MULTI-OBJECT TRACKING WITH
TRACKED OBJECT BOUNDING BOX ASSOCIATION

Nanyang Yang, Yi Wang and Lap-Pui Chau

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
yang0526@e.ntu.edu.sg, wang1241@e.ntu.edu.sg, elpchau@ntu.edu.sg

ABSTRACT

The CenterTrack tracking algorithm achieves state-of-the-art tracking performance using a simple detection model and single-frame spatial offsets to localize objects and predict their associations in a single network. However, this joint detection and tracking method still suffers from high identity switches due to the inferior association method. To reduce the high number of identity switches and improve the tracking accuracy, in this paper, we propose to incorporate a simple tracked object bounding box and overlapping prediction based on the current frame onto the CenterTrack algorithm. Specifically, we propose an Intersection over Union (IOU) distance cost matrix in the association step instead of simple point displacement distance. We evaluate our proposed tracker on the MOT17 test dataset, showing that our proposed method can reduce identity switches significantly by 22.6% and obtain a notable improvement of 1.5% in IDF1 compared to the original CenterTrack's under the same tracklet lifetime. The source code is released at https://github.com/Nanyangny/CenterTrack-IOU.

Index Terms— Multi-object tracking, joint detection and tracking, data association

1. INTRODUCTION

Multi-object tracking (MOT) is a popular topic in computer vision due to its wide application in areas such as transportation and elderly care. Recent progress on joint detection and tracking technique has drawn much research attention in MOT problems. MOT is a task to estimate trajectories for objects of interest through space and time [1]. The rapid development of deep learning has advanced the research on MOT.

MOT is often addressed by the tracking-by-detection paradigm which consists of two parts [2]. First, an object detection algorithm that outputs detection results in the form of bounding boxes location in every frame; then an association algorithm is used to link up the newly detected objects with the existing tracks based on spatial information or extracted re-identification (re-ID) features, or both. Most of the existing MOT solvers use two separate models to perform the two steps respectively. Although there has been significant development

in object detection [3, 4, 5] and re-ID [6, 7] separately to enhance the overall tracking performance, those methods hardly achieve real-time inference speed due to the slow and complex association methods [8, 9, 10] and separate learned models without shared features.

Recent research in simultaneous detection and tracking method [11, 12, 13] provides another viable research direction for MOT tasks. Under this approach, existing detectors are converted into trackers and both tasks are combined in the same framework. Two tasks now share the same set of low-level features, therefore no need for re-computation.

CenterTrack [13], one of the state-of-the-art trackers, adopts the idea of simultaneous detection and tracking methods with point-based detection. In CenterTrack, each object is represented by the center point of its bounding box. This center point is tracked through time. Objects in a frame are represented by a heatmap of points. CenterTrack takes in heatmaps of two consecutive frames and trains the model to output an offset vector from the current object center to its center in the previous frame. A simple greedy matching is performed using the distance between the predicted offset and detected center point in the previous frame to associate object identities. This tracking-conditioned detection framework replaces the need for a motion model [13], which reduces the need for extra computation. However, CenterTrack relies on center displacement offset to associate objects in adjacent frames only, which is not enough to provide robust association ability especially when occlusions occur. Additionally, long-range tracklet association is not explored in [13]. Therefore, in this paper, we propose to integrate the prediction of the tracked object bounding box to the existing CenterTrack tracking model, which enables robust distance cost matrix calculation based on both center displacement and tracked object bounding box prediction to associate objects through long-range tracklets, shown in Figure 1. Our main contributions in this project are:

- Incorporate tracked object bounding box prediction to CenterTrack using robust cost matrix calculation in object association.

- Evaluate the proposed method on MOT17 dataset to obtain a significant reduction in identity switches (IDs)
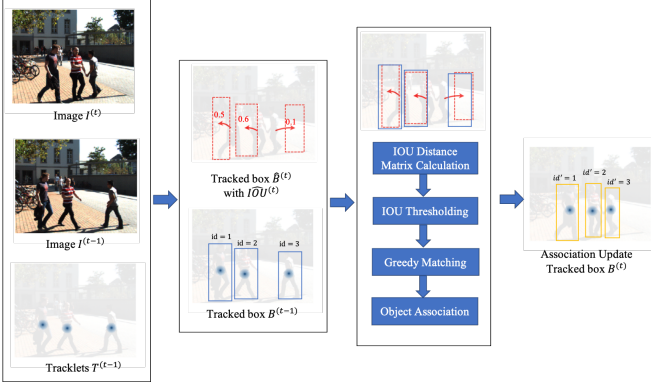
Fig. 1: Overview of our proposed method: A network is configured to predict tracked bounding box and IOU based on the original CenterTrack model. IOU distance cost matrix is calculated between the tracked object bounding boxes $\hat{B}^{(t-1)}$ (in blue) at the previous frame and tracked object bounding boxes prediction $\hat{B}^{(t)}$ (in dotted red) from the current frame, followed by IOU filtering. Finally, a simple greedy matching algorithm is used to associate objects in the current frame.

and notable improvements in accuracy score only with additional two output branches.

This paper is organized as follows: In Section 2 and 3, the detailed methodology of the baseline baseline CenterTrack model and proposed tracking method are introduced. The experiment details and results are described in Section 4. We conclude this paper in Section 5.

2. BASELINE CENTERTRACK ALGORITHM

CenterTrack takes three inputs, the current frame, the previous frame, and a heatmap rendered from tracked object centers. Then, the model outputs a center detection heatmap for the current frame, the bounding box size map, and a center offset map.

Heatmap Representation CenterTrack is built on CenterNet [4]. Objects are presented by the center point of their bounding boxes. For each center $\mathbf{p}$ of class C in each frame, it is rendered as a Gaussian-shaped peak into a heatmap $Y \in [0, 1]^{\frac{W}{R} \times \frac{H}{R} \times C} = R(\{\mathbf{p}_0, \mathbf{p}_1, ...\})$, the rendering function at position $\mathbf{q} \in \mathbb{R}^2$ is defined as:

$$R_q(\{\mathbf{p}_0, \mathbf{p}_1, ...\}) = \max_i \exp\left(-\frac{(\mathbf{p}_i - \mathbf{q})^2}{2\sigma_i^2}\right) \quad (1)$$

where the Gaussian kernel σ_i is a function of the object size [14].

Tracking-conditioned Detection In CenterTrack, two frames are passed to the network: the current frame $I^{(t)} \in \mathbb{R}^{W \times H \times 3}$ and the prior frame $I^{(t-1)} \in \mathbb{R}^{W \times H \times 3}$. This allows the model to estimate the change between the frames

and potentially reason about the occluded objects at time t from visual information at time $t-1$. To help the model detect the objects in the current frame, a heatmap rendered from prior detections $\hat{Y} \in [0, 1]^{\frac{W}{R} \times \frac{H}{R} \times C}$ and size map $\hat{S} \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$, are used to feed the model as well, where R is a downsampling factor and C is the number of classes. To reduce false positive detections, local maxima (peaks) in a 3×3 region are used and only peaks with a confidence score greater than a threshold τ are rendered. The object sizes are extracted from the size map to calculate the objects' bounding boxes.

Object Association CenterTrack predicts a center displacement as two output channels $\hat{D}^{(t)} \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$. For each detected object at location $\hat{\mathbf{p}}$, the predicted $\hat{D}^{(t)}_{\hat{p}^{(t)}}$ shows the difference of object center in the current frame $\hat{p}^{(t)}$ and the previous frame $\hat{p}^{(t-1)}$, $\hat{D}^{(t)}_{\hat{p}^{(t)}} = \hat{p}^{(t)} - \hat{p}^{(t-1)}$. With this center offset prediction, the object center location in the previous frame can be easily tracked. For each detection at $\hat{p}$, a simple greedy matching algorithm is used to associate it with the closet unmatched prior detection at position $\hat{p} - \hat{D}_{\hat{p}}$ [13]. If the distance is more than the bounding box size of objects in the adjacent frames, a new tracklet is spawn. This association method uses displacement distance only.

Objective Functions Focal loss is used as the training objective function to learn the heatmap [15, 14]:

$$L_k = \frac{1}{N} \sum_{xyc} \begin{cases} (1 - \hat{Y}_{xyc})^\alpha \log(\hat{Y}_{xyc}) & \text{if } Y_{xyc} = 1, \\ (1 - Y_{xyc})^\beta (\hat{Y}_{xyc})^\alpha \log(1 - \hat{Y}_{xyc}) & \text{otherwise} \end{cases}$$

where N is the number of objects, Y_{xyc} a ground-truth heatmap corresponding to the annotated centers, and $\alpha = 2$ and $\beta = 4$ are hyperparameters of the function.

The regression objective functions for size and displacement use the L1 loss [13]:

$$L_{size} = \frac{1}{N} \sum_{i=1}^{N} |\hat{S}_{\mathbf{p}_i} - s_i|, \quad (2)$$

where s_i is the bounding box size of the i-th object at location $\mathbf{p}_i$.

$$L_{off} = \frac{1}{N} \sum_{i=1}^{N} |\hat{D}_{\mathbf{p}_i^{(t)}} - (\mathbf{p}_i^{(t-1)} - \mathbf{p}_i^{(t)})|, \quad (3)$$

where $\mathbf{p}_i^{(t)}$ and $\mathbf{p}_i^{(t-1)}$ are tracked centers.

3. PROPOSED METHOD

Our proposed tracked object bounding box prediction (CenterTrack++) is built upon the CenterTrack tracking method in [13]. The original CenterTrack association method only uses single-frame tracked center offsets to associate objects through time, this method may fail in long-range tracklets or when occlusions occur as the occluded object's identity tends to assign to the object that occludes it. Under those
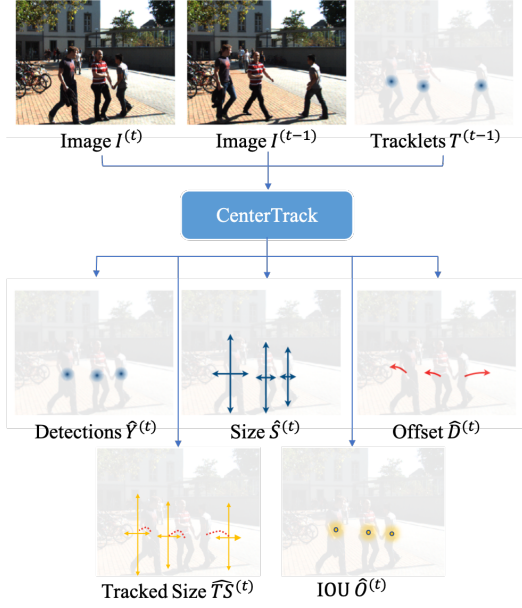
Fig. 2: Illustration of proposed CenterTrack++. Only two new output branches (Tracked Size and IOU) are added to the original CenterTrack [13] framework.

cases, a single center displacement is not sufficient to obtain accurate tracking results. Therefore, additional size prediction is proposed to allow the tracking algorithm to better deal with the identity association by taking the overlapping of prior tracked object bounding boxes and predicted tracked bounding boxes into consideration. With two additional outputs on top of the existing tracking-conditioned CenterTrack method, the number of IDs can be easily reduced and tracking accuracy will then be improved, resulting in a better MOT tracker.

3.1. Tracked Object Bounding Box and IOU Prediction (CenterTrack++)

Inspired by the idea of IOU distance in SORT [16] and IOU-Tracker [17], IOU information is used in the association. To enable IOU distance calculation, prediction of tracked object bounding box in the prior frame with long-range tracklet lifetime is as shown in Figure 2.

Tracked Object Bounding Box Prediction. There are two possible ways to predict tracked object bounding box in the previous frame based on the current frame.

- **Tracking_wh** Similar to the learning of center offset in the original CenterTrack model, the width and height difference $\hat{TS} \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$ of the object's bounding box in the current frame and the previous frame is learned. This is used to predict the bounding box of the tracked object in the prior frame. For each detected object at location $\hat{\mathbf{p}}$, $\hat{TS}_{\hat{p}^{(t)}}^{(t)}$ is the difference of

object size in the current frame $\hat{s}^{(t)} = (\hat{w}^{(t)}, \hat{h}^{(t)})$ and previous frame $\hat{s}^{(t-1)} = (\hat{w}^{(t-1)}, \hat{h}^{(t-1)})$, calculated by $\hat{s}^{(t)} - \hat{s}^{(t-1)}$, where $\hat{w}$ and $\hat{h}$ are width and height of object bounding box at location $\hat{\mathbf{p}}$.

- **Tracking_ltrb** Apart from learning width and height difference, we can also use offsets of the left, top, right, and bottom (ltrb) of the bounding box from the center in the prior frame instead, thus $\hat{TS} \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R} \times 4}$. For each detected object at location $\hat{\mathbf{p}}$, $\hat{TS}_{\hat{p}^{(t)}}^{(t)} = (\hat{x}^{(t-1)} - \frac{\hat{w}^{(t-1)}}{2}, \hat{y}^{(t-1)} + \frac{\hat{h}^{(t-1)}}{2}, \hat{x}^{(t-1)} + \frac{\hat{w}^{(t-1)}}{2}, \hat{y}^{(t-1)} - \frac{\hat{h}^{(t-1)}}{2})$, $\hat{x}$ and $\hat{y}$ are horizontal and vertical coordinates of $\hat{\mathbf{p}}$.

IOU Prediction. To further suppress inaccurate association, the IOU value of the bounding box of the same target in adjacent frames (IOU-adjacent) is learnt to provide a filtering threshold for unlikely associations. It is reasonable to assume that the IOU-adjacent is equal or smaller than IOU between detection in the prior frame and the regressed object bounding box based on the current frame (IOU-tracked). Any IOU-tracked that is smaller than the predicted IOU-adjacent should not be associated. Therefore, we configure the network to learn the IOU-adjacent ($O \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times 1}$). For each detected object at location $\hat{\mathbf{p}}$, $\hat{O}_{\hat{p}_{(t)}}^{(t)} = IOU(b^{(\hat{t-1})}, b^{(\hat{t})})$, where $b^{(\hat{t-1})}$ and $b^{(\hat{t})}$ are tracked bounding box in the previous and current frame at position $\hat{\mathbf{p}}$ respectively. The model is conditioned to reason about how much overlapping of the same object's bounding box between adjacent frames.

Therefore, two additional outputs would be produced by the model under this proposed method. The L1 loss objective function can be applied to them. For IOU prediction:

$$L_{IOU} = \frac{1}{N} \sum_{i=1}^{N} |\hat{O}_{\mathbf{p}_i}^{(t)} - IOU(b_{\mathbf{p}_i}^{(t-1)}, b_{\mathbf{p}_i}^{(t)})|, \quad (4)$$

where $b_{\mathbf{p}_i}^{(t-1)}$ and and $b_{\mathbf{p}_i}^{(t)}$ are tracked ground-truth bounding boxes.

In the case of Tracking_wh approach for tracked object bounding box prediction:

$$L_{tracked\_size} = \frac{1}{N} \sum_{i=1}^{N} |\hat{TS}_{\mathbf{p}_i}^{(t)} - (s_i^{(t-1)} - s_i^{(t)})|. \quad (5)$$

In the case of Tracking_ltrb approach for tracked object bounding box prediction:

$$L_{tracked\_size} = \frac{1}{N} \sum_{i=1}^{N} |\hat{TS}_{\mathbf{p}_i}^{(t)} - (x_i^{(t-1)} - \frac{w_i^{(t-1)}}{2}, \\ y_i^{(t-1)} + \frac{h_i^{(t-1)}}{2}, \\ x_i^{(t-1)} + \frac{w_i^{(t-1)}}{2}, \\ y_i^{(t-1)} - \frac{h_i^{(t-1)}}{2})|. \quad (6)$$

3.2. Association Steps

As explained in [16], IOU distance can implicitly handle short-term occlusions caused by passing targets. Since we can predict the tracked object bounding boxes from the current frame, IOU distance cost matrix can be naturally incorporated into the association process. IOU distance cost is calculated by $1 - IOU(\hat{b}^{(t-1)}, \hat{tb}^{(t)})$, where $\hat{b}^{(t-1)}$ is detected bounding box in the prior frame and $\hat{tb}^{(t)}$ is tracked bounding box prediction from the current frame. If the IOU distance cost is more than $\hat{O}^{(t)}_{\hat{p}^{(t)}_i}$ at the detected point $\hat{p}^{(t)}_i$, we set the corresponding cost to infinity which effectively prevents from unlikely associations. After distance cost matrix computation, a simple greedy matching algorithm is employed to assign object identities.

As displacement and IOU distance cost matrices are now available, we further explored the possibilities of using different combinations and orders of matrices during association. Not only using a single matrix, two matrices can be summed up together to produce a combined matrix for the association. Additionally, we can use two matrices sequentially. Specifically, after one round of a simple greedy matching with one distance cost matrix, a different cost matrix can be used to associate the remaining unmatched detections and tracklets further with another round of greedy matching. Tracking performances of different association methods are reported in section 4 under Ablative Studies.

4. EXPERIMENTS

4.1. Dataset and Metrics

We use MOT17 [18] dataset to train and evaluate the proposed method in our paper. MOT17 contains 7 sequences for training and test respectively. The videos were captured by stationary cameras mounted in high-density scenes with heavy occlusion. Only pedestrians are annotated and evaluated. The video framerate is 25-30 FPS. Since MOT17 does not provide an official validation split, we split each training sequence into halves, the first half for training and the second one for validation in our ablative studies. Our main results are reported on the test set. We followed the same metric used in the original CenterTrack model, the CLEAR metric [19] and Identification F_1 score (IDF1) [20], to evaluate overall tracking accuracy.

4.2. Implementation Details

Our implementation is based on CenterTrack, DLA [21] is used as the network backbone with Adam optimizer [22] at a learning rate of $1.25e - 4$ and a batch size of 16. We use standard data augmentations include horizontal flipping, random resized cropping, and color jittering. For all experiments, we use 70 epochs for the network training. The learning rate

Table 1: Results on MOT17 validation set using the tracking_wh approach.

Association	IDF1↑	MOTA↑	IDs↓	FP↓	FN↓
DIS	69.2	66.2	219	3.9	29.5
IOU	**71.1**	66.7	204	**3.6**	29.3
Combined	70.9	66.2	233	3.9	29.6
DIS→IOU	70.0	66.2	218	3.9	29.5
IOU→DIS	69.8	**66.8**	**185**	**3.6**	**29.2**

Table 2: Results on MOT17 validation set using the tracking_ltrb approach.

Association	IDF1↑	MOTA↑	IDs↓	FP↓	FN↓
DIS	69.2	66.2	219	3.9	29.5
IOU	**72.4**	**66.7**	191	3.8	29.2
Combined	70.8	66.5	236	3.8	29.3
DIS→IOU	70.5	66.6	202	3.8	29.2
IOU→DIS	71.4	**66.7**	**166**	3.8	29.2

drops by $1/10$ at the 60th epoch. We train and test the proposed model with three GTX 1080 Ti GPUs.

The input size is resized to 960×544, with downsampling $R = 4$. Followed the recommended parameters in CenterTrack, we also use random false positive ratio $\lambda_{fp} = 0.1$ and random false negative ratio $\lambda_{fp} = 0.4$ to generate noises in the dataset to train a robust tracking-conditioned object detector. Similarly, we only output tracklets with confidence of $\theta = 0.4$ and above and render heatmap with a threshold $\tau = 0.5$. The network is pre-trained on CrowdHuman dataset [23] before training on MOT17 dataset. However, unlike original CenterTrack, we use long-range tracklets, tracklet lifetime = 30, discarding the unmatched tracklets only after 30 frames.

4.3. Ablative Studies

As described in the previous section, we experimented cost matrix with different combinations and orders: 1. Displacement only (DIS), used in original CenterTrack; 2. IOU only (IOU); 3. IOU and displacement (Combined); 4. IOU first followed by displacement (IOU→DIS); 5. Displace first followed by IOU (DIS→IOU).

The result of IOU only association method in Table 1 and 2 confirm the benefits of using additional tracked object bounding box prediction to reduce IDs of object tracking, both IDF1 improves compared to the baseline DIS CenterTrack tracking algorithm, with a significant 3.2% IDF1 improvement from baseline method using tracking_ltrb approach in Table 2. Since our proposed method focuses on improving CenterTrack's association ability, not detection capability, a small improvement in MOTA is expected and observed as MOTA metrics MOTA penalizes detection errors and IDs while IDF1 focuses on the tracking accuracy of detected objects. [24].

Comparing the overall performance between tracking_wh

4

and tracking_ltrb approach to predict tracked object bounding box, it is observed that the use of ltrb offsets is more effective to regress the width and height of the bounding box of the tracked centers from the current detected centers compared to just learning from the size offset between adjacent frames.

However, the idea of sequential matching using different matrices yields little or no improvement in association accuracy compared to single IOU matching, implying that single IOU matching is sufficient to provide accurate tracking results. Additionally, the idea of a combined distance matrix does not necessarily improve tracking accuracy neither, this could be due to the addition of two distance matrices with different scales, which can be considered as noises to corrupt the associating power of the other matrix.

CenterTrack++ is more robust in object tracking compared to CenterTrack in cases when occlusions occur or objects exit the frame. Figure 3 and 4 demonstrate CenterTrack++'s ability to track objects accurately during those cases while CenterTrack fails.
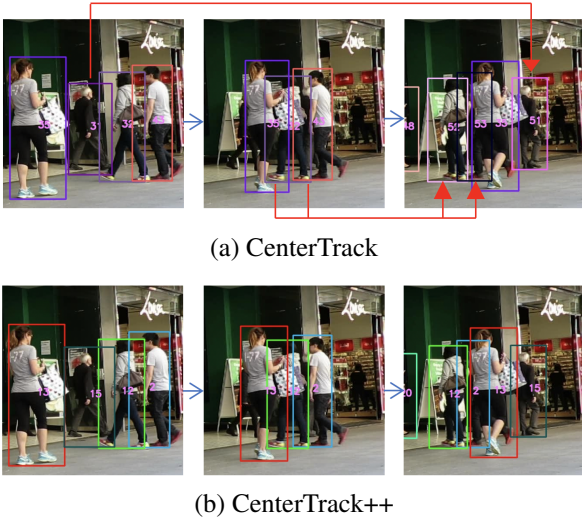

(a) CenterTrack

(b) CenterTrack++

Fig. 3: Comparison of the tracking results on the sequence "MOT17-09" validation when short-term occlusions occur with each arrow representing 10-frame interval.

4.4. Test Result

From the ablative studies, we found out that the use of IOU distance in the association step under the tracking_ltrb approach yields the best performance. We adopt the best method and evaluate its performance on MOT17 test data. The results are shown in Table 3. It is shown that our method can reduce the IDs significantly by 22.6% and obtain a notable improvement of 1.5% in IDF1 score compared to the Original CenterTrack under the same tracklet lifetime. Our method obtains the best performance based on MOTA, IDF1 and IDs evaluation metric among trackers only using spatial features

for association. Compared with FairMOT [2] that employs re-identification based on additional appearance features, our method still obtains better IDs score (2352 vs. 3303).

Table 3: Comparison of the state-of-the-art methods under "private detector" protocol. Note: S=Spatial, A=Appearance.

Tracker	Association Features	MOTA↑	IDF1↑	IDs↓
TubeTK[25]	S	63	58.6	4137
CenterTrack[13]	S	67.8	64.7	3039
Ours	S	**68.1**	**66.2**	**2352**
SST[26]	A	52.4	49.5	8431
CTrackerV1[27]	S+A	66.6	57.4	5529
DEFT[28]	S+A	66.6	65.4	**2823**
FairMOT[2]	S+A	**73.7**	**72.3**	3303

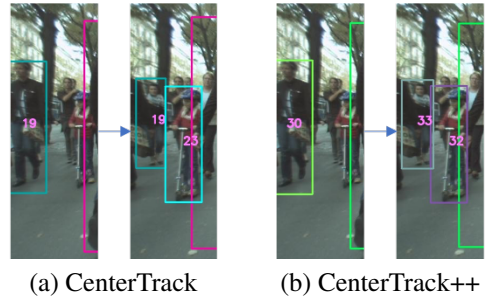(a) CenterTrack   (b) CenterTrack++

Fig. 4: Comparison of the tracking results on the sequence "MOT17-02" validation set when the leftmost person exited the frame. CenterTrack assigns the same ID to different person(ID:19), while CenterTrack++ does not.

5. CONCLUSION

In this paper, we propose tracked object bounding box and overlapping prediction outputs onto the CenterTrack tracking algorithm, which reduces the IDs and improves overall tracking accuracy. The extra prior tracked object bounding box and overlapping prediction enable the use of the IOU distance matrix to associate objects across frames more accurately. Experiments on MOT17 test dataset under private protocol show that our proposed method achieves the best performance among the trackers only using spatial features in the association.

6. REFERENCES

[1] Jianbo Shi and Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 1994, pp. 593–600.

[2] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu, "Fairmot: On the fairness

of detection and re-identification in multiple object tracking," *arXiv preprint arXiv:2004.01888*, 2020.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[4] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl, "Objects as points," 2019.

[5] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," 2018.

[6] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," 07 2018.

[7] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian, "Person re-identification in the wild," 2017.

[8] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krähenbühl, Trevor Darrell, and Fisher Yu, "Joint monocular 3d vehicle detection and tracking," 2019.

[9] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, July 2017, pp. 3701–3710, IEEE Computer Society.

[10] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4705–4713.

[11] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 941–951.

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3057–3065.

[13] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl, "Tracking objects as points," *European Conference on Computer Vision (ECCV)*, 2020.

[14] Hei Law and Jia Deng, "Cornernet: Detecting objects as paired keypoints," 2019.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," 2018.

[16] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464–3468.

[17] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information,"

in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–6.

[18] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[19] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 246309, May 2008.

[20] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," 2016.

[21] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell, "Deep layer aggregation," 2019.

[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.

[23] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun, "Crowdhuman: A benchmark for detecting human in a crowd," 2018.

[24] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *Computing Research Repository*, vol. abs/1609.01775, 2016.

[25] Bo Pang, Yizhuo Li, Yifan Zhang, Muchen Li, and Cewu Lu, "Tubetk: Adopting tubes to track multi-object in a one-step training model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[26] Shijie Sun, Naveed Akhtar, Huansheng Song, Ajmal Mian, and Mubarak Shah, "Deep affinity network for multiple object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, 07 2019.

[27] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu, "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," 2020.

[28] Mohamed Chaabane, Peter Zhang, Ross Beveridge, and Stephen O'Hara, "Deft: Detection embeddings for tracking," *arXiv preprint arXiv:2102.02267*, 2021.