

# DiffusionTrack: Diffusion Model For Multi-Object Tracking

Run Luo<sup>123</sup>, Zikai Song<sup>3\*</sup>, Lintao Ma<sup>3</sup>, Jinlin Wei<sup>34</sup>, Wei Yang<sup>3</sup>, Min Yang<sup>12\*</sup>

<sup>1</sup>Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

<sup>2</sup>University of Chinese Academy of Sciences

<sup>3</sup>Huazhong University of Science and Technology

<sup>4</sup>University of California, Santa Barbara

{r.luo,min.yang}@siat.ac.cn, {skyesong, mltdml, weiyangcs}@hust.edu.cn, jinlinwei@ucsb.edu

## Abstract

Multi-object tracking (MOT) is a challenging vision task that aims to detect individual objects within a single frame and associate them across multiple frames. Recent MOT approaches can be categorized into two-stage tracking-by-detection (TBD) methods and one-stage joint detection and tracking (JDT) methods. Despite the success of these approaches, they also suffer from common problems, such as harmful global or local inconsistency, poor trade-off between robustness and model complexity, and lack of flexibility in different scenes within the same video. In this paper we propose a simple but robust framework that formulates object detection and association jointly as a consistent denoising diffusion process from paired noise boxes to paired ground-truth boxes. This novel progressive denoising diffusion strategy substantially augments the tracker’s effectiveness, enabling it to discriminate between various objects. During the training stage, paired object boxes diffuse from paired ground-truth boxes to random distribution, and the model learns detection and tracking simultaneously by reversing this noising process. In inference, the model refines a set of paired randomly generated boxes to the detection and tracking results in a flexible one-step or multi-step denoising diffusion process. Extensive experiments on three widely used MOT benchmarks, including MOT17, MOT20, and DanceTrack, demonstrate that our approach achieves competitive performance compared to the current state-of-the-art methods. Code is available at <https://github.com/RainBowLuoCS/DiffusionTrack>.

## 1 Introduction

Multi-object Tracking is one of the fundamental vision tasks with applications ranging from human-computer interaction, surveillance, autonomous driving, etc. It aims at detecting the bounding box of the object and associating the same object across consecutive frames in a video sequence. Recent MOT approaches can be categorized into two-stage tracking-by-detection (TBD) methods and one-stage joint detection and tracking (JDT) methods. TBD methods detect the bounding boxes of the objects within a single frame using a detector and associate the same object cross frames by employing supplementary trackers. These trackers encompass a spectrum of techniques, such as motion-

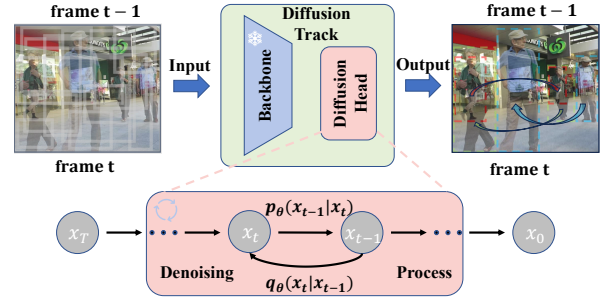


Figure 1: DiffusionTrack formulates object association as a denoising diffusion process from paired noise boxes to paired object boxes within two adjacent frames  $t - 1$  and  $t$ . The diffusion head receives the two-frame image information extracted by the frozen backbone and then iteratively denoises the paired noise boxes to obtain the final paired object boxes.

based trackers (Bewley et al. 2016; Cao et al. 2022; Zhang et al. 2022; Aharon, Orfaig, and Bobrovsky 2022; Zhao et al. 2022; Wojke, Bewley, and Paulus 2017; Zhang et al. 2021; Liu et al. 2023) that employ the Kalman filter framework (Welch, Bishop et al. 1995). In addition, certain TBD approaches establish object associations through the utilization of Re-identification (Re-ID) techniques (Chen et al. 2018; Bergmann, Meinhardt, and Leal-Taixe 2019a), and others that rely on graph-based trackers (He et al. 2021; Rangesh et al. 2021; Li, Gao, and Jiang 2020) that model the association process as minimization of a cost flow problem.

JDT approaches try to combine the tracking and detection process in a unified manner. This paradigm consists of three mainstream strategies: query-based trackers (Sun et al. 2020; Meinhardt et al. 2022; Zeng et al. 2022; Cai et al. 2022; Chen et al. 2021) that adopt unique query implicitly by forcing each query to track the same object, offset-based trackers (Bergmann, Meinhardt, and Leal-Taixe 2019b; Tokmakov et al. 2021; Xu et al. 2022; Zhou, Koltun, and Krähenbühl 2020) utilizing the motion feature to predict motion offset, and trajectory-based trackers (Pang et al. 2020; Zhou et al. 2022) that tackle severe object occlusions via spatial-temporal information. However, most of TBD and JDT ap-

\*co-corresponding author

proaches suffer from the following common drawbacks: (1) Harmful global or local inconsistency plagues both methods. In TBD approaches, the segmentation of detection and tracking tasks into distinct training processes engenders global inconsistencies that curtail overall performance. Although JDT approaches aim to bridge the gap between detection and tracking, they still treat them as disparate tasks through various branches or modules, not fully resolving the inconsistency; (2) A suboptimal balance between robustness and model complexity is evident in both approaches. While the simple structure of TBD methods suffers from poor performance when faced with detection perturbation, the complex design of JDT approaches ensures stability and robustness but compromises detection accuracy compared to TBD methods; (3) Both approaches also exhibit inflexibility across different scenes within the same video. Conventional methods process videos under uniform settings, hindering the adaptive application of strategies for varying scenes and consequently limiting their efficacy.

Recently, diffusion models have not only excelled in various generative tasks but also demonstrated potential in confronting complex discriminative computer vision challenges (Chen et al. 2022; Gu et al. 2022). This paper introduces DiffusionTrack, inspired by the progress in diffusion models, and constructs a novel consistent noise-to-tracking paradigm. DiffusionTrack directly formulates object associations from a set of paired random boxes within two adjacent frames, as illustrated in Figure 1. The motivation is to meticulously refine the coordinates of these paired boxes so that they accurately cover the same targeted objects across two consecutive frames, thereby implicitly performing detection and tracking within a uniform model pipeline. This innovative coarse-to-fine paradigm is believed to compel the model to learn to accurately distinguish objects from one another, ultimately leading to enhanced performance. DiffusionTrack addresses the multi-object tracking task by treating data association as a generative endeavor within the space of paired bounding boxes over two successive frames. Extensive experiments on 3 challenging datasets including MOT17 (Milan et al. 2016), MOT20 (Dendorfer et al. 2020) and DanceTrack (Sun et al. 2022), exhibit the state-of-the-art performance among the JDT multi-object trackers, which is also compared with TBD approaches.

In summary, our main contributions include:

1. We propose DiffusionTrack, which is the first work to employ the diffusion model for multi-object tracking by formulating it as a generative noise-to-tracking diffusion process.
2. Experimental results show that our noise-to-tracking paradigm has several appealing properties, such as decoupling training and evaluation stage for dynamic boxes and progressive refinement, promising consistency model structure for two tasks, and strong robustness to detection perturbation results.

## 2 Related Work

Existing MOT algorithms can be divided into two categories according to the paradigm of handling the detection and as-

sociation, *i.e.*, the two-stage TBD methods and the one-stage JDT methods.

**Two-stage TBD methods** is a common practice in the MOT field, where object detection and data association are treated as separate modules. The object detection module uses an existing detector (Ren et al. 2015; Duan et al. 2019; Ge et al. 2021), and the data association module can be further divided into motion-based methods (Bewley et al. 2016; Wojke, Bewley, and Paulus 2017; Zhang et al. 2022; Aharon, Orfaig, and Bobrovsky 2022; Cao et al. 2022) and graph-based (Zhang, Li, and Nevatia 2008; Jiang et al. 2019; Brasó and Leal-Taixé 2020; Li, Gao, and Jiang 2020; He et al. 2021) methods. *Motion-based methods* integrate detections through a distinct motion tracker across consecutive frames, employing various techniques. SORT (Bewley et al. 2016) initialed the use of the Kalman filter (Welch, Bishop et al. 1995) for object tracking, associating each bounding box with the highest overlap through the Hungarian algorithm (Kuhn 1955). DeepSORT (Wojke, Bewley, and Paulus 2017) enhanced this by incorporating both motion and deep appearance features, while StrongSORT (Du et al. 2022) further integrated lightweight, appearance-free algorithms for detection and association. ByteTrack (Zhang et al. 2022) addressed fragmented trajectories and missing detections by utilizing low-confidence detection similarities. P3AFormer (Zhao et al. 2022) combined pixel-wise distribution architecture with Kalman filter to refine object association, and OC-SORT (Cao et al. 2022) amended the linear motion assumption within the Kalman Filter for superior adaptability to occlusion and non-linear motion. *Graph-based methods*, including Graph Neural Networks (GNN) (Gori, Monfardini, and Scarselli 2005) and Graph Convolutional Networks (GCN) (Kipf and Welling 2016), have been widely explored in MOT, with vertices representing detection bounding boxes or tracklets and edges across frames denoting similarities. This setup allows the association challenge to be cast as a min-cost flow problem. MPNTrack (Brasó and Leal-Taixé 2020) introduced a message-passing network to capture information between vertices across frames, GNMOT (Li, Gao, and Jiang 2020) constructed dual graph networks to model appearance and motion features, and GMTracker (He et al. 2021) emphasized both inter-frame matching and intra-frame context.

**One-stage JDT methods.** In recent years, there have been several explorations into the one-stage paradigm, which combines object detection and data association into a single pipeline. *Query-based methods*, a burgeoning trend, utilize DETR (Carion et al. 2020; Zhu et al. 2020) extensions for MOT by representing each object as a query regressed across various frames. Techniques such as TrackFormer (Meinhardt et al. 2022) and MOTR (Zeng et al. 2022) perform simultaneous object detection and association using concatenated object and track queries. TransTrack (Sun et al. 2020) employs cyclical feature passing to aggregate embeddings, while MeMOT (Cai et al. 2022) encodes historical observations to preserve extensive spatio-temporal memory. *Offset-based methods*, in contrast, bypass inter-frame association and instead focus on regressing past object locations to new positions. This approach

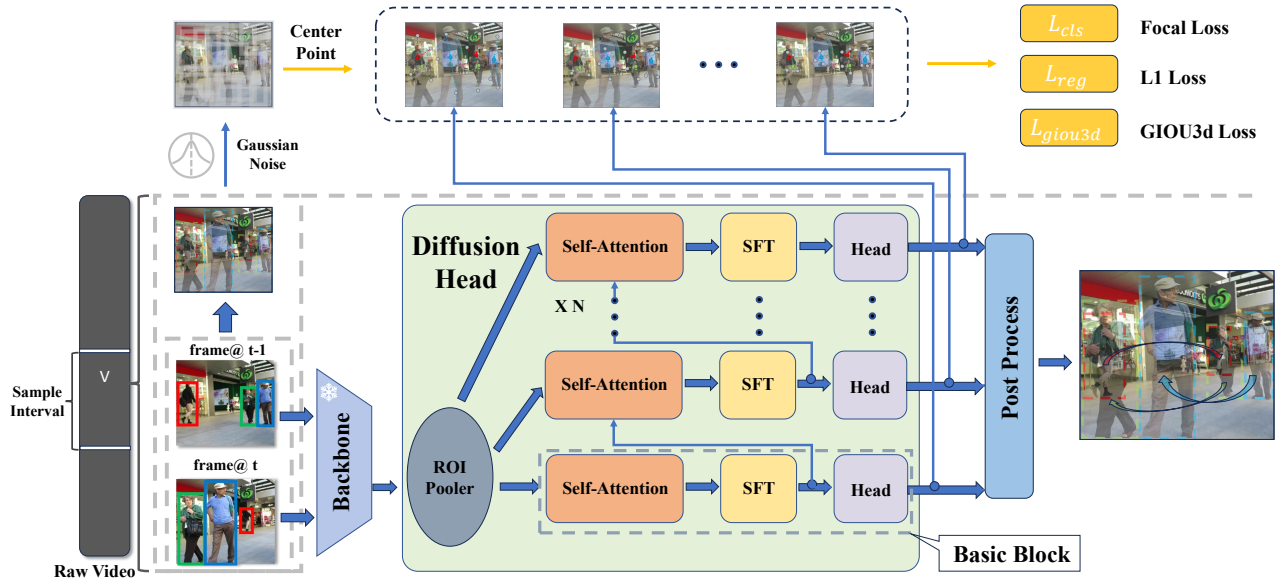


Figure 2: The architecture of DiffusionTrack. Given the images and corresponding ground-truth in the frame  $t$  and frame  $t-1$ , we extract features from two adjacent frames through the frozen backbone, then the diffusion head takes paired noise boxes as input and predicts category classification, box coordinates and association score of the same object in two adjacent frames. During training, the noise boxes are constructed by adding Gaussian noise to paired ground-truth boxes of the same object. In inference, the noise boxes are constructed by adding Gaussian noise to the padded prior object boxes in the previous frame.

includes Tracktor++ (Cai et al. 2022) for temporal realignment of bounding boxes, CenterTrack (Zhou, Koltun, and Krähenbühl 2020) for object localization and offset prediction, and PermaTrack (Tokmakov et al. 2021), which fuses historical memory to reason target location and occlusion. TransCenter (Xu et al. 2022) further advances this category by adopting dense representations with image-specific detection queries and tracking. *Trajectory-based methods* extract spatial-temporal information from historical tracklets to associate objects. GTR (Zhou et al. 2022) groups detections from consecutive frames into trajectories using trajectory queries, and TubeTK (Pang et al. 2020) extends bounding-boxes to video-based bounding-tubes for prediction. Both efficiently handle occlusion issues by utilizing long-term tracklet information.

**Diffusion model.** As a class of deep generative models, diffusion models (Ho, Jain, and Abbeel 2020; Song and Ermon 2019; Song et al. 2020) start from the sample in random distribution and recover the data sample via a gradual denoising process.

However, their potential for visual understanding tasks has yet to be fully explored. Recently, DiffusionDet (Chen et al. 2022) and DiffusionInst (Gu et al. 2022) have successfully applied diffusion models to object detection and instance segmentation as noise-to-box and noise-to-filter tasks, respectively. Inspired by their successful application of the diffusion model, we proposed DiffusionTrack, which further broadens the application of the diffusion model by formalizing MOT as a denoising process. To the best of our knowledge, this is the first work that adopts a diffusion model for the MOT task.

### 3 Method

In this section, we present our DiffusionTrack. In contrast to existing motion-based and query-based methods, we design a consistent tracker that performs tracking implicitly by predicting and associating the same object across two adjacent frames within the video sequence. We first briefly review the pipeline of multi-object tracking and diffusion models. Then, we introduce the architecture of DiffusionTrack. Finally, we present model training and inference.

#### 3.1 Preliminaries

**Multi-object tracking.** The learning objective of MOT is a set of input-target pairs  $(\mathbf{X}_t, \mathbf{B}_t, \mathbf{C}_t)$  sorted by time  $t$ , where  $\mathbf{X}_t$  is the input image at time  $t$ ,  $\mathbf{B}_t$  and  $\mathbf{C}_t$  are a set of bounding boxes and category labels for objects in the video at time  $t$  respectively. More specifically, we formulate the  $i$ -th box in the set  $\mathbf{B}_t$  as  $\mathbf{B}_t^i = (c_x^i, c_y^i, w_i, h_i)$ , where  $(c_x^i, c_y^i)$  is the center coordinates of the bounding box,  $(w_i, h_i)$  are width and height of that bounding box,  $i$  is the identity number respectively. Specially,  $\mathbf{B}_t^i = \emptyset$  when  $i$ -th object miss in  $\mathbf{X}_t$ .

**Diffusion model.** Recent diffusion models usually use two Markov chains: a forward chain that perturbs the image to noise and a reverse chain that refines noise back to the image. Formally, given a data distribution  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ , the forward noise perturbing process at time  $t$  is defined as  $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ . It gradually adds Gaussian noise to the data according to a variance schedule  $\beta_1, \dots, \beta_T$ :

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \quad (1)$$

Given  $\mathbf{x}_0$ , we can easily obtain a sample of  $\mathbf{x}_t$  by sampling a Gaussian vector  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and applying the transforma-

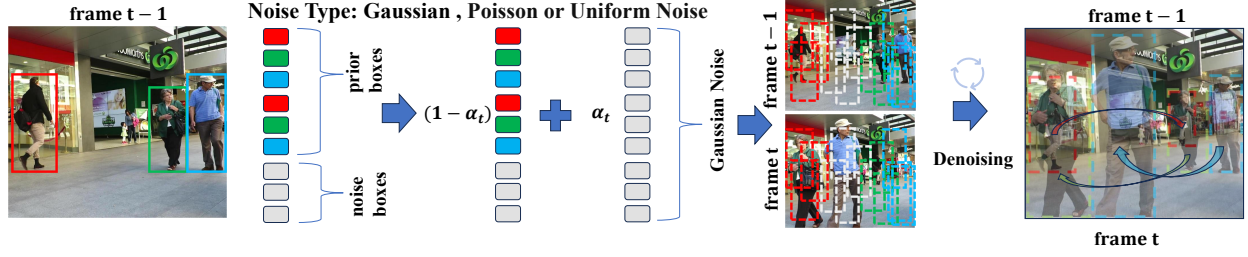


Figure 3: The inference of DiffusionTrack can be divided into three steps: (1) padding repeated prior boxes with given noise boxes until predefined number  $N_{test}$  is reached. (2) adding Gaussian noise to input boxes according to  $\mathbf{B} = (1 - \alpha_t) \cdot \mathbf{B} + \alpha_t \cdot \mathbf{B}_{noise}$  under the control of  $\alpha_t$ . (3) getting tracking results by a denoising process with the number of DDIM sampling steps  $s$ .

tion as follows:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + (1 - \bar{\alpha}_t) \epsilon, \quad (2)$$

where  $\bar{\alpha}_t = \prod_{s=0}^t (1 - \beta_s)$ . During training, a neural network predict  $\mathbf{x}_0$  from  $\mathbf{x}_t$  for different  $t \in \{1, \dots, T\}$ . In inference, we start from a random noise  $\mathbf{x}_T$  and iteratively apply the reverse chain to obtain  $\mathbf{x}_0$ .

### 3.2 DiffusionTrack

The overall framework of our DiffusionTrack is visualized in Figure 2, which consists of two major components: a feature extraction backbone and a data association denoising head (diffusion head), where the former runs only once to extract a deep feature representation from two adjacent input image  $(\mathbf{X}_{t-1}, \mathbf{X}_t)$ , and the latter takes this deep features as condition, instead of two adjacent raw images, to progressively refine the paired association box predictions from paired noise boxes. In our setting, data samples are a set of paired bounding boxes  $\mathbf{z}_0 = (\mathbf{B}_{t-1}, \mathbf{B}_t)$ , where  $\mathbf{z}_0 \in \mathbb{R}^{N \times 8}$ . A neural network  $f_\theta(\mathbf{z}_s, s, \mathbf{X}_{t-1}, \mathbf{X}_t)$   $s = \{0, \dots, T\}$  is trained to predict  $\mathbf{z}_0$  from paired noise boxes  $\mathbf{z}_s$ , conditioned on the corresponding two adjacent images  $(\mathbf{X}_{t-1}, \mathbf{X}_t)$ . The corresponding category label  $(\mathbf{C}_{t-1}, \mathbf{C}_t)$  and association confidence score  $\mathbf{S}$  are produced accordingly. If  $\mathbf{X}_{t-1} = \mathbf{X}_t$ , the multi-object tracking task degenerates into an object detection problem. The consistent design allows DiffusionTrack to solve the two tasks simultaneously.

**Backbone.** We employ the backbone of YOLOX (Ge et al. 2021) as our backbone. The backbone extracts high-level features of the two adjacent frames with FPN (Lin et al. 2017) and then feeds them into the following diffusion head for conditioned data association denoising.

**Diffusion head.** The diffusion head takes a set of proposal boxes as input to crop RoI-feature (Jiang et al. 2018) from the feature map generated by the backbone and sends these RoI-features to different blocks to obtain box regression, classification results, and association confidence scores, respectively. To solve the object tracking problem, we add a spatial-temporal fusion module (STF) and an association score head to each block of the diffusion head.

**Spatial-temporal fusion module.** We design a new spatial-temporal fusion module so that the same paired box can exchange temporal information with each other to ensure that

the data association on two consecutive frames can be completed. Given the RoI-features  $\mathbf{f}_{roi}^{t-1}, \mathbf{f}_{roi}^t \in \mathbb{R}^{N \times R \times d}$ , and the self-attention output query  $\mathbf{q}_{pro}^{t-1}, \mathbf{q}_{pro}^t \in \mathbb{R}^{N \times d}$  at current block, we conduct linear project and batch matrix multiplication to get the object query  $\mathbf{q}^{t-1}, \mathbf{q}^t \in \mathbb{R}^{N \times d}$  as:

$$\begin{aligned} \mathbf{P}_1^i, \mathbf{P}_2^i &= \text{Split}(\text{Linear1}(\mathbf{q}_{pro}^i)), \\ \text{feat} &= \text{Bmm}(\text{Bmm}(\text{Concat}(\mathbf{f}_{roi}^i, \mathbf{f}_{roi}^j), \mathbf{P}_1^i), \mathbf{P}_2^i) \\ \mathbf{q}^i &= \text{Linear2}(\text{feat}), \quad \mathbf{q}^i \in \mathbb{R}^{N \times d} \\ (i, j) &\in [(t-1, t), (t, t-1)] \end{aligned} \quad (3)$$

**Association score head.** In addition to the box head and class head, we add an extra association score head to obtain the confidence score of the data association by feeding the fused features of the two paired boxes into a Linear Layer. The head is used to determine whether the paired boxes output belongs to the same object in the subsequent Non-Maximum Suppression (NMS) post-processing process.

### 3.3 Model Training and Inference

In the training phase, our approach takes a pair of frames randomly sampled from sequences in the training set with an interval of 5 as input. we first pad some extra boxes to original ground-truth boxes appearing in both frames such that all boxes are summed up to a fixed number  $N_{train}$ . Then we add Gaussian noise to the padded ground-truth boxes with the monotonically decreasing cosine schedule for  $\alpha_t$  in time step  $t$ . We finally conduct a denoising process to get association results from these constructed noise boxes. We also design a baseline that only corrupts the ground-truth boxes in frame  $t$  and conditionally denoises the corrupted boxes based on the prior boxes in frame  $t-1$  to verify the necessity of corruption design for both frames in DiffusionTrack.

**Loss Function.** GIoU (Rezatofighi et al. 2019) loss is an extension of IoU loss which solves the problem that there is no supervisory information when the predicted boxes have no intersection with the ground-truth. We extend the definition of GIoU to make it compatible with paired boxes design. 3D GIoU and 3D IoU are the volume-extended versions of the original area ones. For each pair paired  $(\mathbf{T}_d, \mathbf{T}_{gt})$  in the matching set  $\mathbf{M}$  obtained by the Hungarian matching algorithm, we denote its class score, predicted boxes result, and association score as  $(\mathbf{C}_d^{t-1}, \mathbf{C}_d^t)$ ,  $(\mathbf{B}_d^{t-1}, \mathbf{B}_d^t)$ , and  $\mathbf{S}_d$ . The



training loss function can be formulated as:

$$\begin{aligned}
 \mathcal{L}_{cls}(\mathbf{T}_d, \mathbf{T}_{gt}) &= \sum_{i=t-1}^t \mathcal{L}_{cls}(\sqrt{\mathbf{C}_d^i \times \mathbf{S}_d}, \mathbf{C}_{gt}^i) \\
 \mathcal{L}_{reg}(\mathbf{T}_d, \mathbf{T}_{gt}) &= \sum_{i=t-1}^t \mathcal{L}_{reg}(\mathbf{B}_d^i, \mathbf{B}_{gt}^i) \\
 \mathcal{L}_{det} &= \frac{1}{N_{pos}} \sum_{(\mathbf{T}_d, \mathbf{T}_{gt}) \in \mathbf{M}} \lambda_1 \mathcal{L}_{cls}(\mathbf{T}_d, \mathbf{T}_{gt}) + \\
 &\quad \lambda_2 \mathcal{L}_{reg}(\mathbf{T}_d, \mathbf{T}_{gt}) + \lambda_3 (1 - GIou_{3d}(\mathbf{T}_d, \mathbf{T}_{gt}))
 \end{aligned} \quad (4)$$

where  $\mathbf{T}_d$  and  $\mathbf{T}_{gt}$  are square frustums consisting of estimated detection boxes and ground-truth bounding boxes for the same target in two adjacent frames respectively.  $N_{pos}$  denotes the number of positive foreground samples.  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  are the weight coefficients that are assigned as 2, 5 and 2 during training experiments.  $\mathcal{L}_{cls}$  is the focal loss proposed in (Lin et al. 2017) and  $\mathcal{L}_{reg}$  is the  $L_1$  loss.

As shown in Figure.3, the inference pipeline of DiffusionTrack is a denoising sampling process from paired noise boxes to association results. Unlike the detection task that selects random boxes from the Gaussian distribution, the tracking task has prior information about an object in the frame  $t-1$ , so we can use prior boxes to generate initialized noise boxes with a fixed number of  $N_{test}$  as in the training phase to benefit data association. In contrast to DiffusionTrack, we simply repeat the prior box without padding extra random boxes and add Gaussian noise to prior boxes only at  $t$  in the baseline model. Once the association results are derived, IoU is utilized as the similarity metric to connect the object tracklets. To address potential occlusions, a simple Kalman filter is implemented to reassociate lost objects and more details exist in the Appendix.

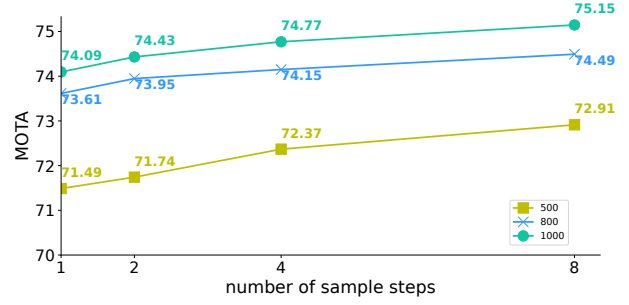
## 4 Experiments

In this section, we first introduce experimental setting and show the intriguing properties of DiffusionTrack. Then we verify the individual contributions in the ablation study and finally present the tracking evaluation on several challenging benchmarks, including MOT17 (Milan et al. 2016), MOT20 (Dendorfer et al. 2020) and DanceTrack (Sun et al. 2022). We also present the comparison with baseline model and carry out a deep analysis for DiffusionTrack.

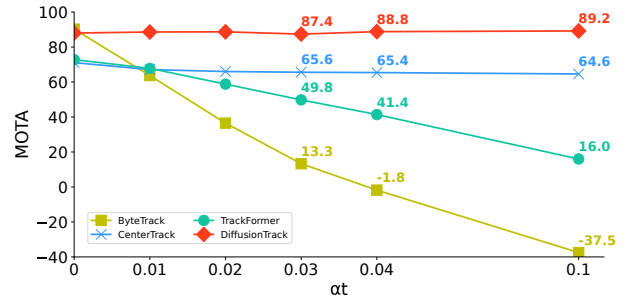
### 4.1 Setting

**Datasets.** We evaluate our method on multiple multi-object tracking datasets including MOT17 (Milan et al. 2016), MOT20 (Dendorfer et al. 2020) and DanceTrack (Sun et al. 2022). MOT17 and MOT20 are for pedestrian tracking, where targets mostly move linearly, while scenes in MOT20 are more crowded. For the data in DanceTrack, the objects have a similar appearance, severe occlusion, and frequent crossovers with highly non-linear motion.

**Metric.** We mainly use Multiple Object Tracking Accuracy (MOTA) (Bernardin and Stiefelhagen 2008), Identity F1 Score (IDF1) (Ristani et al. 2016), and Higher Order Tracking Accuracy (HOTA) (Luiten et al. 2021) for evaluation.



(a) Dynamic boxes and progressive refinement. DiffusionTrack is trained on the MOT17 train-half set with 500 proposal boxes and evaluated on the MOT17 val-half set with different numbers of proposal boxes. More sampling steps and proposal boxes in inference bring performance gain, but the effect is gradually saturated



(b) Robustness to detection perturbation. All trackers are trained on MOT17 training set and evaluated on MOT17 val-half set with little detection perturbation as  $\mathbf{B}_{det} = (1 - \alpha_t) \cdot \mathbf{B}_{det} + \alpha_t \cdot \mathbf{B}_{noise}$ . DiffusionTrack is robust to perturbation attacks with 800 proposal boxes while other approaches are vulnerable.

Figure 4: Intriguing properties of DiffusionTrack. DiffusionTrack obtains performance gain by enlarging proposal box numbers and sampling steps while being robust to detection perturbation compared with the previous tracker.

**Implementation Details.** We adopt the pre-trained YOLOX detector from ByteTrack (Zhang et al. 2022) and train DiffusionTrack on MOT17, MOT20, and DanceTrack training sets in two phases. For MOT17, the training schedule consists of 30 epochs on the combination of MOT17, CrowdHuman, Cityperson and ETHZ for detection and another 30 epochs on MOT17 solely for tracking. For MOT20, we only add CrowdHuman as additional training data. For DanceTrack, we do not use additional training data and only train 40 epochs. We also use Mosaic (Bochkovskiy, Wang, and Liao 2020) and Mixup (Zhang et al. 2017) data augmentation during the detection and tracking training phases. The training samples are directly sampled from the same video within the interval length of 5 frames. The size of an input image is resized to  $1440 \times 800$ . The 236M trainable diffusion head parameters are initialized with Xavier Uniform. The AdamW (Loshchilov and Hutter 2018) optimizer is employed with an initial learning rate of  $1e-4$ , and the learning

prior Info proportion	MOTA	IDF1	HOTA	AssA
0%	71.2	65.9	58.1	54.9
25%	73.6	70.0	60.7	58.4
50%	<b>74.5</b>	71.2	61.8	60.1
75%	74.1	<b>71.4</b>	<b>61.9</b>	<b>60.7</b>
100%	72.9	66.8	58.4	54.7

(a) Proportion of prior information. Using prior information benefit data association.

padding strategy	MOTA	IDF1	HOTA	AssA
Repeat	72.9	66.8	58.4	54.7
Cat Poisson	71.9	67.1	58.9	56.1
Cat Gaussian	<b>73.6</b>	<b>70.0</b>	<b>60.7</b>	<b>58.4</b>
Cat Uniform	71.5	63.9	56.8	52.2
Cat Full	71.2	64.4	57.3	53.7

(b) Box padding strategy. Compared to other padding strategy, concatenating Gaussian noise works best.

perturbation strategy $f(x)$	MOTA	IDF1	HOTA	AssA
0.4	73.0	67.2	58.2	54.2
$x$	73.6	70.0	60.7	58.4
$(e^x - 1)/(e - 1)$	74.3	70.5	61.4	59.7
$\log(x + 1)/\log 2$	<b>74.4</b>	<b>72.0</b>	<b>62.6</b>	<b>61.9</b>

(c) Perturbation schedule. Choosing  $t$  through a logarithmic perturbation strategy works best.

box	sampling step	MOTA	IDF1	HOTA	FLOPs(G)	FPS
500	1	71.5	66.3	58.4	<b>229.6</b>	<b>21.05</b>
500	2	71.7	68.1	59.5	459.2	10.47
800	1	73.6	70.0	60.7	367.3	15.89
1000	1	<b>74.1</b>	<b>70.7</b>	<b>61.3</b>	459.1	13.37

(d) Efficiency comparison. Adopting more proposal boxes and sampling steps brings performance gain at the cost of latency.

Table 1: Ablation experiments. The model is trained on the MOT17 train-half and tested on the MOT17 val-half. Default settings are marked in gray. See Sec 4.3 for more details.

rate decreases according to the cosine function with the final decrease factor of 0.1. We adopt a warm-up learning rate of  $2.5e-5$  with a 0.2 warm-up factor on the first 5 epochs. We train our model on 8 NVIDIA GeForce RTX 3090 with FP32-precision and a constant seed for all experiments. The mini-batch size is set to 16, with each GPU hosting two batches with  $N_{train} = 500$ . Our approach is implemented in Python 3.8 with PyTorch 1.10. We set association score threshold  $\tau_{conf} = 0.25$ , 3D NMS threshold  $\tau_{nms3d} = 0.6$ , detection score threshold  $\tau_{det} = 0.7$  and 2D NMS threshold  $\tau_{nms2d} = 0.7$  for default hyper-parameter setting. The total training time is about 30 hours, and FPS is measured with FP16-precision and batch size of 1 on a single GPU.

## 4.2 Intriguing Properties

DiffusionTrack has several intriguing properties, such as the ability to achieve better accuracy through more boxes or/and more refining steps at the higher latency cost, and strong robustness to detection perturbation for safety application.

**Dynamic boxes and progressive refinement.** Once the model is trained, it can be used by changing the number of boxes and the number of sample steps in inference. Therefore, we can deploy a single DiffusionTrack to multiple scenes and obtain a desired speed-accuracy trade-off without retraining the network. In Figure 4a, we evaluate DiffusionTrack with 500, 800, and 1000 proposal boxes by increasing their sampling steps from 1 to 8, showing that high MOTA in DiffusionTrack could be achieved by either increasing the number of random boxes or the sampling steps.

**Robustness to detection perturbation.** Almost all previous approaches are very sensitive to detection perturbation which poses significant risks to safety-critical applications such as autonomous driving. Figure 4b shows the robustness of the four mainstream trackers under detection perturbation. As can be seen from the performance comparison,

DiffusionTrack has no performance penalty for perturbation, while other trackers are severely affected, especially the two-stage ByteTrack.

## 4.3 Ablation Study

We conduct ablation experiments on several relevant factors in Figure 3 to study DiffusionTrack in detail.

**Proportion of prior information.** In contrast to object detection, multi-object tracking has prior information about the object location in the previous frame  $t - 1$ . When constructing  $N_{test}$  proposal boxes, we can control the proportion of prior information by simply repeating prior boxes. we can find that an appropriate proportion of prior information can improve the tracking performance from Table 1a.

**Box padding strategy.** Table 1b shows different box padding strategies. Our Concatenating Gaussian random boxes outperforms repeating existing prior boxes, concatenating random boxes in different noise types or image-size.

**Perturbation schedule.** Proposal boxes are initialized by adding Gaussian noise to padded prior boxes under the control of  $\alpha_t$ . We need a perturbation schedule to deal with complicated scenes, such as a larger  $\alpha_t$  when facing non-linear object motion. The perturbation schedule can be modeled by  $t$  and formulated as  $t = 1000 \cdot f(x)$ , where  $x$  is the average percentage of object motion cross two frames and  $f$  is the perturbation schedule function. As shown in Table 1c, using a logarithmic function  $f(x) = \frac{\log(x+1)}{\log 2}$  as perturbation schedule works best.

**Efficiency comparison.** Table 1d shows the efficiency comparison with different numbers of proposal boxes and sampling steps. The run time is evaluated on a single NVIDIA GeForce 3090 GPU with a mini-batch size of 1 and FP16-precision. We observe that more refinements cost brings more performance gain and results in less FPS. Diffusion-

	MOT17							MOT20						
Methods	MOTA↑	IDF1↑	HOTA↑	AssA↑	DetA↑	IDs↓	Frag↓	MOTA↑	IDF1↑	HOTA↑	AssA↑	DetA↑	IDs↓	Frag↓
<i>Two-Stage:</i>														
OC-SORT	78.0	77.5	63.2	63.4	63.2	1950	2040	75.7	76.3	62.4	62.5	62.4	942	1086
BoT-SORT	80.5	<b>80.2</b>	<b>65.0</b>	<b>65.5</b>	<b>64.9</b>	1212	<b>1803</b>	77.8	<b>77.5</b>	<b>63.3</b>	62.9	<b>64.0</b>	1313	1545
Bytetrack	80.3	77.3	63.1	62.0	64.5	2196	2277	77.8	75.2	61.3	59.6	63.4	1223	1460
StrongSORT	79.6	79.5	64.4	64.4	64.6	<b>1194</b>	1866	73.8	77.0	62.6	<b>64.0</b>	61.3	<b>770</b>	<b>1003</b>
P3AFormer	<b>81.2</b>	78.1	/	/	/	1893	/	<b>78.1</b>	76.4	/	/	/	1332	/
GMTracker	61.5	66.9	/	/	/	2415	/	/	/	/	/	/	/	/
GNMOT	50.2	47.0	/	/	/	5273	/	/	76.4	/	/	/	/	/
<i>One-Stage:</i>														
TrackFormer	74.1	68.0	57.3	54.1	60.9	2829	4221	68.6	65.7	54.7	53.0	56.7	<b>1532</b>	<b>2474</b>
MeMOT	72.5	69.0	56.9	55.2	/	2724	/	63.7	66.1	54.1	<b>55.0</b>	/	1938	/
MOTR	71.9	68.4	57.2	55.8	/	<b>2115</b>	<b>3897</b>	/	/	/	/	/	/	/
CenterTrack	67.8	64.7	52.2	51.0	53.8	3039	6102	/	/	/	/	/	/	/
PermaTrack	73.8	68.9	55.5	53.1	58.5	3699	6132	/	/	/	/	/	/	/
TransCenter	73.2	62.2	54.5	49.7	60.1	4614	9519	67.7	58.7	/	/	/	3759	/
GTR	<u>75.3</u>	<u>71.5</u>	<u>59.1</u>	<u>57.0</u>	<u>61.6</u>	<u>2859</u>	/	/	/	/	/	/	/	/
TubeTK	63.0	58.6	/	/	/	4137	/	/	/	/	/	/	/	/
Baseline	74.6	66.7	55.9	50.8	61.9	16375	7206	63.3	49.5	42.5	34.7	52.5	9990	6710
DiffusionTrack	<b>77.9</b>	<b>73.8</b>	<b>60.8</b>	<b>58.8</b>	<b>63.2</b>	3819	4815	<b>72.8</b>	<b>66.3</b>	<b>55.3</b>	51.3	<b>59.9</b>	4117	4446

Table 2: Performance comparison to state-of-the-art approaches on the MOT17 and MOT20 test set with the private detections. The best results are shown in bold. The offline method is marked in underline.

Methods	HOTA↑	MOTA↑	DetA↑	AssA↑	IDF1↑
QDTrack	45.7	83.0	72.1	29.2	44.8
TraDes	43.3	86.2	74.5	25.4	41.2
SORT	47.9	<b>91.8</b>	72.0	31.2	50.8
ByteTrack	47.3	89.5	71.6	31.4	52.5
OC-SORT	<b>54.6</b>	89.6	<b>80.4</b>	<b>40.2</b>	<b>54.6</b>
TransTrack	45.5	88.4	75.9	27.5	45.2
CenterTrack	41.8	86.8	78.1	22.6	35.7
GTR	48.0	<u>84.7</u>	<u>72.5</u>	<u>31.9</u>	<u>50.3</u>
Baseline	44.0	79.4	74.1	26.2	40.2
DiffusionTrack	<b>52.4</b>	<b>89.3</b>	<b>82.2</b>	<b>33.5</b>	<b>47.5</b>

Table 3: Performance comparison to state-of-the-art approaches on the DanceTrack test set. The best results are shown in bold. Offline method is marked in underline

Track can flexibly choose different settings for every single frame to deal with complicated scenes within a video.

#### 4.4 State-of-the-art Comparison

Here we report the benchmark results of DiffusionTrack and baseline compared with other mainstream methods on multiple datasets. We evaluated DiffusionTrack on DanceTrack, MOT17, and MOT20 test datasets with 500, 800, and 1000 noise boxes respectively in same default setting.

**MOT17 and MOT20.** We use the standard split and obtain the test set evaluation by submitting the results to the online website. As can be seen from the performance comparison in Table2, our DiffusionTrack achieves state-of-the-art both in MOT17 and MOT20 for one-stage methods with the MOTA

of 77.9 and 72.8 respectively.

**DanceTrack.** To evaluate DiffusionTrack under challenging non-linear object motion, we report results on the DanceTrack in Table 3. DiffusionTrack achieves the state-of-the-art on DanceTrack with HOTA (52.4).

The baseline model has a close performance to DiffusionTrack on MOT17 but performs very poorly on MOT20 and DanceTrack. In our understanding, Baseline simply learns a coordinate regression between boxes  $\mathbf{B}_{t-1}$  and boxes  $\mathbf{B}_t$  at conditioned on the pooled features at time  $t-1$  which can not deal with crowd and non-linear object motion problem. We guess the coarse-to-fine diffusion process is a special data-augmented method that can enable DiffusionTrack to discriminate between various objects.

## 5 Conclusion

In this work, we propose a novel end-to-end multi-object tracking approach that formulates object detection and association jointly as a consistent denoising diffusion process from paired noise boxes to object association. Our noise-to-tracking pipeline has several appealing properties, such as dynamic box and progressive refinement, consistent model structure, and robustness to perturbation detection results, enabling us to obtain the desired speed-accuracy trade-off with same network parameters. Extensive experiments show that DiffusionTrack achieves favorable performance compared to previous strong baseline methods. We hope that our work will provide a interesting insight into multi-object tracking from the perspective of the diffusion model, and that the performance of a wide variety of trackers can be enhanced by local or global denoising processes.

## Acknowledgments

Min Yang was supported by National Key Research and Development Program of China (2022YFF0902100), Shenzhen Science and Technology Innovation Program (KOTD20190929172835662), Shenzhen Basic Research Foundation (JCYJ20210324115614039 and JCYJ20200109113441941). The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

## References

- Aharon, N.; Orfaig, R.; and Bobrovsky, B.-Z. 2022. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*.
- Bergmann, P.; Meinhardt, T.; and Leal-Taixe, L. 2019a. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 941–951.
- Bergmann, P.; Meinhardt, T.; and Leal-Taixe, L. 2019b. Tracking without bells and whistles. In *Proceedings of the ICCV*, 941–951.
- Bernardin, K.; and Stiefelhagen, R. 2008. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008: 1–10.
- Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, 3464–3468. IEEE.
- Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.
- Brasó, G.; and Leal-Taixé, L. 2020. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6247–6257.
- Cai, J.; Xu, M.; Li, W.; Xiong, Y.; Xia, W.; Tu, Z.; and Soatto, S. 2022. MeMOT: multi-object tracking with memory. In *Proceedings of the CVPR*, 8090–8100.
- Cao, J.; Weng, X.; Khirodkar, R.; Pang, J.; and Kitani, K. 2022. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*.
- Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of the ECCV*, 213–229. Springer.
- Chen, L.; Ai, H.; Zhuang, Z.; and Shang, C. 2018. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *2018 IEEE international conference on multimedia and expo (ICME)*, 1–6. IEEE.
- Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2022. Diffusion-det: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*.
- Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the CVPR*, 8126–8135.
- Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; and Leal-Taixé, L. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.
- Du, Y.; Song, Y.; Yang, B.; and Zhao, Y. 2022. Strongsort: Make deepsort great again. *arXiv preprint arXiv:2202.13514*.
- Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the ICCV*, 6569–6578.
- Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.
- Gori, M.; Monfardini, G.; and Scarselli, F. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, 729–734. IEEE.
- Gu, Z.; Chen, H.; Xu, Z.; Lan, J.; Meng, C.; and Wang, W. 2022. DiffusionInst: Diffusion Model for Instance Segmentation. *arXiv preprint arXiv:2212.02773*.
- He, J.; Huang, Z.; Wang, N.; and Zhang, Z. 2021. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In *Proceedings of the CVPR*, 5299–5309.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.
- Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the ECCV*, 784–799.
- Jiang, X.; Li, P.; Li, Y.; and Zhen, X. 2019. Graph neural based end-to-end data association framework for online multiple-object tracking. *arXiv preprint arXiv:1907.05315*.
- Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.
- Li, J.; Gao, X.; and Jiang, T. 2020. Graph networks for multiple object tracking. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 719–728.
- Lin, T.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. *IEEE TPAMI*, PP(99): 2999–3007.
- Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the CVPR*, 2117–2125.
- Liu, Z.; Wang, X.; Wang, C.; Liu, W.; and Bai, X. 2023. SparseTrack: Multi-Object Tracking by Performing Scene Decomposition based on Pseudo-Depth.
- Loshchilov, I.; and Hutter, F. 2018. Decoupled weight decay regularization. In *Proceedings of the ICLR*.



- Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; and Leibe, B. 2021. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129: 548–578.
- Meinhardt, T.; Kirillov, A.; Leal-Taixé, L.; and Feichtenhofer, C. 2022. Trackformer: Multi-object tracking with transformers. In *Proceedings of the CVPR*, 8844–8854.
- Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; and Schindler, K. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.
- Pang, B.; Li, Y.; Zhang, Y.; Li, M.; and Lu, C. 2020. 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*, 6308–6318.
- Rangesh, A.; Maheshwari, P.; Gebre, M.; Mhatre, S.; Ramezani, V.; and Trivedi, M. M. 2021. Trackmpnn: A message passing graph neural architecture for multi-object tracking. *arXiv preprint arXiv:2101.04206*.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the CVPR*, 658–666.
- Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the ECCV*, 17–35. Springer.
- Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Sun, P.; Cao, J.; Jiang, Y.; Yuan, Z.; Bai, S.; Kitani, K.; and Luo, P. 2022. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the CVPR*, 20993–21002.
- Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; and Luo, P. 2020. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*.
- Tokmakov, P.; Li, J.; Burgard, W.; and Gaidon, A. 2021. Learning to track with object permanence. In *Proceedings of the ICCV*, 10860–10869.
- Welch, G.; Bishop, G.; et al. 1995. An introduction to the Kalman filter.
- Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple on-line and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, 3645–3649. IEEE.
- Xu, Y.; Ban, Y.; Delorme, G.; Gan, C.; Rus, D.; and Alameda-Pineda, X. 2022. TransCenter: Transformers with dense representations for multiple-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; and Wei, Y. 2022. Motr: End-to-end multiple-object tracking with transformer. In *Proceedings of the ECCV*, 659–675.
- Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.
- Zhang, L.; Li, Y.; and Nevatia, R. 2008. Global data association for multi-object tracking using network flows. In *2008 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.
- Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the ECCV*, 1–21. Springer.
- Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129: 3069–3087.
- Zhao, Z.; Wu, Z.; Zhuang, Y.; Li, B.; and Jia, J. 2022. Tracking objects as pixel-wise distributions. In *Proceedings of the ECCV*, 76–94. Springer.
- Zhou, X.; Koltun, V.; and Krähenbühl, P. 2020. Tracking objects as points. In *Proceedings of the ECCV*, 474–490. Springer.
- Zhou, X.; Yin, T.; Koltun, V.; and Krähenbühl, P. 2022. Global Tracking Transformers. In *CVPR*.
- Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.