Pedestrian Head Detection and Tracking via Global Vision Transformer

Xuan-Thuy Vo[1], Van-Dung Hoang[2], Duy-Linh Nguyen[1], and Kang-Hyun Jo[1]

[1] Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan (44610), South Korea
[2] Ho Chi Minh City University of Technology and Education, Ho Chi Minh City, Vietnam
Email: xthuy@islab.ulsan.ac.kr,dunghv@hcmute.edu.vn,
ndlinh301@mail.ulsan.ac.kr,acejo@ulsan.ac.kr

Abstract. In recent years, pedestrian detection and tracking have significant progress in both performance and latency. However, detecting and tracking pedestrian human-body in highly crowded environments is a challenging task in the computer vision field because pedestrians are partly or fully occluded by each other. That needs much human effort for annotation works and complex trackers to identify invisible pedestrians in spatial and temporal domains. To alleviate the aforementioned problems, previous methods tried to detect and track visible parts of pedestrians (e.g., heads, pedestrian visible-region), which achieved remarkable performances and can enlarge the scalability of tracking models and data sizes. Inspired by this purpose, this paper proposes simple but effective methods to detect and track pedestrian heads in crowded scenes, called PHDTT (Pedestrian Head Detection and Tracking with Transformer). Firstly, powerful encoder-decoder Transformer networks are integrated into the tracker, which learns relations between object queries and image global features to reason about detection results in each frame, and also matches object queries and track objects between adjacent frames to perform data association instead of further motion predictions, IoU-based methods, and Re-ID based methods. Both components are formed into single end-to-end networks that simplify the tracker to be more efficient and effective. Secondly, the proposed Transformer-based tracker is conducted and evaluated on the challenging benchmark dataset CroHD. Without bells and whistles, PHDTT achieves 60.6 MOTA, which outperforms the recent methods by a large margin. Testing videos are available at https://bit.ly/3eOPQ2d.

Keywords: Pedestrian Head Detection · Pedestrian Head Tracking · Vision Transformer · Crowded Scenes · Surveillance Systems

1 Introduction

Pedestrian detection and tracking are fundamental tasks in visual image and video understanding, which have attracted much attention in recent years. These two tasks have widely applied to many real-world applications such as surveillance systems, action recognition, abnormal detection, robot navigation, human-machine interaction, and autonomous vehicles.

In crowded scenes, pedestrian detection and tracking are challenging missions in computer vision research due to the high density of pedestrians on the road. The tracking performances rely on the level of crowd occlusion. When increasing the high density of pedestrians, trackers produce mislocalized results because a pedestrian is largely occluded with other pedestrians, and the models are ambiguously learned to determine the boundaries of each pedestrian since the appearance features are very identical. As a result, ambiguous learning makes the networks generate more false positives and identity changes during tracking. This reason causes performance degradation. Moreover, annotating pedestrian full-body bounding boxes takes a high cost, and it is difficult to enlarge the scalability of the model and data size. Existing methods [21,23] provide efficient annotations only localizing visible regions of pedestrians and give us a promising opportunity to investigate pedestrian heads detection and tracking in crowded scenes.

Current trends in computer vision utilizing vision Transformer for visual understanding tasks such as image classification [7,14,28], object detection [5,42], multiple object tracking [16, 22, 33, 34], object segmentation [8, 29, 35] bring promising advantages of self-attention and cross-attention mechanisms, and general modeling capacities. Transformer is originally designed for machine translation in natural language processing task, which achieves significant improvements in modeling long-range dependencies in input data. ViT [7] was the first method applying Transformer encoder architecture to vision tasks, which shows its simpleness and effectiveness. To fully leverage both Transformer encoder and decoder into the detector, DETR [5] uses self-attention blocks in Transformer encoder to model the global image features extracted from Convolutional Neural Networks (CNNs) backbone. And then, DETR considers a set of object queries as a set of predictions and learns the relation between image global features and the set of object prediction through cross-attention blocks in Transformer decoder and Hungarian matching to reason about object categories and locations. According to DETR, this work investigate the benefits of Transformer into pedestrian heads detection and tracking tasks, called PHDTT. Firstly, PHDTT performs detection in the adjacent frames based on correlation learning of object queries and image global features. Secondly, PHDTT associates detection results between previous frame and current frame based on track queries. Each object query indicates one object in the current current and each track query represents one object in the previous frame. It means PHDTT considers both detection predictions and association between frames as queries into single end-to-end network to reason about tracking results via global Transformer. Thus, the proposed method is simple and consistent compared to ReID-based methods [12,30,37], and motion-based methods [18,32,39].

We conduct and evaluate the proposed PHDTT on benchmark CroHD [23] for pedestrian detection and tracking tasks in crowded scenes. Without bells and whistles, PHDTT achieves the 60.6 MOTA on the CroHD test set, which surpasses the state-of-the-art head trackers by a large margin. It is noteworthy that PHDTT is the first method leveraging the Transformer encoder and decoder into pedestrian heads detection and tracking. We hope the researcher can use our PHDTT as the baseline for improvement and comparison.

2 Related Works

The pedestrian heads detection and tracking are strongly correlated to multiple object tracking (MOT) tasks. Hence, we briefly summarize the method-based taxonomy of the MOT task based on milestones. Both pedestrian heads tracking and MOT include two essential steps: detection and data association. Therefore, we also provide a review of these two steps.

Multiple Object Tracking. In tracking literature, MOT is grouped into two methods that are tracking by detection and detection by tracking. Firstly, tracking-by-detection methods employ advanced object detectors [5,9,13,19,20,40] to improve tracking performance since data association procedure heavily relies on detection results. Most state-of-the-art trackers [27,30,37–39] use CenterNet [40] as the detector. Recent methods [18,26] utilize the single-stage object detector RetinaNet [13] for the detection step. Due to high efficiency, some existing methods [12,30] use YOLO [19] as the detector, and ByteTrack [36] employs the simple and effective YOLOX [9] for producing detection results. Secondly, detection-by-tracking methods [6, 41] adopt the tracking model such as single object tracking and Kalman filter to improve detection performance.

Data association. Data association is the crucial step in MOT, matching detection results based on similarity scores and ReID-based methods. SORT [3], DeepSORT [31] predicts future object location via Kalman filter and computes the IoU scores between predicted objects and detected objects. And then, these methods apply the Hungarian algorithm to assign each identity to each object based on IoU cost. In the most popular methods [12,17,30,37] add a new ReID branch to the detection network for predicting appearance features and also use the Hungarian algorithm to perform a one-to-one assignment.

Pedestrian heads tracking. In recent years, many researchers [10,24] have paid much attention to pedestrian full-body detection by introducing multi-scale features, loss-based anchor assignments. However, head detection is paid less attention by researchers and needs more investigations. Head detection has been widely used in crowd counting and intelligent surveillance systems. HeadHunter [23] was the first method to solve pedestrian heads detection and tracking in crowded environments, applying Faster R-CNN [20] for performing detection. Since the head regions are very small, HeadHunter uses high input resolution to detect small heads. To do that, extracted features from the backbone are upsampled to higher resolutions via the Context Sensitive module and transposed convolutions. For data association, HeadHunter utilizes IoU and center distance similarity and Hungarian matching to perform assignments between the same targets.

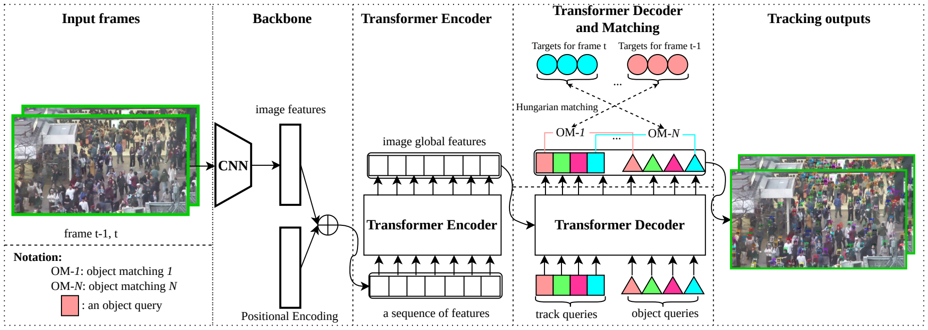


Fig. 1. The overall architecture of the proposed PHDTT. This network includes six essential procedures: input frames, backbone, Transformer encoder, Transformer decoder, Object matching, and tracking outputs.

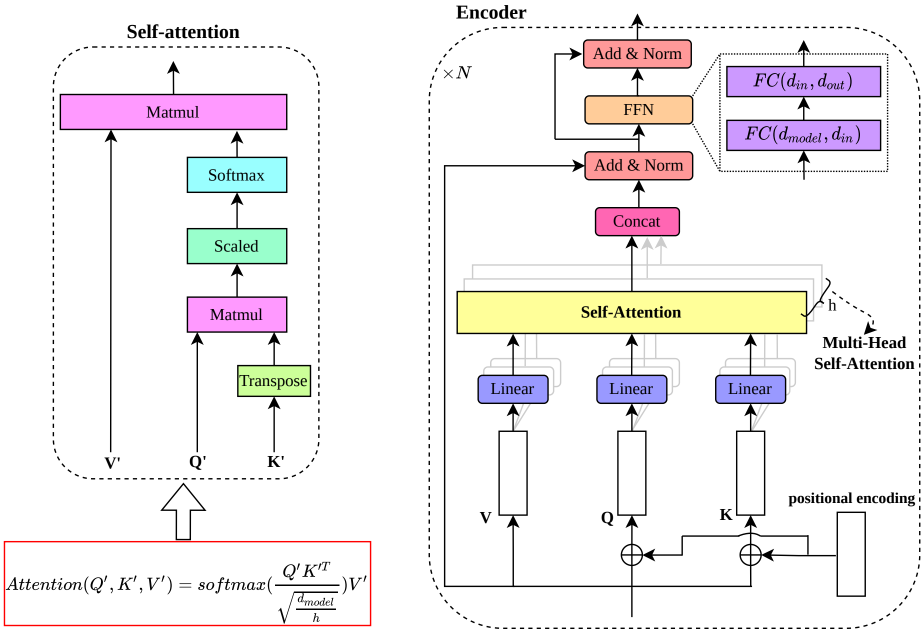
Vision Transformer. ViT [7] was the first method that brings the original Transformer encoder from natural language processing to computer vision, which shows the promising improvements in both modeling capacity and performance on par with advanced CNNs. DETR [5] applies Transformer encoder and decoder and Hungarian matching to reason about detection prediction through the relation between object queries and image global features, which achieves high efficiency and simple architecture.

3 The proposed method

The overall architecture of the proposed method is shown in Fig. 1. The input of this system takes adjacent frames e.g., frame $t-1$: $I^{t-1} \in \mathbb{R}^{3\times H\times W}$ and frame t: $I^{t} \in \mathbb{R}^{3\times H\times W}$. The input features are extracted by CNNs network, denoted by $F \in \mathbb{R}^{2048\times\frac{H}{32}\times\frac{W}{32}}$. We take this feature F from stage 5 of the backbone network, and this feature has a low resolution suitable for generating a sequence of features. Because the model complexity of Transformer encoder and decoder networks [5] quadratically grows with the increase of the input feature sizes.

Transformer Encoder. Firstly, the extracted feature F is mapped from channel 2048 to 256 by using 1×1 convolution and flatten to dimension $S \in \mathbb{R}^{N\times d_{model}}$, where $N = \frac{H}{32} * \frac{W}{32}$ is a sequence dimension and $d_{model} = 256$ is embedding dimension. Secondly, because of the flattened features, the order of sequences is lost. Accordingly, positional encoding is supplemented with input features to learn the relationship between sequences, followed by [25]. The core element of the Transformer encoder is the self-attention module that models long-range dependencies in the input sequences. The detailed architecture of the self-attention operation and Transformer encoder are shown in Fig. 2(a) and Fig. 2(b), respectively.

Similar to [5, 7, 25], the proposed PHDTT creates the query matrix Q, key matrix K, and value matrix V from the feature F. And then, we linearly project



(a) Self-attention operation (b) Transformer Encoder

Fig. 2. The detailed architecture of the self-attention operation (a) and Transformer encoder (b). V, Q, K are value matrix, query matrix, and key matrix. Matmul indicates matrix multiplication. h denotes the number of self-attention blocks inside the multi-head self-attention module. concat is concatenation oepration. FFN is feed-forward network, contains two stacked fully-connected layers to map the low dimension d_{model} to high dimension d_{in} and map back to original dimension d_{model}. N is the number of multi-head self-attention modules in one Transformer encoder network.

each matrix embedding to lower dimension $\frac{d_{model}}{h}$ and the model can perform all self-attention blocks in a parallel way. Multi-Head self-attention module includes h self-attention blocks. In each self-attention block, we compute attention weights defined as follows,

$$Attention(Q^{'}, K^{'}, V^{'}) = softmax(\frac{Q^{'}K^{'}}{\sqrt{\frac{d_{model}}{h}}})V^{'}, \tag{1}$$

where $Q^{'}, K^{'}, V^{'}$ are the projected query, key, and value matrices, respectively. The core idea of self-attention block is to learn correlation between query-key elements through dot-product operation. Generally speaking, one query position globally gathers all information of key positions and output which positions are important to be emphasized in value features. After using FFN, we can model image global features.

Transformer Decoder. The detailed architecture of the Transformer decoder is illustrated in Fig. 3. The main purpose of the Transformer decoder is

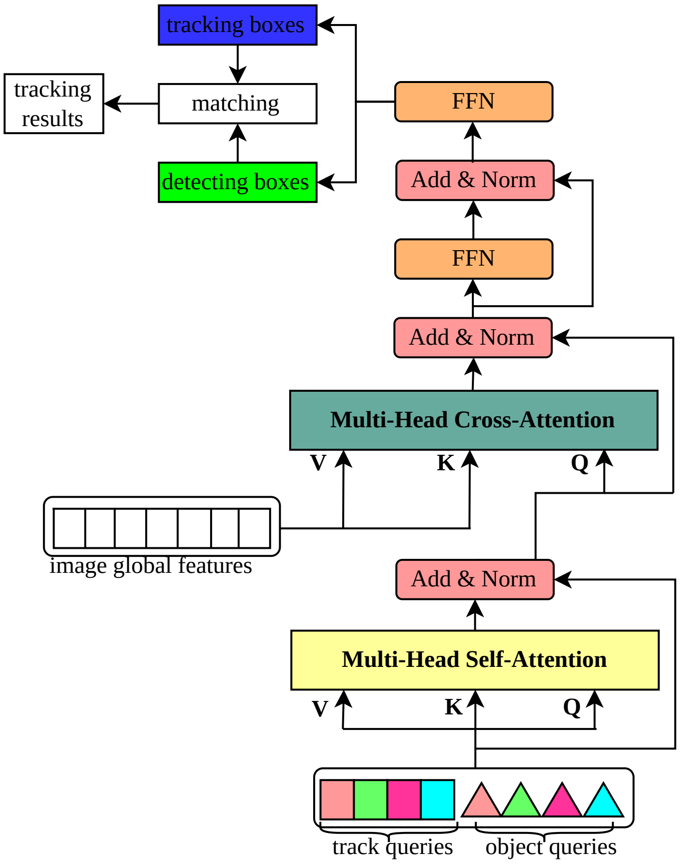


Fig. 3. The detailed architecture of the Transformer decoder. Multi-head self-attention block is computed similar to the computation in the Transformer encoder. The main component of the Transformer decoder is the multi-head cross-attention block that learns the relation between image global features and combined queries.

to learn the relation between object queries *vs.* image global features to reason about detecting boxes for frame t, track queries *vs.* image global features to reason about tracking boxes, and object queries *vs.* track queries to facilitate data association step and also improve detection task. Firstly, we generate a set of learned object queries with dimension N responsible for predicting N objects at frame t. Detected objects at the previous frame $t-1$ (i.e., track queries) are combined with object queries and passed to Transformer decoder architecture to model the object similarities of two adjacent frames through multi-head self-attention. These similar features are set as query matrix in the multi-head cross-attention module, and the image global features are set as key and value matrices. Secondly, FFN is used to generate the final prediction with $2N$ bounding boxes: N detecting boxes and N tracking boxes. To match objects with the same targets, we compute IoU scores between each paired prediction (i.e., tracking box *vs.* detecting box). If the IoU score is greater than a certain threshold, this paired box is assigned as the positive sample otherwise. Finally, we apply the Hungarian matching algorithm for assigning positive samples to box targets that follow the procedure in DETR [5].

Matching algorithm. N detecting boxes, and N tracking boxes are assigned to ground truth boxes via Hungarian matching. The cost for the assignment is linear combination of classification and localization losses, defined as,

$$\mathcal{L}_{match} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{loc}\mathcal{L}_{loc}, \tag{2}$$

where $\mathcal{L}_{cls} = -\log \hat{p}(c_i)$ is a negative log-likelihood for computing classification cost in which $\hat{p}(c_i)$ denotes the probability of class c_i. $\mathcal{L}_{loc} = \lambda_{L1}\left\|b_{i,t} - \hat{b}_{i,t}\right\| + \lambda_{giou}\mathcal{L}_{giou}(b_{i,t}, \hat{b}_{i,t})$ is the localization cost that is linear combination of $\mathcal{L}1$ and $\mathcal{L}_{giou}$ costs. $\lambda_{cls}, \lambda_{L1}, \lambda_{giou}$ are the balancing terms.

4 Experiments

4.1 Dataset and Evaluation Metrics

We conduct and evaluate the proposed PHDTT on the challenging benchmark dataset CroHD [23]. This dataset recorded in crowded scenarios consists of four training videos corresponding to 5740 training frames and five testing videos corresponding to 5723 testing frames. The average of pedestrian head density over all videos is approximately 178 pedestrians per frame. The performances on the test set are submitted to the evaluation system[3], and our evaluated results under the name PHDTT are available at this link[4].

For performance evaluation, the proposed method is measured by standard metrics: Multiple Object Tracking Accuracy (MOTA) proposed by [2], the ratio of correctly identified detections IDF1 [2], Higher Order Tracking Accuracy (HOTA) defined by [15], and IDEucl [23]. Extra metrics used to evaluate all aspects of the proposed method are Mostly tracked targets (MT), Mostly lost targets (ML), the total number of false positives (FP), the total number of false negatives (FN), and the number of identity switches (ID Sw.).

4.2 Implementation Details

The experiments are implemented by the deep learning Pytorch framework. The used backbone network for feature extraction is ResNet-50 [11] pre-trained on ImageNet dataset for weight initialization in tracking model. We train the model on Crowdhuman [21] for 150 epochs and fine-tune the trained model on the training set of CroHD for 20 epochs. We use the GPU Tesla V100 device with Cuda 10.2, and CuDNN 7.6.5 to train the tracking model with batch size of 16

[3] https://motchallenge.net/

[4] https://motchallenge.net/results/Head_Tracking_21/

Table 1. Comparison with state-of-the-art tracking methods on CroHD test set

Method	MOTA	IDF1	IDEucl	HOTA	MT	ML	FP	FN	ID Sw.
SORT [3]	46.4	48.4	58.0	-	2.1	9.2	-	-	**649**
V_IOU [4]	53.4	35.4	34.3	-	3.4	7.8	-	-	1,890
Tracktor [1]	58.9	38.5	31.8	-	5.3	**5.0**	-	-	3,474
HeadHunter [23]	57.8	**53.9**	**54.2**	**36.8**	31.9	19.9	**51,840**	**299,459**	4,394
Our PHDTT	**60.6**	47.9	52.6	36.1	**47.1**	8.0	132,714	184,215	15,004

Table 2. The detailed performance on each video of the CroHD test set

Video	MOTA	IDF1	IDEucl	HOTA	MT	ML	FP	FN	ID Sw.
HT21-11	79.0	63.7	65.5	47.8	84	4	1,485	4,620	232
HT21-12	72.0	57.6	63.1	40.7	514	11	36,824	48,584	2,352
HT21-13	37.3	24.9	27.0	20.2	192	63	66,256	60,498	19,005
HT21-14	68.4	59.7	55.1	43.4	223	59	11,514	46,604	2,551
HT21-15	55.1	44.3	52.4	33.1	91	51	16,635	23,909	4,863

and learning rate schedule followed by [5]. The number of object queries is set to $N = 500$. The optimizer is AdamW to minimize the training objectives,

$$\mathcal{L}_{training} = \lambda_{cls}\mathcal{L}_{focal} + \lambda_{giou}\mathcal{L}_{giou}, \tag{3}$$

where $\mathcal{L}_{focal}$ is the Focal loss [13] for classification loss and $\mathcal{L}_{giou}$ is the GIoU localization loss.

5 Results

As shown in Table 1, our proposed PHDTT surpasses all state-of-the-art trackers by a large margin. Specifically, the proposed method achieves 60.6 MOTA that outperforms SORT [3] by 14.2 MOTA, V_IOU [4] by 7.2 MOTA, Tracktor [1] by 1.7 MOTA, and HeadHunter [23] by 2.8 MOTA. It demonstrates the effectiveness and generalization capacity of the tracker PHDTT. The proposed method joins both detection and association tasks into the single end-to-end network that each task can help to learn another task efficiently. While all methods treat two tasks independently, the total inference time is a sum of two tasks. Based on detected bounding boxes, SORT [3] uses the Kalman filter to predict future motion and the Hungarian matching algorithm to associate future predicted boxes and detected boxes. V_IOU [4] reduces the tracking fragmentation and ID switches by combining a single object tracker into Hungarian matching, improving the data association task. Instead of using available detection results, HeadHunter [23] utilizes Faster R-CNN as the baseline detector. To meet the requirement of small head sizes, Headhunter introduces a Context-Sensitive Prediction Module and uses transposed convolutions to balance the high-level and low-level features of the backbone and to upsample multi-level features to higher spatial resolutions.

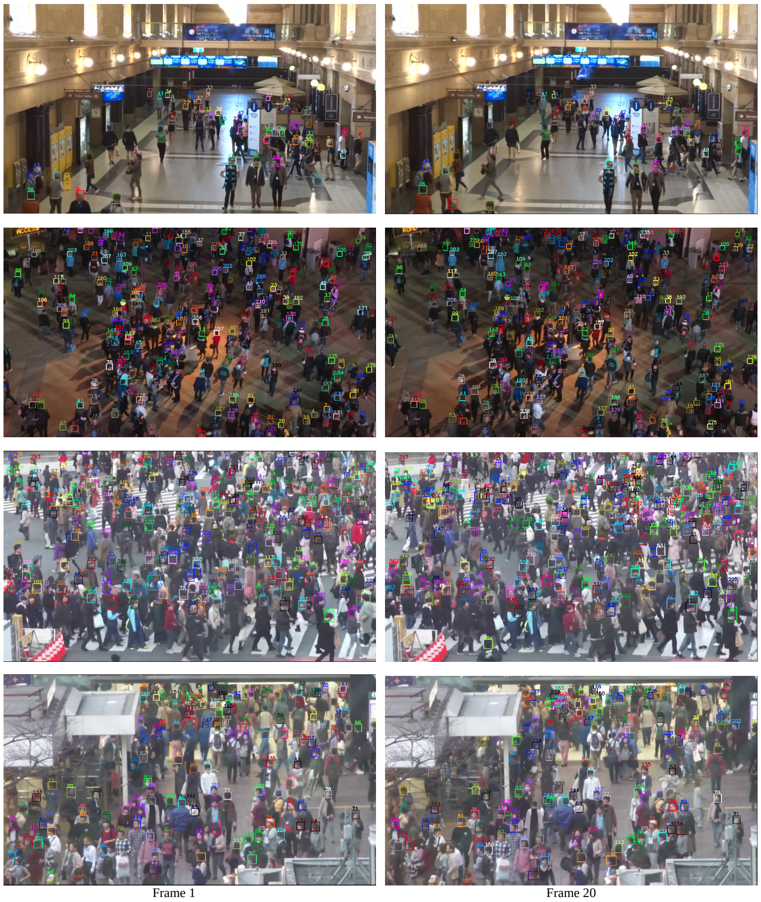


Fig. 4. The qualitative results of the proposed method on the CroHD validation set with different crowded scenes. Each number indicates each identity of each person.

Table 2 shows the detailed performances on five sequences of the CroHD test set. Specifically, PHDTT achieves 79.0 MOTA, 72.0 MOTA, 37.3 MOTA, 68.4 MOTA, and 55.1 MOTA on video HT21-11, HT21-12, HT21-13, HT21-14, and HT21-15, respectively. The proposed method poorly performs on the sequence HT21-15 and HT21-13 because both videos contain the highest pedestrian density, motion blur, and crowd occlusion problem. We will consider this challenging problem for future researches.

The qualitative results of our proposed PHDTT are shown in Fig. 4. The visualization performances are evaluated under crowded conditions with different scenes such as indoor, outdoor night, and day. As a result, the PHDTT model can detect small pedestrian heads and track assigned pedestrian heads accordingly. It is noteworthy that the proposed method only uses the single-level feature and does not increase the spatial resolution for detecting small objects. Thus, our approach is efficient while HeadHunter [23] uses transposed convolution to upsample the multi-level features to higher spatial dimensions.

6 Conclusion

This paper leverages the powerful Transformer encoder and decoder architectures into pedestrian head detection and tracking tasks. The track queries and object queries are responsible for detection predictions in adjacent frames. The Transformer network not only extracts the long-range dependencies in image features but also learns the object relations between frames to reason about tracking boxes and detecting boxes. The query-key mechanism facilitates the data association procedure since the object similarity across frames is learned through the Transformer decoder. Integrating the matching step into the detection network can improve the overall performance and enhance the model's capacity learning at the current frame. Without bells and whistles, the proposed method surpasses the existing methods by a large margin, becomes the state-of-the-art tracker. To the best of our knowledge, there are no existing methods that apply the benefits of the Transformer encoder and decoder to solve pedestrian head detection and tracking researches. We believe that our proposed PHDTT can serve as the simple baseline for pedestrian head detection and tracking tasks.

Acknowledgement

This results was supported by "Region Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE)(2021RIS-003).

References

1. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 941–951 (2019)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016)
4. Bochinski, E., Senst, T., Sikora, T.: Extending iou based multi-object tracking by visual information. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2018)
5. 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)

6. Chu, P., Wang, J., You, Q., Ling, H., Liu, Z.: Transmot: Spatial-temporal graph transformer for multiple object tracking. arXiv preprint arXiv:2104.00194 (2021)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
8. Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Instances as queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6910–6919 (2021)
9. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
10. Ge, Z., Wang, J., Huang, X., Liu, S., Yoshie, O.: Lla: Loss-aware label assignment for dense pedestrian detection. arXiv preprint arXiv:2101.04307 (2021)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Liang, C., Zhang, Z., Lu, Y., Zhou, X., Li, B., Ye, X., Zou, J.: Rethinking the competition between detection and reid in multi-object tracking. arXiv preprint arXiv:2010.12138 (2020)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
14. 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 (ICCV). pp. 10012–10022 (October 2021)
15. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision **129**(2), 548–578 (2021)
16. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. arXiv preprint arXiv:2101.02702 (2021)
17. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 164–173 (2021)
18. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: European Conference on Computer Vision. pp. 145–161. Springer (2020)
19. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
21. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
22. Sun, P., Jiang, Y., Zhang, R., Xie, E., Cao, J., Hu, X., Kong, T., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple-object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020)

23. Sundararaman, R., De Almeida Braga, C., Marchand, E., Pettre, J.: Tracking pedestrian heads in dense crowd. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3865–3875 (2021)
24. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: A simple and strong anchor-free object detector. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
26. Vo, X.T., Tran, T.D., Nguyen, D.L., Jo, K.H.: Regression-aware classification feature for pedestrian detection and tracking in video surveillance systems. In: International Conference on Intelligent Computing. pp. 816–828. Springer (2021)
27. Wang, Q., Zheng, Y., Pan, P., Xu, Y.: Multiple object tracking with correlation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3876–3886 (2021)
28. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 568–578 (October 2021)
29. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8741–8750 (2021)
30. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 107–122. Springer (2020)
31. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017)
32. Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: An online multi-object tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12352–12361 (2021)
33. 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)
34. Zeng, F., Dong, B., Wang, T., Chen, C., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. arXiv preprint arXiv:2105.03247 (2021)
35. Zhang, W., Pang, J., Chen, K., Loy, C.C.: K-net: Towards unified image segmentation. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021)
36. 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)
37. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021)
38. Zheng, L., Tang, M., Chen, Y., Zhu, G., Wang, J., Lu, H.: Improving multiple object tracking with single object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2453–2462 (2021)
39. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision. pp. 474–490. Springer (2020)
40. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)

41. Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., Yang, M.H.: Online multi-object tracking with dual matching attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 366–382 (2018)
42. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)