DynamicTrack: Advancing Gigapixel Tracking in Crowded Scenes

Yunqi Zhao*
Tsinghua University
Beijing, China
yq-zhao22@mails.tsinghua.edu.cn

Yuchen Guo*
Tsinghua University
Beijing, China
yuchen.w.guo@gmail.com

Zheng Cao
Biren Technology
Beijing, China
zcao@birentech.com

Kai Ni
HoloMatic Technology Co.,Ltd.
Beijing, China
nikai@holomatic.com

Ruqi Huang
Tsinghua University
Beijing, China
ruqihuang@sz.tsinghua.edu.cn

Lu Fang†
Tsinghua University
Beijing, China
fanglu@tsinghua.edu.cn

***Abstract*—Tracking in gigapixel scenarios holds numerous potential applications in video surveillance and pedestrian analysis. Existing algorithms attempt to perform tracking in crowded scenes by utilizing multiple cameras or group relationships. However, their performance significantly degrades when confronted with complex interaction and occlusion inherent in gigapixel images. In this paper, we introduce DynamicTrack, a dynamic tracking framework designed to address gigapixel tracking challenges in crowded scenes. In particular, we propose a dynamic detector that utilizes contrastive learning to jointly detect the head and body of pedestrians. Building upon this, we design a dynamic association algorithm that effectively utilizes head and body information for matching purposes. Extensive experiments show that our tracker achieves state-of-the-art performance on widely used tracking benchmarks specifically designed for gigapixel crowded scenes.**

***Index Terms*—Multi-object Tracking, Gigapixel Image, Crowded Scenes, Contrastive Learning**

I. Introduction

As the development of imaging devices, the acquisition of gigapixel images [1], [2] has become increasingly feasible. Gigapixel images, with large spatial coverage and high imaging quality, have substantial potential applications in smart cities, such as traffic monitoring [3] and pedestrian surveillance [4] in crowded scenes.

Although gigapixel images offer richer semantic information and finer-grained targets for crowded scenes compared to megapixel images, they also introduce complex interactions and severe occlusion issues, posing new challenges for tracking. Some researchers attempt to address severe occlusion in crowded scenes through multi-camera tracking [5]. However, the rigid separation of continuous space by multiple cameras leads to the dispersion of spatial information. Others seek robust tracking by leveraging interaction information provided by group relationships [6] among pedestrians, yet capturing group relationships in crowded scenes proves challenging. The existing issues with current methods make it difficult to apply gigapixel images in practice.

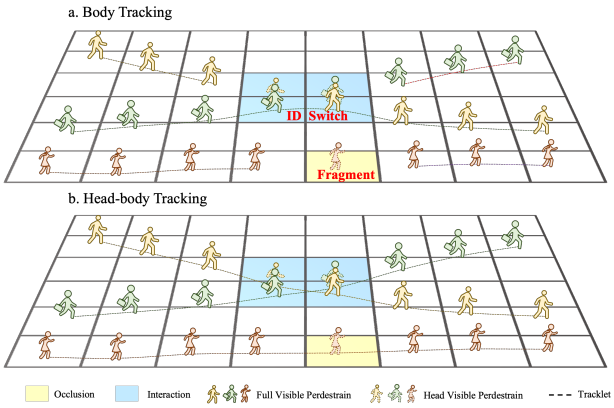

Fig. 1. The comparison between body tracking and head-body tracking: a. Body tracking encounters ID switch and fragment in interactive and occluded scenarios. b. Head-body tracking is robust in crowded scenes.

In this paper, we present DynamicTrack that comprises dynamic detection and association modules to address the above mentioned challenges in gigapixel crowded scenes. Head features are less likely to be obscured and thus provide robust trajectory cues, enhancing tracking accuracy in crowded scenes. Based on this observation, we propose a dynamic detector for head-body joint detection, facilitating the joint tracking of heads and bodies. Specifically, we incorporate embedding learning into the dynamic detector and utilize the embedding loss derived from contrastive learning for feature learning. To fully leverage the distinct characteristics of the head and body in crowded scenarios, we propose a dynamic association algorithm that incorporates head features into the matching process. The dynamic association algorithm treats the body as the core and the head as the support, which combines fine-grained local head features with global body information. We conduct extensive experiments to evaluate the performance of our proposed tracker. The results demonstrate

*Co-first author, †Corresponding author.

This work is supported in part by National Natural Science Foundation of China (NSFC) under contract No. 62125106, 61860206003, 62088102 and U21B2013.

that our dynamic detector achieves the best performance in joint detection on the Crowdhuman [7] dataset. Moreover, our proposed DynamicTrack outperforms state-of-the-art methods on MOT20 [8] and PANDA [9] datasets.

Our contributions are summarized as the following:

1) Dynamic Detection: We propose a dynamic detector based on contrastive learning, which enables joint head and body detection for gigapixel tracking.
2) Dynamic Association: We introduce a dynamic association algorithm to fully exploit the potential of both head and body cues for joint matching.
3) We demonstrate the superior performance of our tracker on widely used tracking benchmarks designed for crowded scenes.

II. Related Work

With the development of imaging devices, obtaining gigapixel images [1], [2] is no longer difficult. Wang et al. introduced the PANDA [9] dataset, designed to address visual tasks under gigapixel resolution. Although PANDA offers rich semantic annotations and fine-grained information, its crowded scenes introduce complex interaction and occlusion, posing significant challenges to tracking. In this work, we incorporate joint head-body tracking into the tracking framework, which enables robust gigapixel tracking in crowded scenes. Some studies have attempted joint detection of head and body. PedHunter [10] employs a mask-guided module to leverage the head information to enhance the representation learning of pedestrian features. Double Anchor R-CNN [11] presents a double-anchor RPN to capture body and head parts in pairs. JointDet [12] detects head and body simultaneously and performs relational learning between them for joint detection. These methods indirectly utilize head information for better detection but have limitations in achieving end-to-end optimization. Some researchers have attempted to integrate head information into tracking frameworks. Sun et al. [13] utilize the harder-to-obscure head feature as the basis for tracking and replace the association result with the matching body. However, tracking based on the head is not suitable for tasks focused on the body as the target. Zhang et al. [14] perform head-body matching based on a positional prior followed by joint head-body tracking. However, the relative position prior of the head and body is not robust under occlusion.

III. DynamicTrack

Our goal is to design a gigapixel tracker for crowded scenes. We introduce a dynamic tracking framework into the traditional Separate Detection and Embedding tracking framework [16], [17] and the proposed *DynamicTrack* framework is shown in Fig. 2. First, we implement an end-to-end dynamic detector based on contrastive learning which is capable of detecting both the body and head of a pedestrian. Then, we propose a dynamic association algorithm that can simultaneously utilize head and body features for robust tracking.

A. Dynamic Detection

The key challenge in gigapixel tracking is dealing with crowded scenes that involve complex interaction and occlusion among pedestrians. To tackle this issue, we have developed an end-to-end dynamic detector capable of simultaneously capturing both the head and body of a pedestrian. Our approach is based on the understanding that head features are less prone to occlusion in crowded environments, and they can thus offer more comprehensive and reliable features for subsequent tasks. To achieve joint detection, we draw inspiration from the concept of associative embedding learning [18] and utilize the associative embedding technique to establish the relationship between the head and body of a pedestrian. As illustrated in Fig. 3, we incorporate a parallel branch into the Faster R-CNN [15] framework. This additional branch is placed after the ROI feature and functions as the embedding module, producing an embedding vector for each instance. To optimize this embedding module, we introduce an Associative Embedding Loss (AML). The AML aims to encourage embeddings from the same pedestrian to be pulled closer together while pushing apart the embeddings belonging to different individuals.

Preliminaries: Given the ground truth annotations $\mathcal{G}$:

$$\mathcal{G} = \left\{ (g_n^{(b)}, g_n^{(h)}) \mid g_n^{(b)} \in \mathcal{G}^{(b)}, g_n^{(h)} \in \mathcal{G}^{(h)} \right\} \tag{1}$$

where $\mathcal{G}^{(b)}$ represents the set of body boxes, $\mathcal{G}^{(h)}$ represents the set of head boxes, and $(g_n^{(b)}, g_n^{(h)})$ represents the matched body and head pair. And the predicted body set is $(\mathcal{D}^{(b)}, \mathbf{e}^{(b)})$ and head set is $(\mathcal{D}^{(h)}, \mathbf{e}^{(h)})$, where $\mathcal{D}$ represents the detection results and $\mathbf{e}$ represents the corresponding embedding features.

Pulling Loss: To ensure that the embedding vectors of positive pairs are close to each other, we design pulling losses for various cases: body and body (bb), head and head (hh), and matched body and head (bh). The pulling loss functions for three cases are as follow:

$$\begin{cases} L_{pull}^{bb} = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} e^{d_{ij}} \|\mathbf{e}_i^{(b)} \mathbf{e}_j^{(b)}\|^2 \\ L_{pull}^{hh} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} e^{d_{ij}} \|\mathbf{e}_i^{(h)} \mathbf{e}_j^{(h)}\|^2 \\ L_{pull}^{bh} = \frac{1}{M} \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} \|\mathbf{e}_i^{(b)} \mathbf{e}_j^{(h)}\|^2 \end{cases} \tag{2}$$

where M and N represent the number of the predicted body set $\mathcal{D}^{(b)}$ and head set $\mathcal{D}^{(h)}$. In the bb and hh cases, we want to mitigate the influence of negative samples that may be geometrically distant. To achieve this, we introduce a distance-aware weighting penalty $e^{d_{ij}}$, where $d_{i,j}$ signifies the distance between respective bounding boxes i and j. By combining these components, we can define the pulling loss as follow:

$$L_{pull} = \mu(L_{pull}^{bb} + L_{pull}^{hh}) + \beta L_{pull}^{bh} \tag{3}$$

In practical implementation, we set μ to 1.0 and β to 1.5.

Pushing Loss: To ensure that the distance between the embedding vectors of negative pairs are as large as possible, we also design pushing losses for various cases: body and body

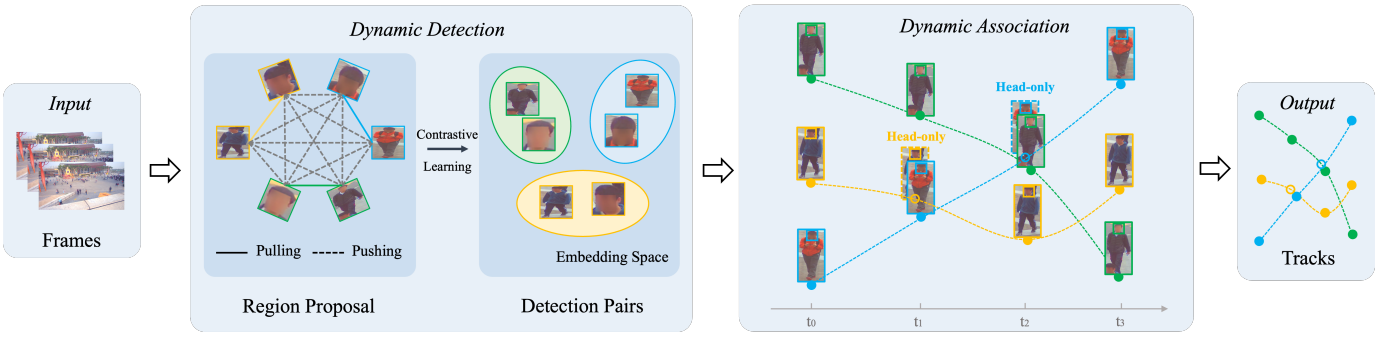

Fig. 2. Overview of DynamicTrack framework for gigapixel tracking. *Dynamic Detection:* Contrastive learning-based detector achieves simultaneous detection of both the body and the head for pedestrian tracking. *Dynamic Association:* Dynamically utilizing head and body of the same identity for matching to achieve robust tracking in crowded scenes."

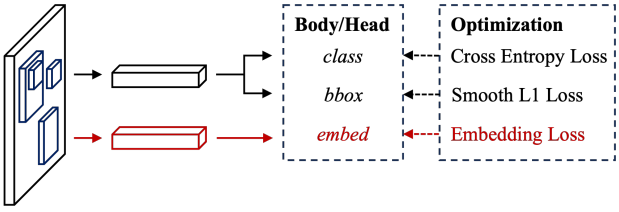

Fig. 3. The framework of our dynamic detector for head-body detection consists of a modified version of the classical two-stage detector, Faster-RCNN [15]. We introduce an additional branch for embedding learning and leverage an associative embedding loss based on contrastive learning for supervision.

(bb), head and head (hh), and matched body and head (bh). However, if the distance between the feature vectors exceeds a threshold σ, we consider the pair as an "easy" negative pair and exclude it from further processing. The pushing loss functions for three cases are as follow:

$$\begin{cases} L_{push}^{bb} = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \| \max(0, \delta - \mathbf{e}_i^{(b)} \mathbf{e}_j^{(b)}) \|^2 \\ L_{push}^{hh} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \| \max(0, \delta - \mathbf{e}_i^{(h)} \mathbf{e}_j^{(h)}) \|^2 \\ L_{push}^{bh} = \frac{1}{M} \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} \| \max(0, \delta - \mathbf{e}_i^{(b)} \mathbf{e}_j^{(h)}) \|^2 \end{cases} \quad (4)$$

where σ is the threshold (which we set to 2 by default), and M and N are similar to the settings in the pulling loss. By combining these components, we can define the pushing loss as follow:

$$L_{push} = \mu(L_{push}^{bb} + L_{push}^{hh}) + \beta L_{push}^{bh} \quad (5)$$

The the weights μ and β in the pushing loss are the same as those used in the pulling loss.

Associative Embedding Loss: Given the pulling loss L_{pull} and pushing loss L_{push}, we can obtain the Associative Embedding Loss by combining them with weighting coefficients σ and τ as follow:

$$Loss_{AML} = \sigma L_{pull} + \tau L_{push} \quad (6)$$

B. Dynamic Association

With the approach outlined in dynamic detection part, we are able to obtain matched head and body detections $\mathcal{D}$. It is worth to note that certain bodies or faces may be absent, which can be represented as $\varnothing$.

$$\mathcal{D} = \{(d_0^{(b)}, d_0^{(h)}), (d_1^{(b)}, \varnothing), (\varnothing, d_2^{(h)}), \ldots, (d_n^{(b)}, d_n^{(h)})\} \quad (7)$$

To effectively utilize the information from both head and body detections, we propose a novel dynamic association algorithm based on the idea of cascade matching. Cascade matching is a technique employed in DeepSORT [16] to facilitate the matching of historical frames. In our dynamic association algorithm, we devise three distinct cases: matched head and body, mismatched body, and mismatched head. By employing the dynamic association algorithm for each of these cases, we are able to effectively leverage the information from both the head and body in occluded environments. Firstly, we associate the matched body and head detection boxes with tracklets to preserve comprehensive information. Subsequently, we merge mismatched body detection boxes with unmatched tracklets to establish a solid foundation for pedestrian tracking. Finally, we reconcile unmatched head detection boxes with unmatched tracklets, enabling us to effectively recover highly obscured objects. The pseudo-code of Dynamic association is shown in Algorithm 1.

The input of the dynamic association algorithm consists of a video sequence, denoted as $\mathcal{V}$, along with matched body and head detection boxes, represented by $\mathcal{D}$. It is important to note that the presence of occlusion may result in the absence of either the head or body in the detection. The objective of this algorithm is to output tracks, denoted as $\mathcal{T}$, for each object in the video. Each track contains the bounding box coordinates and identity of the object in each frame. To achieve this, we first divide all the detection boxes in each frame into three categories: $\mathcal{D}^{(bh)}$ represents the matched body and head, $\mathcal{D}^{(b)}$ represents the mismatched body, and $\mathcal{D}^{(h)}$ represents the mismatched head. Once the detection boxes are separated, we apply the association function to each category to associate the detection boxes with their corresponding tracks.

The initial association is carried out between the matched body-head detection boxes and all the tracks in $\mathcal{T}$. The

Algorithm 1 Dynamic association for head-body detections

Input: Body and head detections $\mathcal{D}$
Output: Tracks $\mathcal{T}$ of the video

function DYNAMIC ASSOCIATION($\mathcal{D}$)
  Initialization: $\mathcal{T} \leftarrow \varnothing$
  for frame f_k in video $\mathcal{V}$ **do**
    $\mathcal{D}^{(bh)}, \mathcal{D}^{(b)}, \mathcal{D}^{(h)} \leftarrow \mathcal{D}$
    /* matched body and head */
    $\mathcal{T}_{remain}, D^{(bh)}_{remain} \leftarrow$ Asso($\mathcal{T}$, $\mathcal{D}^{(bh)}$)
    /* mismatched body */
    $\mathcal{T}_{remain}, D^{(b)}_{remain} \leftarrow$ Asso($\mathcal{T}_{remain}$, $\mathcal{D}^{(b)}$)
    /* mismatched head */
    $\mathcal{T}_{remain}, D^{(h)}_{remain} \leftarrow$ Asso($\mathcal{T}_{remain}$, $\mathcal{D}^{(h)}$)
    /* delete unmatched tracks */
    $\mathcal{T} \leftarrow \mathcal{T} - \mathcal{T}_{remain}$
    /* initialize new tracks */
    $\mathcal{T} \leftarrow \mathcal{T} + \mathcal{D}^{(bh)}_{remain} + \mathcal{D}^{(b)}_{remain}$
  end for
end function
function ASSO($\mathcal{T}$, $\mathcal{D}$)
  $\mathcal{D}_{high}, \mathcal{D}_{low} \leftarrow \mathcal{D}$
  Associate $\mathcal{T}$ and $\mathcal{D}_{high}$
  $\mathcal{D}_{remain} \leftarrow$ remaining object boxes from $\mathcal{D}_{high}$
  $\mathcal{T}_{remain} \leftarrow$ remaining tracks from $\mathcal{T}$
  return $\mathcal{T}_{reamin}$, $\mathcal{D}_{remain}$
end function

similarity matrix are computed using the Intersection over Union (IOU) and Re-ID feature distances between the matched detection boxes $\mathcal{D}^{(bh)}$ and the predicted boxes of tracks $\mathcal{T}$. The Hungarian Algorithm is then employed to complete the matching process based on the similarity matrix. Any unmatched detections are stored in $D^{(bh)}_{remain}$ and the unmatched tracks are stored in $\mathcal{T}_{remain}$. Next, the second association is performed between the mismatched body detection boxes $D^{(b)}_{remain}$ and the remaining tracks $\mathcal{T}_{remain}$ after the first association. Similarity metrics are computed in the same manner as the first association, and the Hungarian Algorithm is applied for the second matching. The unmatched detections are saved in $D^{(b)}_{remain}$, and the unmatched tracks from the second association are stored in $\mathcal{T}_{remain}$. Finally, the third association takes place between the mismatched head detection boxes $D^{(h)}_{remain}$ and the remaining tracks $\mathcal{T}_{remain}$ after the second association. The matching process is conducted in the same way as described above. Any unmatched detections are kept in $D^{(h)}_{remain}$ and the unmatched tracks from the third association are stored in $\mathcal{T}_{remain}$.

After the association, the unmatched tracks will be deleted from the tracklets. For each track in the unmatched tracks $\mathcal{T}_{remain}$ after the third association, only when it exists for more than a certain number of frames, i.e. 10, we delete it from the tracks $\mathcal{T}$. Finally, we initialize new tracks from the unmatched detection boxes $D^{(bh)}_{remain}$ and $D^{(b)}_{remain}$ after the third association. It is worth noting that we did not consider unmatched head detections $D^{(h)}_{remain}$, since heads are only used as supplementary information for tracking purposes, and introducing them into the main tracking framework would introduce more noise. As a result, the output of each individual frame will consist of the bounding boxes and identities of the tracks $\mathcal{T}$ in the current frame.

TABLE I
RESULTS ON MOT20 VAL SET. ALL TRACKERS UTILIZE THE SAME DETECTION RESULTS OBTAINED FROM THE DYNAMIC DETECTOR.

Tracker	MOTA↑	IDF1↑	HOTA↑	IDs↓
ByteTrack	68.5	71.4	57.6	3942
OCSORT	68.3	68.9	56.1	4037
StrongSORT	67.7	69.7	56.8	3253
BotSORT	69.4	71.8	57.7	3168
DynamicTrack	70.2	72.1	57.9	3376

TABLE II
RESULTS ON PANDA TEST SET. ALL TRACKERS UTILIZE THE SAME DETECTION RESULTS OBTAINED FROM THE DYNAMIC DETECTOR.

Tracker	MOTA↑	IDF1↑	HOTA↑	IDs↓
ByteTrack	39.3	34.6	33.6	26265
OCSORT	43.3	34.8	36.0	44634
StrongSORT	55.8	53.9	48.1	11478
BotSORT	57.3	57.4	50.8	10235
DynamicTrack	60.4	59.2	53.9	7646

IV. EXPERIMENT AND RESULT

A. Experimental Setup

We construct dynamic detector based on the well-known Faster-RCNN [15] architecture. We train the dynamic detector using the CrowdHuman dataset [7] and the same training weights were utilized for subsequent experiments. CrowdHuman is primarily focused on pedestrian detection in crowded scenes and provides precise annotations for both human body and head. To evaluate the association module, we conduct joint head and body tracking experiments on the widely-used MOT20 dataset, which includes challenging crowded scenes. Furthermore, we assess the performance of our DynamicTrack on the PANDA dataset [9]. The PANDA dataset is a gigapixel multi-object tracking dataset specifically designed for highly crowded and challenging scenes. When evaluating the detection performance, we utilize two widely used metrics: AP and MR^{-2}. For assessing the body-face association performance, we employ mMR^{-2}, a metric proposed in [19]. This metric quantifies the proportion of body-face pairs that are miss-matched. When evaluating the tracking performance, we primarily rely on three widely-uesd evaluation metrics: MOTA, IDF1, and HOTA. MOTA predominantly assesses detection performance, while IDF1 emphasizes association performance. HOTA aims to strike a balance between accurate detection and association effects.

B. Results of DynamicTrack

Tracking performance on MOT20. In Tab. I, we provide the tracking results on the MOT20 dataset and compare them with the state-of-the-art two-stage methods on MOT20. We

Fig. 4. Visualization results of DynamicTrack. We have selected gigapixel sequences from the test set of PANDA to demonstrate the effectiveness of DynamicTrack in handling complex crowded scenarios. In our visualizations, we utilize customizable visualization windows represented by green and blue rectangles. Additionally, we use colors to indicate different identities, with the same bounding box color indicating the same identity.

Fig. 5. Visualization of detection results and matching head and body pairs on CrowdHuman test set.

use the same detector trained on the CrowdHuman dataset as a baseline. Obviously, our method achieves higher performance of MOTA which is the primary evaluation metric.

Tracking performance on PANDA. In Tab. II, we provide the gigapixel tracking results on the PANDA dataset. Our approach, DynamicTrack, is compared with state-of-the-art two-stage trackers including motion-based methods like ByteTrack [17] and OCSORT [20], as well as appearance-based methods like BotSORT [21] and StrongSORT [22]. From the results, it can be observed that DynamicTrack achieves comparable performance to other state-of-the-art methods.

Qualitative results. Fig. 4 showcases the tracking results under gigapixel sequences. Analysis from the results obtained on Trian Station Square, Dongmen Street and Huaqiangbei indicate that accurate tracking can be achieved in crowded scene of gigapixel sequences. This includes successful tracking of both sparse, large targets in the foreground as well as dense, small targets in the background.

C. Ablation Study

Ablation study of Dynamic Detection. Tab. V presents the detection results on the CrowdHuman dataset. In this table, we compare our novel embedding-based method with the traditional position-based method. The position-based method

TABLE III
ABLATION STUDY OF THE DIFFERENT MODULES OF DYNAMICTRACK ON MOT20 TEST SET.

Body	Head	MOTA↑	IDF1↑	HOTA↑	IDs
✓		68.3	71.3	57.4	4059
✓	✓	70.2	72.1	57.9	3376

TABLE IV
ABLATION STUDY OF THE DIFFERENT MODULES OF DYNAMICTRACK ON PANDA TEST SET.

Body	Head	MOTA↑	IDF1↑	HOTA↑	IDs↓
✓		55.7	55.3	48.5	10384
✓	✓	60.4	59.2	53.9	7646

calculates the IOU distance between the head and body detections, and then utilizes the Hungarian algorithm to select the best matching pairs. From the results, it is evident that our dynamic detector outperforms the position-based approach by a significant margin of **14.22%** with respect to mMR^{-2}. Moreover, our dynamic detector maintains competitive detection performance. Fig. 5 showcases the detection results and matching results in crowded scenarios. It illustrates that our embedding-based method performs better, especially in situations involving complex occlusions.

Ablation study of Dynamic Association. To evaluate the impact of incorporating head information for tracking, we conducted experiments on the widely used MOT20 and PANDA dataset, which consist of crowded scenes. Specifically, we compare the effects of body-based tracking and body-head tracking. The results are summarized in Tab. III and Tab. IV. From the results, it is clear that the inclusion of head information in tracking yields notable improvements. In MOT20 dataset, the head-body tracking method outperforms the body-based tracking method by 1.9 in terms of MOTA. Similarly, in PANDA dataset, the head-body tracking method surpasses the body-based tracking method by 4.7 in terms of MOTA. These results indicate that introducing head information can lead to significant performance gains, particularly in occluded environments.

V. CONCLUSION

In this paper, we address the challenging task of gigapixel tracking in crowded scenes by introducing DynamicTrack. To enhance the robustness to occlusion, we incorporate head information in addition to the traditional body-based features and leverage contrastive learning for dynamic detection. Moreover, we propose dynamic association algorithms for body-head tracking to overcome the challenges posed by gigapixel crowded sequences. Experimental results on benchmark MOT20 and PANDA have shown that our approach outperforms the state-of-the-art trackers in crowded scenes. The future plan is to integrate the dynamic detection framework into the latest transformer-based detectors to further enhance the tracking performance in crowded scenes.

TABLE V
RESULTS ON THE CROWDHUMAN VALIDATION SET. POS: POSITION-BASED METHOD. EMB: EMBEDDING-BASED METHOD.

Method	Class	AP ↑	MR^{-2} ↓	mMR^{-2} ↓
Emb	Head	0.727	0.557	56.57
	Body	0.867	0.459	
Pos	Head	0.743	0.532	70.79
	Body	0.867	0.441	

REFERENCES

[1] D. J. Brady, M. E. Gehm, R. A. Stack, D. L. Marks, D. S. Kittle, D. R. Golish, E. Vera, and S. D. Feller, "Multiscale gigapixel photography," *Nature*, 2012.

[2] X. Yuan, M. Ji, J. Wu, D. J. Brady, Q. Dai, and L. Fang, "A modular hierarchical array camera," *Light: Science & Applications*, 2021.

[3] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. Hwang, "Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *CVPR*, 2019.

[4] Z. Cheng, L. Qin, Q. Huang, S. Jiang, S. Yan, and Q. Tian, "Human group activity analysis with fusion of motion and appearance information," *ACM MM*, 2011.

[5] R. Eshel and Y. Moses, "Tracking in a dense crowd using multiple cameras," *IJCV*, 2010.

[6] F. Zhu, X. Wang, and N. Yu, "Crowd tracking with dynamic evolution of group structures," in *ECCV*, 2014.

[7] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.

[8] P. Dendorfer, A. Osep, A. Milan, K. Schindler, D. Cremers, I. D. Reid, S. Roth, and L. Leal-Taixé, "Motchallenge: A benchmark for single-camera multiple target tracking," *IJCV*, 2020.

[9] X. Wang, X. Zhang, Y. Zhu, Y. Guo, X. Yuan, L. Xiang, Z. Wang, G. Ding, D. Brady, Q. Dai *et al.*, "Panda: A gigapixel-level human-centric video dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[10] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Pedhunter: Occlusion robust pedestrian detector in crowded scenes," in *AAAI*, 2020.

[11] K. Zhang, F. Xiong, P. Sun, L. Hu, B. Li, and G. Yu, "Double anchor r-cnn for human detection in a crowd," *arXiv preprint arXiv:1909.09998*, 2019.

[12] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Relational learning for joint head and human detection," in *AAAI*, 2020.

[13] Z. Sun, J. Chen, M. Mukherjee, H. Wang, and D. Zhang, "An improved online multiple pedestrian tracking based on head and body detection," in *MSN*, 2021.

[14] Y. Zhang, H. Chen, Z. Lai, Z. Zhang, and D. Yuan, "Handling heavy occlusion in dense crowd tracking by focusing on the heads," in *AJCAI*, 2023.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015.

[16] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *ICIP*, 2017.

[17] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *ECCV*, 2022.

[18] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006.

[19] J. Wan, J. Deng, X. Qiu, and F. Zhou, "Body-face joint detection via embedding and head hook," in *ICCV*, 2021.

[20] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric sort: Rethinking sort for robust multi-object tracking," in *CVPR*, 2023.

[21] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "Bot-sort: Robust associations multi-pedestrian tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[22] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "Strongsort: Make deepsort great again," *TMM*, 2023.