P3AFormer Supplementary

Zelin Zhao[1]*, Ze Wu[2], Yueqing Zhuang[2], Boxun Li[2], and Jiaya Jia[1,3]

[1] The Chinese University of Hong Kong
[2] MEGVII Technology
[3] SmartMore

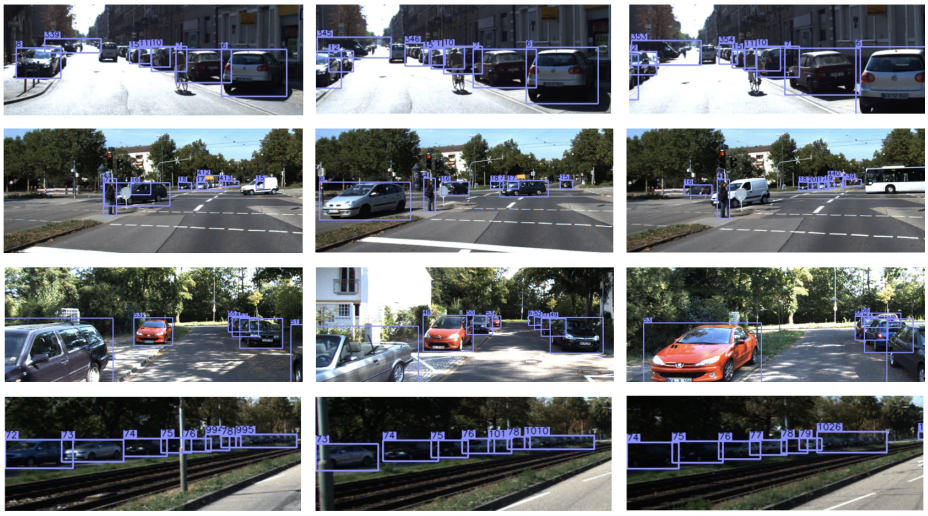

Fig. 1. Visualization of tracking results of P3AFormer on `KITTI-val`.

1 Extended related work

This section mentions some related work not included in the main text due to spatial constraints.

1.1 Pixel-wise tracking

Bibby et al. [1] uses pixel-wise posterior to model object relationships across frames, and their work is not a deep learning method. After that, an unpublished work from Song et al. [14] proposes to adopt pixel-wise information in single-object tracking, which relies on segmentation annotations. A recent benchmark STEP [15,7] requires segment and track every pixel, which is different from standard MOT settings.

* The work was done when Zelin Zhao took internship at SmartMore.

1.2 Understanding vision transformers

Several recent findings of vision transformers inspire the design of the P3AFormer. First, local inductive biases can improve the training of vision transformer [13,10], which supports the mask attention [5]. Second, the training of vision transformers can be unstable due to negative Hessian eigenvalues [13]. Heavy data augmentations [13,6] can mitigate this effect. Moreover, pixel-wise techniques may further smooth the loss landscapes [13], which motivates our training scheme.

2 Methodology details

We provide more details of the methodology parts in this section.

2.1 Pixel-wise similarity function

During pixel-wise propagation, we adopt a pixel-wise similarity function proposed by [21]. Given two feature maps $\mathbf{P}_l^{(t)}$ and $\mathbf{P}_l^{(t-1)->(t)}$, the pixel-wise similarity of each location p is computed as:

$$w^{(t-1)->(t)}(p) = \exp\left(\frac{\mathbf{P}_l^{(t)}(p) \cdot \mathbf{P}_l^{(t-1)->(t)}(p)}{\left|\mathbf{P}_l^{(t)}(p)\right| \left|\mathbf{P}_l^{(t-1)->(t)}(p)\right|}\right). \tag{1}$$

Algorithm 1: Pixel-wise association at timestep t.

Input: tracks, center heatmaps confidence scores, bounding boxes;
Output: updated tracks;
Predict new locations of tracks via Kalman Filter;
Match the tracks with the predicted heatmaps via the Hungarian algorithm;
for *all unmatched objects* **do**
    initialize a new track for it if its confidence is larger than the threshold η_s;
end
Remove a track if it's dead for n_k frames;

2.2 Pixel-wise Association Algorithm

Here we sketch the pixel-wise association in Algorithm 1. This algorithm depicts the same procedure as **Fig.3.** of the main text.

3 Experimental details

3.1 Training process.

The input image is of shape 1440×800 for MOT17/MOT20 and 1280×384 for KITTI. Following [8,19], we use data augmentation, such as Mosaic [2] and

Mixup [18,11]. We use AdamW [12] with an initial learning rate of 6×10^{-5}. We adopt the poly learning rate schedule [4] with weight decay 1×10^{-4}. The full training procedure lasts for 200 epochs. The P3AFormer models are all trained with eight Tesla V100 GPUs. The specific configurations of the losses are provided in the supplementary. The run-time analysis of different models is provided in the supplementary.

3.2 Loss configurations

The weight for the cross-entropy loss is 0.1, the focal loss is 0.5, and we use 1.0 for the size loss.

3.3 Ablation studies

We specify the details of the ablated models in this subsection.

Vanilla Model When we remove all the pixel-wise techniques from P3AFormer, the model is reduced to a vanilla deformable DETR [20] and the association strategy is purely based on the detected bounding boxes.

Pro. We add feature propagation to the vanilla model, which means the model consumes two frames as input via the backbones and uses the pixel-wise feature propagation to align the pixel-level feature embeddings. The pixel-level embeddings are sent to the DETR decoder [3] to get the final predictions. The rest parts are the same as the vanilla model.

Pre. We leverage our proposed pixel decoder and the object decoder to get the object centers and sizes. Those predictions are directly sent to the tracker, and the tracking is purely based on the bounding boxes.

Pre.+Ass. We output the center heatmaps via the Pre. Model and track objects via the pixel-wise association algorithm.

Pro.+Pre. We adopt multi-frame input in the Pre. model and track objects based on bounding boxes.

Pro.+Pre.+Ass. This is the full P3AFormer model.

3.4 Generalization experiments

We give more details of the generalization experiments. Our implementation is based on their official released code. The vanilla Tractor model predicts the temporal realignment of bounding boxes and uses a re-id network to enhance the association of the objects.

Table 1. Running time of different models on the MOT17 dataset.

Model	Time (ms)
TransCenter [16]	112.4
MOTR [17]	132.3
P3AFormer (ours)	108.2

Table 2. Validating the effectiveness of the matching threshold η_m on `MOT17-val`.

η_m	MOTA ↑	IDF1 ↑
0.75	76.3	75.1
0.65	78.4	76.0
0.55	76.6	75.7

Table 3. Validating the effectiveness of the track initialization threshold η_s on `MOT17-val`.

η_s	MOTA ↑	IDF1 ↑
0.90	77.8	72.9
0.80	78.4	76.0
0.70	74.2	73.6

Table 4. Validating the effectiveness of the kill-dead threshold n_k on `MOT17-val`.

n_k	MOTA ↑	IDF1 ↑
40	75.1	72.6
30	78.4	76.0
20	73.4	76.4

+Pro. We adopt the pixel-wise feature propagation to align the feature maps of consecutive frames extracted by the backbone before sending them into heads.

+Pro.+Pre. We change the shape of heads to output pixel-wise center heatmaps and sizes. The rest parts are the same as Tractor.

+Pro.+Pre.+Ass. We use our proposed pixel-wise association algorithm to associate the pixel-wise predictions from the modified Tracktor model. Note that the reID and motion models are not used in this setting.

3.5 Run-time analysis of tracking models

We evaluate the running time of different models: TransCenter [16], MOTR [17] and ours. The results are presented in Table 1. The running time is averaged for each frame. We observe that our pixel-wise techniques do not increase the overall running time (because our pixel-wise association techniques can be implemented efficiently via matrix operations). Although the transformer-based

approaches are generally slower than the highly optimized detectors [19], we believe transformers can become more efficient [9] in the future.

3.6 Ablation studies on hyperparameters

We change various hyperparameters in pixel-wise association and the results are presented in Table 2, Table 3 and Table 4. We find that the P3AFormer can work well under a variety of hyper-parameters.

3.7 Visualizations

We provide more visualizations on the KITTI dataset in Figure 1. We found that P3AFormer can track small objects of different classes on the KITTI dataset.

References

1. Bibby, C., Reid, I.: Robust real-time visual tracking using pixel-wise posteriors. In: European Conference on Computer Vision. pp. 831–844. Springer (2008)
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. arXiv preprint arXiv:2112.01527 (2021)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Fong, W.K., Mohan, R., Valeria, H.J., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. IEEE Robotics and Automation Letters (2022)
8. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
9. Kitaev, N., Kaiser, Ł., Levskaya, A.: Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451 (2020)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
11. Liu, Z., Li, S., Wu, D., Chen, Z., Wu, L., Guo, J., Li, S.Z.: Unveiling the power of mixup for stronger classifiers. arXiv preprint arXiv:2103.13027 (2021)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
13. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=D78Go4hVcxO`
14. Song, Y., Li, C., Wang, Y.: Pixel-wise object tracking. arXiv preprint arXiv:1711.07377 (2017)
15. Weber, M., Xie, J., Collins, M., Zhu, Y., Voigtlaender, P., Adam, H., Green, B., Geiger, A., Leibe, B., Cremers, D., et al.: Step: Segmenting and tracking every pixel. arXiv preprint arXiv:2102.11859 (2021)
16. Xu, Y., Ban, Y., Delorme, G., Gan, C., Rus, D., Alameda-Pineda, X.: Transcenter: Transformers with dense queries for multiple-object tracking. arXiv preprint arXiv:2103.15145 (2021)
17. Zeng, F., Dong, B., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. arXiv preprint arXiv:2105.03247 (2021)
18. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

19. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. arXiv preprint arXiv:2110.06864 (2021)
20. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
21. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 408–417 (2017)