Chained-Tracker: Chaining Paired Attentive Regression Results for End-to-End Joint Multiple-Object Detection and Tracking (Supplementary Material)

Jinlong Peng[1] *, Changan Wang[1] *, Fangbin Wan[2], Yang Wu[3] **, Yabiao Wang[1], Ying Tai[1], Chengjie Wang[1], Jilin Li[1], Feiyue Huang[1], and Yanwei Fu[2]

[1] Tencent Youtu Lab {jeromepeng, changanwang, caseywang, yingtai, jasoncjwang, jerolinli, garyhuang}@tencent.com
[2] Fudan University {fbwan18, yanweifu}@fudan.edu.cn
[3] Nara Institute of Science and Technology yangwu@rsc.naist.jp

1 Overview

This supplementary material includes:
(1) The detailed design of the CTracker network architecture. (Sec. 2)
(2) The details of data augmentation in training. (Sec. 3.1)
(3) The details of Chained-Anchors setting. (Sec. 3.2)
(4) The detailed experiment results of CTracker and the qualitative comparison with other SOTA methods, including POI [1] and Tracktor [2]. (Sec. 4)
(5) The experiment of adding the appearance feature to CTracker. (Sec. 5)

2 Details of Network Architecture

As in Fig. 1, we refer to Resnet50 [3] and FPN [4] to build multi-scale feature representations at five scale levels, we denote them as $\{P_2, P_3, P_4, P_5, P_6\}$.

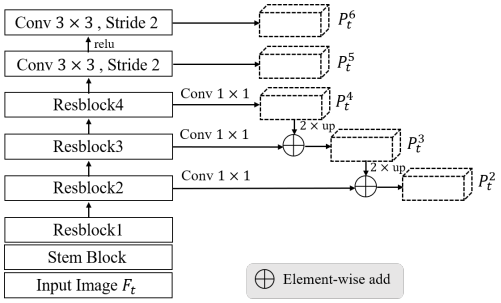

Fig. 1. The detailed architecture of backbone in CTracker network.

* *Equal contribution.*
** *Corresponding author: Yang Wu (wuyang0321@gmail.com)*

Then we combine the features from two adjacent frames at each scale for subsequent prediction, as in Fig. 2,. With the combined features, we apply two parallel branches to perform object classification and ID verification. The two branches consist of four consecutive 3×3 conv layers interleaved with ReLU activations to perform feature learning for specific tasks, above which a 3×3 conv with Sigmoid activation is appended to predict the confidence.

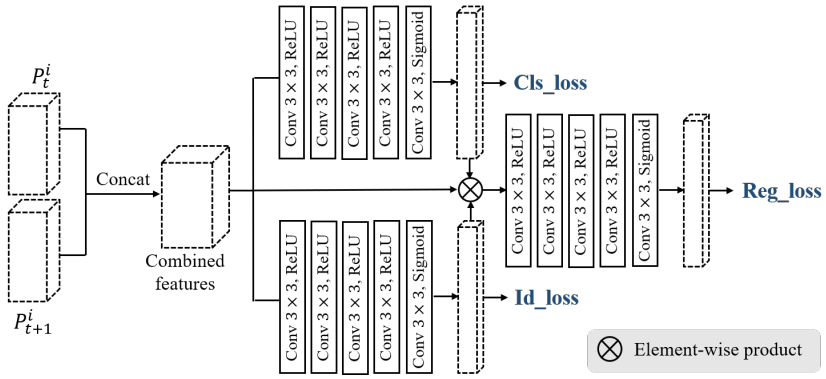

Fig. 2. The detailed architecture of prediction head in CTracker network. P_t^i and P_{t+1}^i are the multi-scale feautres of two adjacent frames, where $i \in \{2, 3, 4, 5, 6\}$.

Finally, we gather the above two predictions by multiplication to get the joint attention map. Since the attention map has the same spatial size as the combined features but with only single channel, we first apply broadcasting on the attention map so that they have compatible shapes, then we perform the attention guidance using element-wise product. With the attention-assistant features, we use the paired boxes regression branch with four conv layers to generate paired boxes for objects of interest. All the box pairs generated from the five scales are post-processed with soft-nms [5] together.

3 Details of Implementation

3.1 Data Augmentation

In order to construct a robust model for objects with different motion speed, we randomly sample two frames with a temporal interval of no more than 3 frames, then we reverse the order of the two frames with 50% probability to form a training pair (*i.e.*, $1 <= |\delta| <= 3$ in Sec. 3.4 of the main text). To further prevent over-fitting, each frame in the pair will be applied with the same data augmentations as follows:
(1) Randomly apply some photometric distortions introduced in SSD [6].
(2) Randomly crop a patch of the size determined by multiplying a random factor in the interval [0.3, 0.8] with the image's shorter side. Note that we only keep those ground truths whose IoMs (Intersection over Min-area) with the cropped

Table 1. Detailed tracking results of CTracker on MOT16 test dataset.

Sequence	MOTA↑	IDF1↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓
MOT16-01	42.0	39.3	79.9	30.4%	30.4%	713	2918	77
MOT16-03	83.6	65.5	78.3	81.1%	0.7%	5600	11024	520
MOT16-06	54.7	52.8	77.1	27.6%	24.0%	795	4158	273
MOT16-07	52.7	41.4	78.6	22.2%	13.0%	587	6884	249
MOT16-08	37.2	35.2	81.8	19.0%	33.3%	499	9824	190
MOT16-12	46.7	53.5	78.5	19.8%	37.2%	112	4250	59
MOT16-14	43.7	43.0	77.1	12.8%	32.9%	628	9247	529
Total	67.6	57.2	78.4	32.9%	23.1%	8934	48305	1897

Table 2. Detailed tracking results of CTracker on MOT17 test dataset.

Sequence	MOTA↑	IDF1↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓
MOT17-01	51.2	44.4	78.7	25.0%	29.2%	202	2891	54
MOT17-03	84.9	66.5	77.9	83.1%	0.7%	5133	10211	479
MOT17-06	56.1	55.2	78.2	29.7%	24.3%	516	4398	261
MOT17-07	50.2	41.0	79.3	21.7%	23.3%	424	7761	228
MOT17-08	31.6	29.6	81.2	14.5%	42.1%	405	13828	212
MOT17-12	47.0	55.7	79.2	18.7%	35.2%	91	4432	69
MOT17-14	39.5	42.7	77.4	10.4%	30.5%	657	9976	540
Total	66.6	57.4	78.2	32.2%	24.2%	7428	53497	1843

patch are greater than 0.2.
(3) With 20% probability, expand the cropped patch using a random factor ranging in [1, 3] by padding with the mean pixel value from ImageNet.
(4) Flip the expanded patch randomly and resize it to a square patch with the size equivalent to the half of the original image's shorter side.

3.2 Chained-Anchors Setting

To determine the scales of Chained-Anchors, we run k-means clustering [7] on all ground truth bounding boxes in the dataset, then we pick five cluster centroids as the scales for Chained-Anchors in different levels of FPN. In our experiments, we use Chained-Anchors of scales $\{38, 86, 112, 156, 328\}$ for $\{P_2, P_3, P_4, P_5, P_6\}$ respectively, and the same ratio of 2.9 is taken for Chained-Anchors of all scales.

4 Detailed Experiment Results

The detailed experiment results of CTracker on MOT16 [8] test dataset and MOT17 test dataset are displayed in Table 1 and Tabel 2.

Moreover, we select two representative qualitative cases to compare our CTracker with the private detection online SOTA method POI [1] and the public detection online SOTA method Tracktor [2]. Fig. 3 displays the tracking results of POI and our CTracker in sequence MOT16-03. In Fig. 3(a), using the POI method, long-term cross-frame tracking drift occurs in several trajectories, which are marked with yellow dotted circles. While in Fig. 3(b), using our CTracker method, there is no long-term cross-frame tracking drift in all the trajectories. For simplicity and efficiency, we focus on the short-term tracking based on the

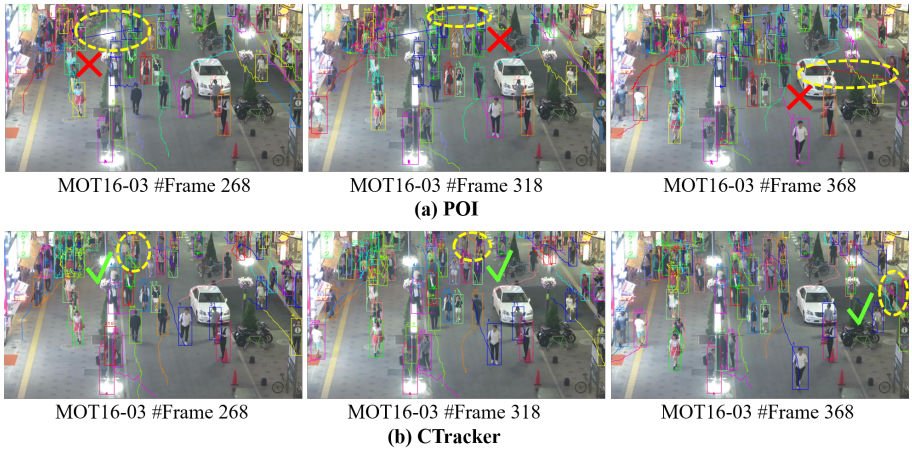

Fig. 3. Qualitative comparison of POI (a) and our CTracker(b).

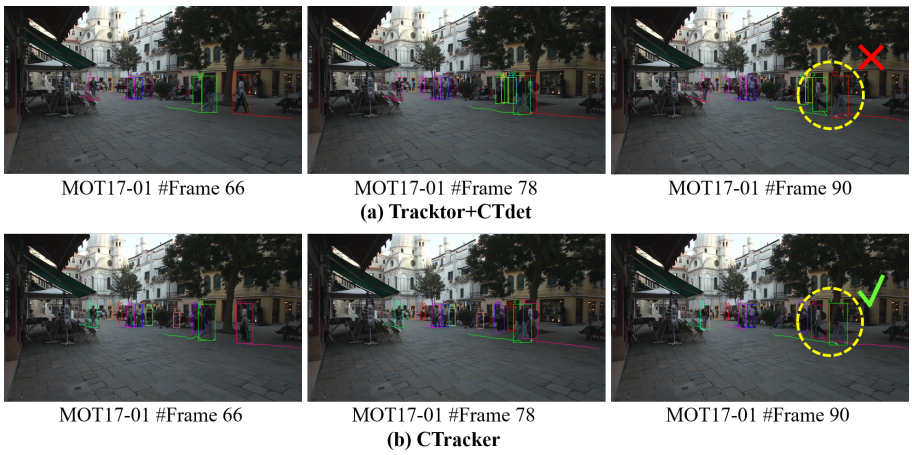

Fig. 4. Qualitative comparison of Tracktor (a) and our CTracker(b).

Chained-Anchors and abandon using the patch-level ReID features of the detected boxes like POI to enhance long-term cross-frame tracking, which may reduce some trajectory integrity to a certain degree while improve the trajectory accuracy greatly. In Fig. 4(a), using the Tracktor method with the same detection of our CTracker, there is a ID switch of trajectory 2 and trajectory 17 due to the occlusion, which is marked with a yellow dotted circle. While in Fig. 4(b), using our CTracker method, the two trajectories representing the same pedestrians are generated correctly due to the accurate box pair association in the CTracker network, which demonstrates the effectiveness of our CTracker in the hard occlusion scene. More complete and clear visualization tracking comparison is displayed in the video attachments.

5 Appearance Feature Experiment

In the main text, to keep the simplicity and efficiency of our CTracker, we abandon using the patch-level ReID features of the detected boxes like other MOT methods to enhance cross-frame data association. In fact, we conduct a appearance feature experiment though we think that it is not related to our main innovations. In the node chaining module, expect for the IoU affinity, we calculate the appearance similarity by adding in the appearance features (256-dim vector from the feature map before the output convolution in the ID verification branch). On MOT16, MOTA increases from 67.6 to 68.5, IDF1 increases from 57.2 to 61.8, IDS decreases from 1897 to 983. While the tracking speed decreases from 34.4fps to 29.2fps. Therefore, We can get better tracking performance when speed loss is acceptable, demonstrating the good expandability of CTracker.

References

1. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: Poi: multiple object tracking with high performance detection and appearance feature. In: ECCV. (2016)
2. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV. (2019)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)
4. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. (2017)
5. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms – improving object detection with one line of code. In: ICCV. (2017)
6. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. (2016)
7. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: CVPR. (2017)
8. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)