input Modality

Our model associate objects with two input modalities: visual and spatial modalities. In Table 1, we compare the effect of removing each modality to study if our model can jointly exploit two modalities. We show that removing either input modality leads to a clear performance drop, which also causes slower convergence during training. In particular, removing visual information leads to a larger IDF1 drop of 6% (68.9-62.9) as appearance information is vital in tracking over occlusion. In contrast, using two modalities together leads to the best overall performance, yielding an IDF1 of 68.9% and a HOTA of 60.9%. These results suggest that instead of replying on one single modality, our model indeed utilizes the complementary information between the two input modalities to achieve better tracking performance.

5. Selecting Frame Pairs for PCL

As explained in section 3.2.2 of the paper, we compute path consistency loss with sampled start, end frames t_s, t_e and query objects $o_i^{t_s}$. It is important for the sampled data to satisfy that (i) the end frame is as far from the start frame as possible to support learning long-distance matching; (ii) the query object is visible in the intermediate and end frames to obtain meaningful association.

Unfortunately, the visibility/presence of objects in each frame is unknown in unsupervised setting. Thus, we estimate it with bounding box overlaps (IoU) between objects and derive a sample strategy, which selects query objects to determine the start and end frames, and finds all possible groups of query object, start and end frames in the input video clip to make the best utilization of training data. Note that we only pick one query object in each start frame to have the minimal constraint on selecting the end frame.

Specifically, we start with the first object in frame 1, i.e., use the object o_1^1 as query object and frame 1 as start frame. o_1^1 is assumed to exist in frame 2 if its IoU with the closest object in frame 2, o_u^2, is higher than a threshold, σ. Similarly, the query object exists in frame 3 if o_u^2 has high IoU with o_v^3, its closest object in frame 3. We repeat this process until no object with high IoU is found and use the last frame as the end frame. We also mark $\{o_1^1, o_u^2, o_v^3, \ldots\}$ as used, signaling they do not need to be re-selected as query objects, as they are likely instances of the same object. Similarly, we select the remaining objects in frame 1 and subsequent frames as query objects while skipping the used objects. Hence, each unique object in the video clip should be selected as query object for approximately one time while start and end frames are temporally disclose. With our approach, the median of temporal distances between (t_s, t_e) is 36 frames while query objects are present in 98% of the end frames. As a result, we observe 50% of the frame skipping in paths is longer than 8 frames, providing hard training samples for learning long distance matching.

Note the sample strategy does not provide pseudo labels to our model. It only chooses the start and end frames while object associations are learned using PCL. Moreover, our method only requires the query object is present up-to the end frame and is not affected if the object is still present after the frame. Yet in pseudo-label methods [2–4], such scenario means the object will form a new tracklet after the frame, thus is assigned with different pseudo IDs before and after the end frame.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 2

[2] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. Simple unsupervised multi-object tracking. *CoRR*, abs/2006.02609, 2020. 2

[3] Yu-Lei Li. Unsupervised embedding and association network for multi-object tracking. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization, 2022.

[4] Kai Liu, Sheng Jin, Zhihang Fu, Ze Chen, Rongxin Jiang, and Jieping Ye. Uncertainty-aware unsupervised multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9996–10005, 2023. 2

[5] Sha Meng, Dian Shao, Jiacheng Guo, and Shan Gao. Tracking without label: Unsupervised multiple object tracking via contrastive similarity learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16264–16273, 2023.

[6] Bing Shuai, Xinyu Li, Kaustav Kundu, and Joseph Tighe. Id-free person similarity learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14669–14679, 2022. 2

[7] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[8] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021.

[9] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020.

[10] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *CVPR*, 2022. 2