Online Multi-Object Tracking With Pose-Guided Object Location and Dual Self-Attention Network

Xin Zhang, Shihao Wang, Yuanzhe Yang,
Chengxiang Chu, and Zhong Zhou (✉)

State Key Laboratory of Virtual Reality Technology and Systems,
Beihang University, Beijing, P.R.China
zz@buaa.edu.cn

Abstract. The recent trend in Multi-Object Tracking (MOT) is heading towards using deep learning to detect objects and extract features. Although tracking frameworks using detection network have achieved outstanding performance in object locating on MOT, it is still challenging for crowded occlusion. In this paper, we propose to alleviate this difficulty by combining bounding boxes from outputs of both object detection and pose estimation. The motivation behind generating redundant candidates is that object detection and pose estimation can complement each other in tracking scenes. In order to get optimal tracking objects from candidates, we present Soft-Pose-NMS. For similarity calculation, we design a Dual Self-Attention Network (DSAN) with the self-attention mechanism. The network generates the self-attention map that enables the network to focus on the object area of detection and tracklet images. Simultaneously, the network can extract the temporal self-attention feature map to suppress noisy images in the tracklet. Experiments are conducted on the MOT benchmark datasets. Results show that our tracker achieves competitive results and is state-of-the-art in half of the metrics.

Keywords: Multi-Object Tracking · Person Re-identification · Dual Self-Attention Network .

1 Introduction

Multi-object tracking (MOT) is one of the most fundamental computer vision tasks, aiming to generate the trajectory information of all interested objects across video frames. It has attracted much attention because of its broad application such as intelligent video analysis, autonomous driving and smart city. The current MOT studies mainly adopt the "tracking-by-detection" strategy that applies the detector to locate objects in each frame and associates objects among the different frames to generate object trajectories[5, 25, 31].

Despite the encouraging progress made in the past few years, there are two significant problems with "tracking-by-detection" strategy. One is that tracking results heavily rely on the quality of object detection, which by itself is hard to

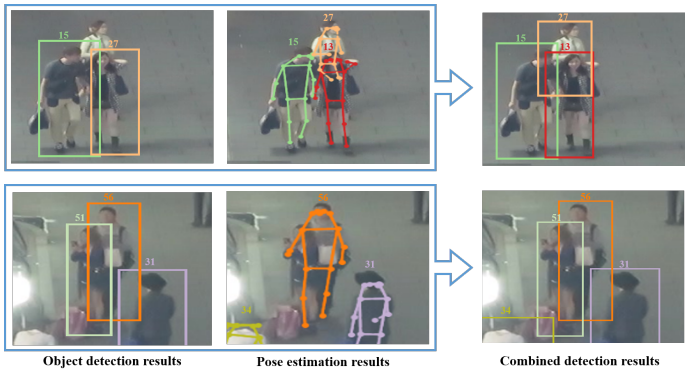

Fig. 1. Object locating with pose guiding. In applying only one kind of detection result, the bounding boxes are mislabeled due to heavy occlusion. Object detection result and pose estimation result can complement each other to locate objects correctly.

generate reliable results across frames. Taking the tracking scenes in the MOT16 dataset as examples, during the crowd scenes, the bounding boxes based on one kind of detection method of the occluded objects is usually unreliable, posing drifting and ID-switching in tracking, as shown in Fig. 1. To alleviate such issues, recent research [24] introduces the object location information from an instance segmentation method to locate the tracking objects. In this paper, we combine the merits of multi-person pose estimation and object detection in a unified framework to introduce object joint points information. We use the pedestrian joint points information to assist in locating the object and alleviate unreliable detection.

On the other hand, for similarity computation in MOT, we need to compare the current detect object with a sequence of previous observations in the trajectory. One of the most commonly track objects in MOT is pedestrians, so the re-identification[22, 16] is commonly used for similarity calculation with challenging factors including occlusion, partial loss and pose variation[31], as shown in Fig. 1. To alleviate such issues, [31, 7] propose the feature extraction network that introduces attention mechanism[27] to extract detection and tracklet appearance features. Additionally, inspired by [29], we introduce the self-attention mechanism, which calculates the self-attention map for detection image and tracklet images, respectively. Moreover, our network is end-to-end, which can alleviate training complexity and extract more robust features.

The main contributions of this paper can be summarized as follows.

1. A new detection strategy is proposed to combine object detection and pose estimation results. The strategy takes advantage of both object detection and pose estimation to handle unreliable detection in online MOT.

2. We design a Dual Self-Attention Network (DSAN), introducing the self-attention mechanism to allocate different attention values to each location in the object image and exploit self-attention temporal feature from the tracklet.

3. Experimental results demonstrate that our tracker achieves competitive performance on the MOT benchmark dataset and is state-of-the-art in half of the metrics.

2 Related Work

In recent years there has been an explosion of technological progress in MOT driven primarily by object detection strategy. Sanchez-Matilla et al. [20] exploited multiple detectors to improve detection performance in MOT. Chen et al. [5] combined detection and predicted bounding boxes by Kalman filter as tracking candidate set for quality evaluation and used different strategies for data association. Although these methods alleviate the unreliable detection results, they still use one kind of detection information. Hence these methods cannot effectively alleviate the issue of missing detection. There are also several works that use other category location information to determine the coordinates of the tracking candidates[6, 10, 24, 13]. Voigtlaender et al. [24] proposed MOTS task and TrackR-CNN network to merge segmentation and multi-object tracking. The network employed top-down segmentation information instead of detection information to locate the object. Nevertheless, the top-down object location information introduced in the above methods still depends on the quality of the object detection results[24, 8]. On the contrary, we propose the Soft-Pose-NMS detection strategy to introduce object joint points information from the bottom-up pose estimation method. The bottom-up object location information is not affected by the object detection performance and can provide additional object position information, and thereby it can effectively improve the object detection results in MOT.

For object feature extraction and similarity computation, Mahmoudi et al. [17] applied CNN extracted appearance features along with position features to calculate more accurate similarity score. Chu et al. [7] introduced a Spatial-Temporal Attention Mechanism (STAM) to handle the tracking drift caused by the occlusion and interaction among objects. Zhu et al. [31] proposed a Dual Matching Attention Networks (DMAN) with both spatial and temporal attention mechanisms to perform the tracklet data association. In this paper, we integrate both spatial and temporal self-attention mechanisms into the proposed MOT framework. Our framework differs from the state-of-the-art DMAN [31] method. First, the spatial attention in the DMAN corresponds to the detection image and trajectory images. Since the attention map is affected by different trajectory images, it becomes unreliable when other objects appear in the trajectory image. In contrast, we exploit the image itself to generate the self-attention map, which is demonstrated to be more robust to inter-object occlusion and noisy detection. Second, the DMAN needs to be divided into two steps to train the model, while our spatial and temporal self-attention map can be end-to-end trained.

3 Proposed Method

Our online tracking framework consists of three tasks, object detection, similarity calculation and trajectory management. We first measure all tracking objects by the proposed Soft-Pose-NMS detection strategy that introduces object pose information. Then we use the Dual Self-Attention Networks (DSAN) to extract feature and compute the similarity score of the detection image and tracklet images. Finally, we update the tracking state of objects and trajectories.

3.1 Soft-Pose-NMS Object Detection Strategy

Given a new frame, we get the joint points of each object through the pose estimation network [15]. Nonetheless, there are abnormal points in these joint points, as shown in Fig. 2. Therefore, the Soft-Pose-NMS detection strategy is designed to generate accurate joint points-based bounding boxes with pose estimation results and determine tracking candidates by screening two types of bounding boxes. These bounding boxes are adopted to alleviate detection failures in crowded scenes.

First, we obtain the primary detection-based bounding box set PB_{det} by object detection method. It is necessary to generate a sufficient number of detection bounding boxes to filter and obtain accurate tracking bounding boxes. Therefore, we set a lower confidence threshold T_{detcon} to generate the detection-based bounding box set B_{det} form PB_{det}.

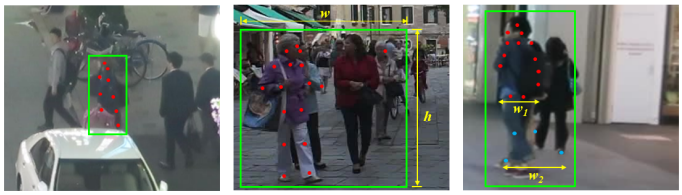

Fig. 2. The bounding box results based on pose estimation. (a) shows the result missing part of the object joint points. (b) shows the results of abnormal joint points with large offsets. (c) shows the result of abnormal joint points with small offsets. Red points and blue points are the clustering result of the object joint points and W_i is the width of two point-groups.

Second, a primary joint points-based bounding box PB_{jpi} is generated by expanding the coordinates of the joint points. Here we define NP_{PBjpi} as the number of joint points and AR_{PBjpi} as the aspect ratio for the PB_{jpi}. Then the primary joint points-based bounding boxes set PB_{jp} can be defined as:

$$PB_{jp} = \{PB_{jp1} ... PB_{jpi}\}, NP_{jpi} > T_{njp} \, and \, AR_{jpi} < T_{ratio} \tag{1}$$

where T_{njp} is threshold for the number of joint points, T_{ratio} is threshold of the aspect ratio. We set T_{njp}=8 and T_{ratio}=0.6 to generate PB_{jp}. However, the joint points-based bounding box coordinate shifting still exists in PB_{jp}, as shown in Fig. 2(c). We observe that this shifting only appears on the abscissa. In order to deal with this joint points drift issue to get exact width value for joint points-based bounding box. First, we use the clustering algorithm to cluster the joint points of each bounding box PB_{jpi} in PB_{jp} into two point groups. Then we calculate the width ratio of the two points groups. Here we define w_1 and w_2 as the width of two point group width, respectively, as shown in Fig. 2(c). We define R_w as the width ratio of w_1 and w_2. Therefore, the width of ith joint points-based bounding box W_{PBjpi} can be generated by the following formula:

$$W_{PBjpi} = \begin{cases} w_1 & R_w > Tw_{ratio} \\ w_2 & R_w \leq Tw_{ratio} \end{cases} \tag{2}$$

where Tw_{ratio} as the threshold of the width ratio. We analyse the position of the drift joint point and set Tw_{ratio} to 2.

After recalculating the width of each joint point-based bounding box, we get the final joint point-based bounding box set B_{jp}. In order to combine detection based bounding boxes and screen unreliable bounding boxes, we need to calculate a reasonable confidence score to the ith joint points-bounding box B_{jpi} in B_{jp}. Directly using the average score of each joint point in joint points-based bounding box B_{jpi} as corresponding confidence value will cause confidence bias. Therefore, we propose a function to explicitly encode pose information of each joint point into the confidence maps. We expand the total variance and make the scoring probability distribution distance of different pedestrians farther. The confidence of B_{jpi} is defined as:

$$CB_{jpi} = \frac{1}{n} \sum_{n}^{i=1} \tan h \frac{s_i}{\sigma} \tag{3}$$

where CB_{jpi} is the confidence of ith joint points-based bounding box B_{jpi}, σ is a data-driven parameter used to control the degree of score suppression and s_i is the score of each joint point. The scores are averaged after $\tan h$ function mapping to generate the confidence CB_{jpi} and the final joint points-based bounding box set B_{jp}.

In order to measure tracking objects bounding box set B_{track}. First, we fuse the detection-based bounding box set B_{det} and the joint points-based bounding box set B_{jp} to generate the all candidates bounding box set B_{can} of current frame. Second, we sort all the bounding boxes according to the confidence and output the bounding box B_{max} with the maximum confidence as tracking objects. Then, we re-assign the confidence of remaining bounding boxes as:

$$CB_{cani} = \begin{cases} CB_{cani} & IoU_{mi} < T_{IoU} \\ CB_{cani}(1 - IoU_{mi}) & IoU_{mi} > T_{IoU} \end{cases} \tag{4}$$

where CB_{cani} indicates the confidence of ith bounding box B_{cani} in candidates bounding box set B_{can}, IoU_{mi} indicates the IoU of bounding box B_{max} and B_{cani}, T_{IoU} indicates the threshold of IoU. Finally, we delete the candidates that confidence less than the confidence threshold T_{con}, until B_{can} is empty.

Algorithm 1 : The Soft-Pose-NMS detection strategy

Input: The primary detection-based bounding box set PB_{det} and the primary joint points-based bounding box set PB_{jp} of current frame in tracking video.

Output: Tracking objects bounding box set B_{track}=$\{B_{track1},...,B_{tracki}\}$ of the current frame.

Generate detection-based bounding box set $B_{det} = \{B_{det1},...,B_{detj}\}$, $CB_{detj} > T_{detcon}$ (CB_{detj} is confidence of detection-based bounding box B_{detj});
Generate joint points-based bounding box set PB_{jp} by Ep.(1);
// Calculate the coordinates of joint points-based bounding boxes
for each PB_{jpi} in PB_{jp} **do**
    Cluster the joint points of PB_{jpi} into two groups;
    Calculate the width W_{PBjpi} for PB_{jpi} by Ep.(2);
end for
$B_{jp} = PB_{jp}$
for each B_{jpi} in B_{jp} **do**
    Calculate the confidence CB_{jpi} for PB_{jpi} by EP.(3);
end for
$B_{can} = B_{det} \cup B_{jp}$;
$B_{track} \leftarrow \{\}$
while B_{can} is not empty **do**
    B_{can} = Sort(B_{can})
    $B_{max} = B_{can}[0]$
    B_{track}.append(B_{mix})
    $B_{can} = B_{can}$ - B_{mix}
    for each B_{cani} in B_{can} **do**
        Update confidence of bounding box in B_{can} by Ep.(4);
        if $CB_{cani} < T_{con}$ **then**:
            delete B_{cani}
        end if
    end for
end while
return B_{track};

3.2 Feature extraction with Dual Self-Attention Network

Extracting more discriminative appearance feature is the critical component of calculating accurate similarity scores. Moreover, the challenge is that object and tracklet images may undergo occlusion and noise in the tracking scene. To alleviate such issues, we design a Dual Self-Attention Network (DSAN) with self-attention mechanisms. Fig. 3 illustrates the architecture of our network.

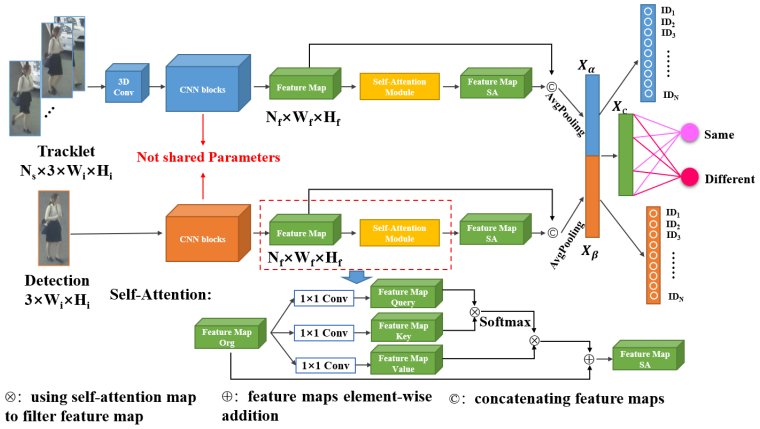

Fig. 3. The architecture of the proposed DSAN. It contains two branches. Given an image of tracking object bounding box and sequence of object tracklet images as inputs. The network extracts the detection and tracklet self-attention feature maps and predicts the probability that the detection and the tracklet are the same object by the combined feature map X_c.

In this work, we use the DenseNet-101[12] as backbone network and introduce the self-attention mechanism to extract tracking object and tracklet feature map. The self-attention mechanism can enlarge the receptive field and get contextual information which enables the network to pay more attention to the object area in the detection and tracklet images. We convolve the tracklet image in the temporal direction by the 3D convolutional layer to exploit the temporal feature of the object. The self-attention map is applied to the feature maps from the last convolutional layer of the DenseNet-101 to compute the self-attention feature map. We apply the detection self-attention feature map X_α and tracklet self-attention feature map X_β for re-identification training and combined feature X_c for binary classifier training to predict whether detection and tracklet are the same object. Furthermore, we will apply the similarity probability P_{same} that predicted by the network to calculate the similarity score between the detection and trajectory.

To infer the self-attention maps of the detection and tracklet, we transform the backbone network feature maps into query feature map f_q, key feature map f_k and value feature map f_v respectively. After that, we use the feature map f_q and f_k to calculate the attention map as the following formula:

$$\beta_{i,j} = \frac{exp(S_{ij})}{\sum_{i=1}^{N} exp(S_{ij})}, S_{ij} = f_q(x_i)^T f_k(x_j) \tag{5}$$

where $\beta_{i,j}$ indicates the attention value of the other jth position in the image on the ith pixel. Then we multiply $\beta_{i,j}$ with f_v to get the self-attention masked feature map f_{org}^{att} that weight by the self-attention map, where:

$$f_{org}^{att} = \sum_{i=1}^{N} \beta_{ij} f_v \tag{6}$$

Additionally, we add the feature map f_{org}^{att} and f_{org}. Therefore the final self-attention feature map f_{sa} is given by:

$$f_{sa} = \theta f_{org}^{att} + f_{org} \tag{7}$$

where θ is a learnable scalar, to gradually emphasize the importance of self-attention feature map.

The training objective of each feature map in DSAN can be modelled as a multi-task training. The joint objective can be written as a weighted linear sum of losses:

$$L_{total} = \alpha L_{sig} + (1 - \alpha) L_{seq} + \beta L_{same} \tag{8}$$

where L_{sig} and L_{seq} are used for re-id training and calculated by the cross-entropy loss function. L_{same} is used for the binary classification training and applying the contrastive loss to calculate. α and β are loss weights. We utilize the ground-truth bounding boxes and objects identity provided in the MOT16 training set to generate detection images and object trajectories for training the network.

3.3 Data Association and Trajectory Management

For data association, we calculate the similarity score between the detection and tracklet feature map firstly, by the following formula:

$$S_{dt} = w_1 dist(f_\alpha, f_\beta) + w_2 P_{same} \tag{9}$$

where w_1 and w_2 are similar score weights, S_{dt} is the final similar score of detection and tracklet. Then tracker generates affinity matrix with the similar scores. Meanwhile, we apply the Hungarian algorithm and affinity matrix to associate the detection and tracklet. Last, the tracker associates the remaining detection with unassociated tracklet based on IoU between detection and tracklets, with a threshold T_{IoUa}. For trajectory management, we initial the trajectory for detection, which is not associated with any trajectory in any of the first T_{init} frames. Trajectories are terminated if they are not associated for T_{term} frames.

4 Experiments

4.1 Implementation Details

To validate the effectiveness of the proposed online tracking approach, we design experiments on popular MOT datasets, MOT16 and MOT17[18]. We employ Pifpaf in [15] to estimate the objects pose information, and use SDP[28] detection

results that officially provided by MOT16 and MOT17 as the object detection results. We set T_{IoU}=0.95 and T_{con} = 0.5 for filtering repetitive bounding box to generate the tracking object set B_{track} and select 5 observations from the 20 most recent frames as tracklet input for DSAN. We set T_{IoUa}=0.7 for data association. For trajectory management, we set the threshold T_{init}=3 for trajectory initialization and T_{term}=10 for trajectory termination.

4.2 Performance on MOT Benchmark Datasets

In order to measure the accuracy of tracking results, we adopt multiple metrics used in the MOT benchmark[2] to evaluate the proposed tracking method, including Multiple Object Tracking Accuracy (MOTA), ID F1 score (IDF, the ratio of correct detections over the average number of ground-truth and computed detections), MT (the ratio of Mostly Tracked objects), Ml (the ratio of Mostly Lost objects), the number of False Negatives (FN), the number of False Positives (FP), the number of ID Switches (IDS), the number of fragments (Frag). Table 1 and Table 2 present the tracking performance on the MOT16 and MOT17 datasets, respectively.

Table 1. Tracking performance on MOT16 dataset. The arrow each metric indicates that the higher (↑) or lower (↓) value is better.

Method	Mode	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓	Frag↓
EDMT[4]	Batch	45.3	47.9	17.0%	39.9%	11122	87899	639	946
QuadMOT[21]	Batch	44.1	38.3	14.6%	44.9%	6388	94775	745	1096
LMP[23]	Batch	48.8	51.3	18.2%	40.1%	6654	86245	481	595
DMAN[31]	Online	46.1	54.8	17.4%	42.7%	7909	89874	744	1616
Tracktor++[1]	Online	56.2	54.9	20.7%	35.8%	**2394**	76844	617	1068
CNNMTT[17]	Online	65.2	62.2	32.4%	21.3%	6578	55896	946	2283
TrctrD16[26]	Online	54.8	53.4	19.1%	37.0%	2955	78765	645	1515
RAR16wVGG[9]	Online	63.0	63.8	**39.9%**	22.1%	13663	53248	482	1251
MPNTrack[3]	Online	58.6	61.7	27.3%	34.0%	4949	70252	354	**684**
Tube_TK_POL[19]	Online	66.9	62.2	39.0%	**16.1%**	11544	47520	1236	1444
Ours	Online	**67.7**	**66.4**	37.9%	18.6%	11453	**42494**	**334**	902

Table 2. Tracking performance on MOT17 dataset.

Method	Mode	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓	Frag↓
EDMT[4]	Batch	50.0	51.3	21.6%	36.3%	32279	247297	2264	3260
MHT_DAM[14]	Batch	50.7	47.2	20.8%	36.9%	22875	252889	2314	2865
Tube_TK_POI[19]	Online	63.0	58.6	31.2%	**19.9%**	27060	177483	4137	5727
CTTrack17[30]	Online	**67.8**	64.7	34.6%	24.6%	**18498**	**160332**	3039	6102
Ours	Online	67.3	**65.9**	**37.9%**	20.7%	20574	195176	**2031**	**2681**

Quantitative results and comparison with the other tracking methods are shown in Table 1 and Table 2. As shown in Table 1, our tracking method achieves a comparable MT, ML, FP, Frag score and performs favourably against the state-of-the-art methods in terms of MOTA, IDF1, FN and IDs on the MOT16 dataset. Our tracker upgrades MOTA to 67.7, IDF1 to 66.4 and reduces FN to 42494, IDs to 334. Meanwhile, our tracker achieves the best performance in IDF1 and IDs among online and batch methods, demonstrating the merits of our tracker in object identity matching and the stability of multi-object tracking. MOTA and FN correspond to the object detection capability. Therefore, the improvement of MOTA and FN demonstrates the merits of our Soft-Pose-Nms detection strategy in object locating for MOT. Similarly, Table 2 shows that our tracker outperforms existing online trackers on half of the metrics and achieves the best performance in terms of IDF1, MT, IDs and Frag on the MOT17 dataset.

In addition, as shown in Table 1, our tracker has a high FP. According to this phenomenon, the detection strategy proposed in this paper is combining the object detection results and pose estimation results. This can alleviate unreliable detection and complement missing object. Second, we find that only the moving pedestrians are recorded as tracking object ground-truth in MOT16 and MOT17. Nevertheless the detection strategy proposed in this paper can detect and track these small-scale pedestrians, occluded pedestrians, stationary pedestrians and pedestrians who are not recorded as tracking objects. Therefore, our detection strategy will cause the phenomenon of high FP, and the similar situation exists in [4, 5] too. This phenomenon also reflects the effectiveness of the detection strategy proposed in this paper.

4.3 Ablation studies

In order to verify the effectiveness of the proposed detection strategy and evaluate its contribution, we use different object detection results and conduct ablation experiments in the MOT16 dataset. We choose Mask R-CNN[11] and SDP[28] as bounding box-based object detection method and PifPaf[15] as pose estimation method. In addition, to exclude the disturbance of other factors, we use DeepSORT[25], the more common method in MOT, for tracking.

Table 3. Evaluation tracking results on MOT16 dataset with different detection method. Ours (M+P) indicates combining Mask R-CNN detection results and PifPaf pose estimation results. Ours (S+P) indicates combining SDP detection results and PifPaf pose estimation results.

Method	MOTA ↑	IDF1↑	MT↑	ML↓	FP↓	FN ↓	IDs↓
Mask R-CNN[11]	40.2	52.6	21.5%	26.9%	14266	51234	528
SDP[28]	60.7	62.4	31.3%	20.9%	**3417**	38041	462
PifPaf[15]	37.6	51.8	14.5%	32.1%	14652	53729	537
Ours(M+P)	43.8	55.8	22.6%	22.8%	15226	46270	511
Ours(S+P)	**64.3**	**65.9**	**34.4%**	**15.6%**	5115	**35732**	**433**

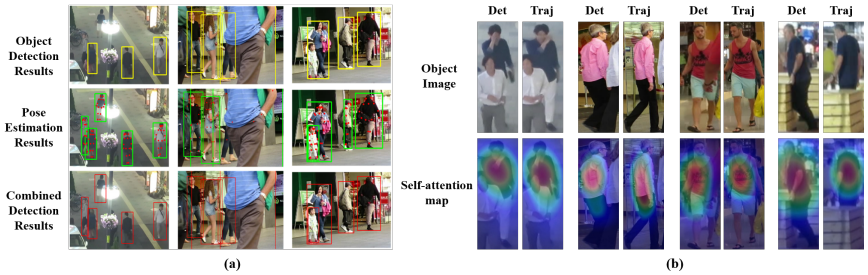

Fig. 4. Visualization of pose-guided object locating results and self-attention maps.

The experiment results are shown in Table 3. The comparison between our detection strategy and object detection methods and pose estimation method confirms that our detection strategy performs best. Our detection strategy improves 3.6 in MOTA, 3.5 in IDF1, 3.1% in MT with the second best detection method and effectively reduced FN demonstrating the merits of our detection strategy in locating the objects. By combining object detection results and pose estimation results, our detection strategy can reduce unreliable detections and alleviate missing detections, as shown in Fig. 4(a).

Table 4. Evaluation results on MOT16 with different feature representations.

Method	MOTA↑	IDF1↑	IDs↓
DenseNet121[12]	61.7	63.1	548
PCB[22]	62.6	64.3	482
Ours(TBSAN)	**65.7**	**68.7**	**455**

To demonstrate the contribution of the proposed DSAN network in our method, we compare representations learned by DSAN with PCB, DenseNet-121. Moreover, we use SDP[28] detection result, provide by MOT16 officially, for tracking. The experiment results are shown in Table 4. It can be seen that the IDF1, IDs and MOTA of DSAN are better than other methods. Our tracker upgrades MOTA to 65.7, IDF1 to 68.7 and reduces IDs to 455, which demonstrates the effectiveness of our feature extraction network.

Fig. 4(b) shows the visualization results of the self-attention feature map from DSAN. In Fig. 4(b), each group consists of four images. The top row of each group shows an image pair from the same object, while the bottom row presents corresponding self-attention feature maps. It can be seen that our self-attention feature map focus more explicitly on object regions and suppress noise and occlusion, which enhances the power of extracting discriminative features.

5 Conclusions

This paper presents a detection strategy and a feature extraction network to improves two main components of most online trackers, detection and feature extraction. The tracker locates joint points of objects with pose estimation results. Then generating optimal object bounding boxes by proposed Soft-Pose-NMS method, which also helps alleviate typical difficulties in tracking such as occlusion handling and track drifting. In this paper, the tracker learns the discriminative self-attention maps from the MOT dataset with the Self-Attention mechanism to calculate more accurate similarity scores. The experimental results on MOT Challenge datasets demonstrated that the proposed tracking framework leads to competitive performance improvement through extensive experiments.

Acknowledgment

This work was supported by National Key R&D Program of China (Grant No.2018YFB2100603) and National Natural Science Foundation of China (Grant No.61872024). The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestion.

References

1. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: Proceedings of the IEEE 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. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6247–6257 (2020)
4. Chen, J., Sheng, H., Zhang, Y., Xiong, Z.: Enhancing detection model for multiple hypothesis tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 18–27 (2017)
5. Chen, L., Ai, H., Zhuang, Z., Shang, C.: Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2018)
6. Choi, W.: Near-online multi-target tracking with aggregated local flow descriptor. In: Proceedings of the IEEE international conference on computer vision. pp. 3029–3037 (2015)
7. Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., Yu, N.: Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4836–4845 (2017)
8. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334–2343 (2017)

9. Fang, K., Xiang, Y., Li, X., Savarese, S.: Recurrent autoregressive networks for online multi-object tracking. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 466–475. IEEE (2018)
10. Fragkiadaki, K., Shi, J.: Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement. In: CVPR 2011. pp. 2073–2080. IEEE (2011)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
13. Keuper, M., Tang, S., Andres, B., Brox, T., Schiele, B.: Motion segmentation & multiple object tracking by correlation co-clustering. IEEE transactions on pattern analysis and machine intelligence **42**(1), 140–153 (2018)
14. Kim, C., Li, F., Ciptadi, A., Rehg, J.M.: Multiple hypothesis tracking revisited. In: IEEE International Conference on Computer Vision (2015)
15. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11977–11986 (2019)
16. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1487–1495 (2019)
17. Mahmoudi, N., Ahadi, S.M., Rahmati, M.: Multi-target tracking using cnn-based features: Cnnmtt. Multimedia Tools and Applications **78**(6), 7077–7096 (2019)
18. Milan, A., Leal-Taixe, L., Reid, I., Roth, S., Schindler, K.: MOT16: A Benchmark for Multi-Object Tracking. arXiv e-prints arXiv:1603.00831 (Mar 2016)
19. Pang, B., Li, Y., Zhang, Y., Li, M., Lu, C.: Tubetk: Adopting tubes to track multi-object in a one-step training model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6308–6318 (2020)
20. Sanchez-Matilla, R., Poiesi, F., Cavallaro, A.: Online multi-target tracking with strong and weak detections. In: European Conference on Computer Vision. pp. 84–99. Springer (2016)
21. Son, J., Baek, M., Cho, M., Han, B.: Multi-object tracking with quadruplet convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5620–5629 (2017)
22. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 480–496 (2018)
23. Tang, S., Andriluka, M., Andres, B., Schiele, B.: Multiple people tracking by lifted multicut and person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3539–3548 (2017)
24. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: Mots: Multi-object tracking and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7942–7951 (2019)
25. 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)
26. Xu, Y., Osep, A., Ban, Y., Horaud, R., Leal-Taixé, L., Alameda-Pineda, X.: How to train your deep multi-object tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6787–6796 (2020)

27. Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., Dai, Q.: Stat: spatial-temporal attention mechanism for video captioning. IEEE transactions on multimedia **22**(1), 229–241 (2019)
28. Yang, F., Choi, W., Lin, Y.: Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2129–2137 (2016)
29. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning. pp. 7354–7363. PMLR (2019)
30. Zhou, X., Koltun, V., Krhenbühl, P.: Tracking objects as points. arXiv arXiv:2004.01177 (2020)
31. 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)