DIMATrack: Dimension Aware Data Association for Multi-Object Tracking

Shu Liu[1], Melikamu Liyih Sinishaw[2], Luo Zheng[1*]
[1]School of Computer Science and Engineering, Central South University, China
[2]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
{sliu35, Loki369}@csu.edu.cn, melikamuliyih29@gmail.com

Abstract

Multi-Object Tracking (MOT) is crucial for real-world applications like video surveillance, where it aims to detect and maintain consistent identifiers for objects across video frames. However, MOT methods often struggle with objects that are heavily overlapped due to occlusion or exhibit diverse poses due to non-linear motion. In this paper, we propose a robust tracking-by-detection method named DIMATrack. It incorporates the Kalman Filter for precise trajectory prediction, and a novel Dimension Aware Intersection-over-Union (DIMA-IoU) metric for enhanced data association. DIMA-IoU improves upon standard IoU by integrating both height-aware and width-aware measurements, improving association accuracy in complex scenarios and during occlusions. By integrating these components, DIMATrack effectively leverages weak cues that are often overlooked by conventional methods, which rely on appearance or spatial information. Extensive experiments on three benchmarks demonstrate the superior performance of our DIMATrack, particularly in challenging tracking environments. The code is available at https://github.com/Melikamuliyih/DIMATrack.

Keywords: ***Multi-object Tracking, Tracking-by-detection, Dimension Aware, Data Association.***

1. Introduction

Multi-Object Tracking (MOT) is a long-standing challenge in computer vision, critical for various applications such as autonomous driving, action recognition, smart elderly care, and human-computer interaction. It aims to detect and track all specific objects frame by frame, which plays an essential role in video understanding.

The prevailing tracking-by-detection methods [2, 37, 14, 30] divide MOT into two sub-tasks: detecting objects in each frame and associating these detections over time. Such approach relies heavily on spatial and appearance information. However, these strong cues fail in complex scenarios like high occlusion and dynamic poses, where objects heavily overlap. Weak cues, such as width state, height state, and velocity direction, can effectively mitigate ambiguous associations when strong cues become unreliable.

Previous works [5, 22] have recognized the potential of weak cues. However, their effectiveness is restricted to specific object interactions. Motion information is vital in MOT, especially for dynamic scenes, yet insufficient alone. Intersection-over-Union (IoU) has been the primary metric for data association. It operates on the image plane and thus falter with dynamic target movements or occlusions. We observe that the weak cues of height and width information from bounding boxes, could effectively address ambiguous associations in the above scenarios.

To advance the state-of-the-art performance in MOT, this paper introduces DIMATrack, a simple yet powerful tracker. It utilizes the high-performance YOLOv7 detector [27] to capture detection boxes, and associate them with our innovative Dimension Aware IoU (DIMA-IoU). We consider DIMA-IoU, the average of height-aware state and width-aware state as potential types of weak cues. Both state properties of objects robustly handle the complexities introduced by diverse poses and highly overlaps due to occlusion and clustering, as they contain the depth information. A Kalman filter complements this approach by predicting object trajectories, enhancing the tracking accuracy further.

Our evaluations on three benchmarks demonstrate that DIMATrack significantly outperforms existing methods in all main MOT metrics (Figure 1). Its simplicity, online operational capability, and efficiency ensure its suitability for real-time applications. The method's generalizability and ease of integration make it particularly attractive for diverse MOT scenarios, including edge device implementations. The main contributions of this work can be summarized as follows:

- We propose an online tracking-by-detection method called DIMATrack, which employs an optimized Kalman filter state vector for enhanced box localization, improving the overall tracking-by-detection framework.
- We introduce DIMA-IoU, a simple and efficient method that averages height-aware IoU and width-

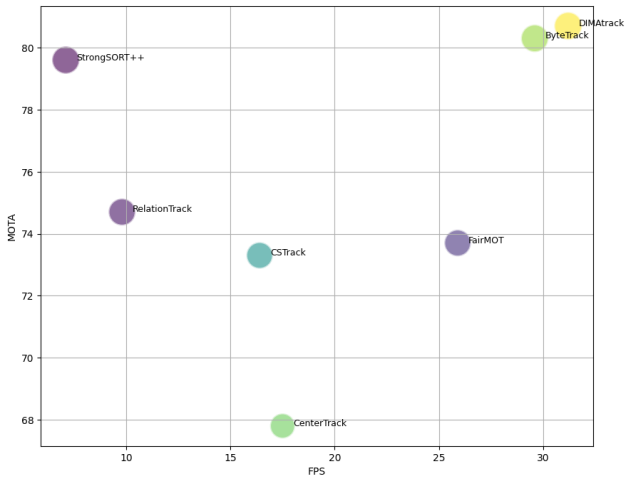

Figure 1. MOTA-FPS-IDF1 comparisons of various trackers on MOT17 dataset. The horizontal and vertical axes represent FPS (running speed) and MOTA, respectively, and the circle radius corresponds to IDF1. Our DIMATrack achieves 80.7% MOTA and 79.0% IDF1 with a running speed of 30 FPS, surpassing all these trackers. More details are available in Table 1.

aware IoU to address challenges related to object overlap and pose diversity caused by occlusions and clustering.

- We demonstrate the consistent and significant improvements across multiple benchmarks and representative trackers, substantiating DIMATrack's superior performance and adaptability.

2. Related Work

Object detection and data association stand as two pivotal components within the realm of multi-object tracking. Detection tasks involve estimating the bounding boxes of objects, while association tasks entail assigning identities to these detected objects.

2.1. Tracking-by-Detection

The tracking-by-detection methods [2, 30, 9, 37] which have been widely used in multi-object tracking, typically involve a two-step process of detection and association. Tracking-by-detection methods make a clear distinction between the detection and tracking of objects. The basic concept involves localizing all objects within each frame using an object detector, followed by associating the detected objects across frames based on features like position and appearance.

With the rapid advancements in object detection [24, 25], an increasing number of methods are turning to more robust detectors to achieve higher tracking performance. The one-stage object detector RetinaNet [13] has been adopted by several methods, including [16, 20]. CenterNet [41] has emerged as the most favored detector among many methods [40, 38] due to its simplicity and efficiency. Additionally, the YOLO series detectors [21, 3, 11] have gained popularity among a large number of methods [37, 29, 13] for their optimal balance of accuracy and speed. While many of these methods directly utilize detection boxes from individual images for tracking, this approach often results in low-score detections and missed detections, leading to subpar tracking performance. ByteTrack [37] addresses this issue by associating high and low score detections at different stages using IoU. However, it falls short when dealing with highly overlapped objects or objects in diverse poses.

2.2. Data Association

Data association lies at the heart of multi-object tracking, where it initially calculates the similarity between tracklets and detection boxes, subsequently aligning them based on their similarity. Its purpose is to match multiple targets between frames, including assigning IDs for new measurements, the creation of new tracks and the elimination of old tracks [38, 37].

The assessment of similarity holds significant importance in determining the outcomes of object tracking processes. Typically, detection-centric methodologies rely on the IoU metric to gauge similarity for sequential matching. FairMOT [38] incorporates both Mahalanobis distance and Cosine distance to evaluate object similarity during initial matching, subsequently employing IoU distance for secondary matching. Similarly, JDE [29] integrates appearance and motion characteristics to measure similarity during initial matching and utilizes IoU distance for subsequent matching. SORT [2] utilizes IoU distance as the similarity metric for the Hungarian algorithm, while DeepSort [30] utilizes Cosine distance and IoU distance for the nearest neighbor algorithm. ByteTrack [37] follows a two-stage matching approach, distinguishing between high-scoring and low-scoring boxes using IoU distance. However, SampleTrack [12] contends that none of these similarity metrics offer an optimal representation. In multi-object tracking, matching failures often stem from inaccuracies in predictions by the Kalman filter [4], particularly as target loss duration increases. This results in inaccuracies in motion cues and IoU distance, leading to linear assignment errors. To address this challenge, the average of the height aware and width aware for tracking-by-detection paradigm introduced a two-stage association strategy, featuring an innovative similarity matrix that incorporates the cosine matrix to evaluate target distances, thereby mitigating incorrect assignments and ensuring robust tracking.

Matching strategy various approaches exist post-similarity computation. SORT [2] employs a one-shot matching approach, while DeepSORT [30] introduces a cas-

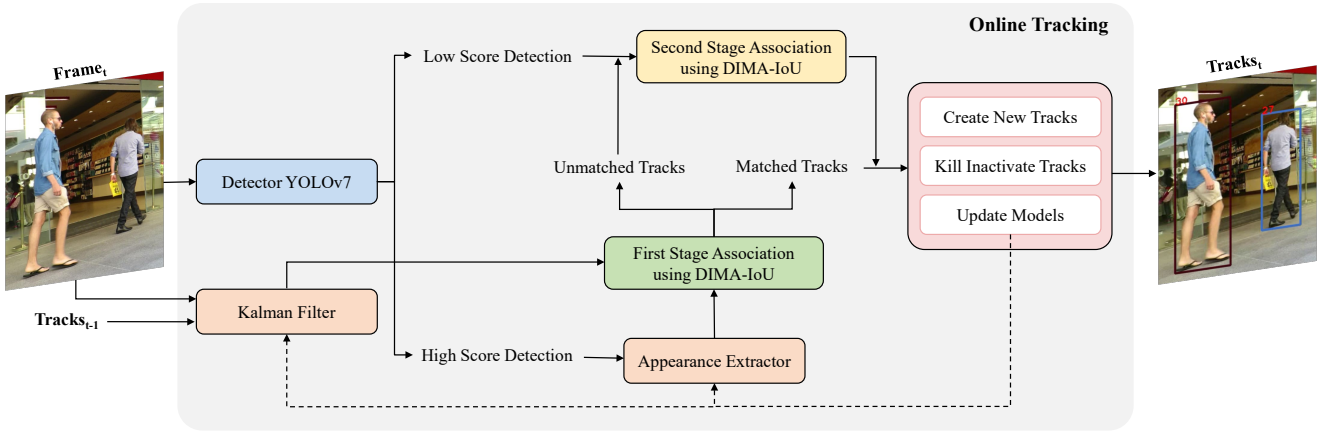

Figure 2. Overview of our proposed tracking-by-detection method — DIMATrack. It leverages the Kalman filter to predict object trajectories and the DIMA-IoU for data association in both the first and second stages.

caded matching strategy, initially pairing detection boxes with recent tracklets before considering those that were previously lost. MOTDT [6] utilizes appearance similarity for initial matching and subsequently employs IoU similarity for unmatched tracklets. QuasiDense [19] converts appearance similarity into probabilities and employs nearest neighbor search for matching. Attention mechanisms [26] facilitate implicit association by directly propagating boxes between frames. Recent innovations, like those presented in [18, 36], introduce track queries to anticipate tracked object locations in subsequent frames, implicitly conducting matching during attention interaction. Despite advancements in association methods, the quality of detection boxes sets the upper limit for data association. Therefore, our focus is on optimizing the utilization of detection boxes across varying confidence levels during the matching process

3. Proposed Method

In this section, we introduce main enhancements and advancements in multi-object tracking within the framework of tracking-by-detection methods. We present a novel state-of-the-art tracker, called DIMATrack, which incorporates innovative techniques to enhance both accuracy and robustness in object tracking. DIMATrack utilizes the Kalman filter for precise box localization and DIMA-IoU for improved data association.

3.1. Overall Architecture of DIMATrack

The overall architecture of our tracking-by-detection method is illustrated in Figure 2. To improve object feature extraction, we adopt appearance extractor for a more robust baseline model [17] instead of relying solely on a high-performance detector. We use the Kalman filter to predict the trajectory of the object referencing the previous tracklets. Inspired by ByteTrack [37], which retains all detection boxes and separates them into two stages for high-score and low-score detections, it uses conventional IoU to associate the tracklets in both stages.

For data association, our DIMATrack incorporates the average of height-aware and width-aware IoU metrics in both the first and second association stages. First, we associate the high-score detection boxes with the tracklets. However, some tracklets remain unmatched when they cannot find an appropriate high-score detection box. This typically occurs during occlusion, instances of highly overlapped objects, or diverse poses, and is addressed using DIMA-IoU. Subsequently, we associate the low-score detection boxes with these unmatched tracklets to recover the objects identified in the low-score detection boxes, while simultaneously filtering out background using DIMA-IoU. Additionally, we employ conventional IoU for retrieving lost tracklets.

3.2. Kalman Filter

The discrete Kalman filter with a constant-velocity model is commonly used for modeling object motion in the image plane. Kalman filter [4] is a linear estimator for dynamical systems discretized in the time domain. It operates based on the state estimation from the previous time step and the current measurement to estimate the state for the new time step. It keeps track of two main variables: the posterior state estimate x and the posterior estimate covariance matrix P of the state. As introduced in [29, 30], we use the state vector x as eight tuples, $x = [x_c, y_c, a, h, \hat{x_c}, \hat{y_c}, \hat{a}, \hat{h}]^T$, where (x_c, y_c) represents the 2D coordinates of the object's center in the image plane, h denotes the scale (area) of the bounding box, and a refers to its aspect ratio. Directly estimating the bounding box's width and height leads to improved performance. We choose to define the Kalman filter's state vector as in Eq. (1) with

eight tuples and Kalman filter's measurement vector as in Eq. (2).

$$x_t = [x_c(t), y_c(t), w(t), h(t), \hat{x}_c(t), \hat{y}_c(t), \hat{w}(t), \hat{h}(t)]^T \tag{1}$$

$$z_t = [z_{x_c}(t), z_{y_c}(t), z_w(t), z_h(t)]^T \tag{2}$$

In the context of MOT, the SORT algorithm [2] employs time-independent process noise covariance (Q) and measurement noise covariance (R) matrices for the Kalman Filter. However, DeepSORT [30] proposes a different approach, suggesting that Q and R should be dynamically adapted based on estimated elements (likely from the state vector) and measurement elements. Thus, the time-dependent process noise covariance Q_t and measurement noise covariance R_t matrices are shown in Eq. (3) and (4), respectively.

$$\begin{aligned} Q_t = \text{diag}\Big(& (\sigma_p \hat{w}_{t-1|t-1})^2, (\sigma_p \hat{h}_{t-1|t-1})^2, \\ & (\sigma_p \hat{w}_{t-1|t-1})^2, (\sigma_p \hat{h}_{t-1|t-1})^2, \\ & (\sigma_v \hat{w}_{t-1|t-1})^2, (\sigma_v \hat{h}_{t-1|t-1})^2, \\ & (\sigma_p \hat{w}_{t-1|t-1})^2, (\sigma_p \hat{h}_{t-1|t-1})^2 \Big) \end{aligned} \tag{3}$$

$$\begin{aligned} R_t = \text{diag}\Big(& (\sigma_m z_w(t))^2, (\sigma_m z_h(t))^2, \\ & (\sigma_m z_w(t))^2, (\sigma_m z_h(t))^2 \Big) \end{aligned} \tag{4}$$

Following the settings in [30], we adopt noise factors of $\sigma_p = 0.05$, $\sigma_v = 0.00625$, and $\sigma_m = 0.05$ due to our matching frame rate of 30 FPS. It's important to note that we adjusted the process noise covariance matrix (Q) and measurement noise covariance matrix (R) to account for slight differences in our state vector (x) compared to [30]. Additionally, to prevent box shape deformation during long predictions in case of track loss, we implemented a logic mechanism similar to the approach presented in [37].

3.3. DIMA-IoU

Identifying temporally stable object properties is crucial for effective multi-object tracking. The height and width states offer valuable information that compensates for the absence of strong discriminative cues, enhancing object differentiation in challenging scenarios. Our method proposes Dimension Aware IoU, which is the average of a height-aware IoU and width-aware IoU metric to enhance association accuracy in scenarios with highly overlapping or clustered objects. Figure 3 illustrates the benefits of DIMA-IoU in resolving object overlap during tracking. The left frame shows the input with multiple objects overlapping, demonstrating the challenge of distinguishing between them. In the middle frame, the height-aware and width-aware states of the overlapped objects are used for tracking. This step is crucial for distinguishing individual objects despite the overlap. The right frame presents the final result after applying DIMA-IoU, where bounding boxes are clearly defined around each object, accurately reflecting their dimensions and trajectories. This sequence highlights how DIMA-IoU demonstrates clear advantages in handling overlapping objects compared to traditional IoU methods.

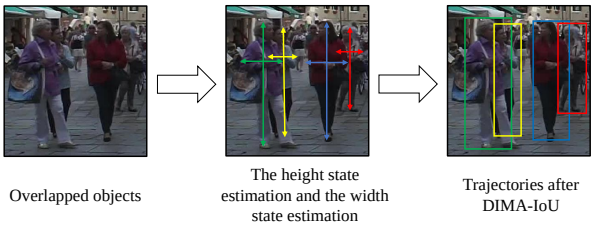

Figure 3. The benefits of DIMA-IoU are illustrated, where the vertical line represents the height estimation and the horizontal line represents the width estimation. The bounding boxes depict the result after estimating the height and width state for the overlapped objects.

Inspired by [33], Height-Aware IoU (HAIoU) information offers valuable clues for distinguishing such objects, where traditional appearance-based cues might be unreliable. This benefit stems from two key advantages: firstly, object height often reflects depth information, making it effective in differentiating significantly overlapped objects in datasets like DanceTrack. Secondly, height is generally less susceptible to variations in object pose, leading to a more robust and accurate representation of object association.

Width-Aware IoU (WAIoU) can be a valuable tool for data association in specific contexts, It excels at capturing horizontal alignment between bounding boxes, making it useful when that's a prime concern. Furthermore, it can complement traditional IoU by providing a more nuanced view of overlap, focusing specifically on the horizontal dimension. WAIoU offers a valuable metric for data association, particularly when horizontal alignment is a critical factor. Width state also contributes to enhancing the association, especially when the object's movement is regular. By incorporating HAIoU alongside WAIoU, a more comprehensive assessment of bounding box overlap becomes possible.

The work [33] focused solely on height-modulated IoU. However, it faces challenges when the target object is obstructed by taller objects. To address this issue, we propose DIMA-IoU. DIMA-IoU utilizes the average of height-aware IoU and width-aware IoU to improve the tracking performance.

We denote two bounding boxes as b1 and b2. Each box is defined by its top-left corner coordinates ($x11$, $y11$) and

bottom-right corner coordinates ($x12$, $y12$). Areas of these boxes are represented by A and B, respectively. Therefore, the conventional IoU calculated as Eq. (5) and WAIoU and HAIoU calculated as Eq. (9) and Eq. (11), respectively.

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

$$overlap_{width} = \max(0,\, \min(x12,\, x22) - \max(x11,\, x21)) \tag{6}$$

$$union_{width} = (x12 - x11) + (x22 - x21) - overlap_{width} \tag{7}$$

$$\text{IoU}_W = \frac{overlap_{width}}{union_{width}} \tag{8}$$

$$\text{WAIoU} = \text{IoU}_W \cdot \text{IoU} \tag{9}$$

$$\text{IoU}_H = \frac{\min(y12,\, y22) - \max(y11,\, y21)}{\max(y12,\, y22) - \min(y11,\, y21)} \tag{10}$$

$$\text{HAIoU} = \text{IoU}_H \cdot \text{IoU} \tag{11}$$

Therefore, the average of the height and width aware IoU (DIMA-IoU) calculated as Eq. (12):

$$\text{DIMA-IoU} = \frac{(\text{HAIoU} + \text{WAIoU})}{2} \tag{12}$$

4. Experiments and Results

In this section, we present the experimental results on three benchmarks to demonstrate the effectiveness of our DIMATrack tracking method. The datasets, evaluation metrics, and implementation details are first introduced. The experiments aim to serve four main purposes: 1) to compare our method with the state of the art; 2) to demonstrate the advantage of DIMA-IoU; 3) to verify the generality of our design; 4) to provide both qualitative and quantitative analysis.

4.1. Experimental Settings

4.1.1 Datasets

We test our DIMATrack on several MOT benchmarks, including MOT17 [7], MOT20 [8], and DanceTrack [23]. MOT17 serves as a widely recognized standard benchmark for multi-object tracking, primarily characterized by linear motion patterns. MOT20 was specifically designed to assess algorithms in scenarios with densely packed objects and significant occlusions. DanceTrack presents one of the most demanding challenges in the MOT domain, featuring a wide array of complex, non-linear motion patterns, along with frequent interactions and occlusions. Notably, the detection task in DanceTrack is relatively straightforward, making it an ideal metric for evaluating association performance. Given the distinct characteristics of these benchmarks, our primary focus is on comparing our method's performance on DanceTrack, aiming to enhance association performance in challenging conditions with limited cues. We leverage MOT17 and MOT20 to assess the generalization capability of our approach across various scenarios. The MOT17 validation set adheres to a commonly used convention, where the training set is divided into two halves for training and validation purposes.

4.1.2 Evaluation Metrics

We employ the CLEAR metrics [1], encompassing MOTA, FN, FP, FPS, and IDF1 [31], to comprehensively assess various facets of tracking performance. MOTA is derived from FP, FN, and IDs, with its calculation prioritizing detection performance due to the typically larger quantities of FP and FN compared to IDs. MOTA consolidates three distinct error metrics: ID switches, false positives, and false negatives - into a unified score, obtained by summing these metrics and dividing by the total number of objects across all frames. MOTA is expressed as $\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDS}_t)}{\sum_t \text{GT}_t}$. Where FN_t false negative at frame t, FP_t is the false positive at frame t. Conversely, IDF1 evaluates identity preservation capability and emphasizes association performance. Offering enhanced measures of ID matching consistency over MOTA, IDF1 amalgamates ID precision (IDP) and ID recall (IDR) into a single value using the harmonic mean, computed as $\text{IDF1} = \frac{2 \times \text{IDP} \times \text{IDR}}{\text{IDP} + \text{IDR}}$. where IDP and IDR are defined as per the precision and recall definitions, respectively.

4.2. Implementation Details

We employ a YOLOv7 [27] detector with a YOLOX-X backbone and pre-trained weights from COCO for our task. To enhance performance, we train the model for 60 epochs on a combined dataset consisting of MOT17, CrowdHuman, Cityscapes, and ETHZ. During training, images are resized to $1440X800$ pixels and the shortest side is further adjusted between 576 and 1024 pixels for multi-scale training. Mosaic and Mixup data augmentation techniques are also applied. The training process leverages an NVIDIA GeForce RTX 3090Ti GPU with a batch size of 30. We utilize the SGD optimizer with weight decay and momentum for optimization. The initial learning rate is set to $1e^{-3}$ with a warm-up period of 1 epoch, followed by a cosine annealing schedule. Training takes approximately 11 hours.

Following the evaluation method in [37], we measure the model's frame rate (FPS) using FP16 precision and a batch

Table 1. Comparison results on MOT17 dataset. The best results are shown in bold. The (↑) indicates the higher is better and (↓) indicates the lower is better.

Tracker	MOTA (↑)	IDF1 (↑)	FP (↓)	FN (↓)	FPS (↑)
TrackFormer [18]	74.1%	68.0%	34602	108777	-
MOTR [36]	73.4%	68.6%	-	-	-
MOTRv2 [39]	78.6%	75.0%	-	-	-
CenterTrack [40]	67.8%	64.7%	18498	160332	17.5
QDTrack [19]	67.8%	66.3%	26589	146643	-
FairMOT [38]	73.7%	72.3%	27507	117477	25.9
CSTrack [16]	74.9%	72.6%	23847	114303	16.4
SimpleTrack [12]	75.3%	76.3%	22317	116010	-
RelationTrack [35]	73.8%	74.7%	27999	118623	9.8
SORT [2]	33.4%	39.8%	7318	32615	-
UCMCTrack [34]	80.5%	**81.1%**	-	-	-
ByteTrack [37]	80.3%	77.3%	25491	83721	29.6
GHOST [22]	78.7%	77.1%	-	-	-
ColTrack [15]	78.8%	73.9%	-	-	-
MeMOTR [10]	72.8%	71.5%	-	-	-
OC-SORT [5]	78.0%	77.5%	15129	107055	-
StrongSORT++ [9]	79.6%	79.5%	27876	86205	7.1
Hybrid-SORT [33]	79.9%	78.7%	-	-	-
DIMATrack (Ours)	**80.7%**	79.0%	**2572**	**7398**	**31.2**

size of 1 on a single GPU. The default thresholds for high detection scores and low detection scores are set at 0.6 and 0.1, respectively, with a trajectory initialization score of 0.7, unless specified otherwise. During the linear assignment step, if the IoU between the detection box and the tracklet box falls below 0.25, the matching is rejected. Lost tracklets are retained for 30 frames in case they reappear.

4.3. Benchmark Results

We compare DIMATrack against state-of-the-art methods on MOT17, MOT20 and DanceTrack. our approach achieves the overall superior performance to the others. Our tracking-by-detection approach consistently outperforms the baseline ByteTrack [37] in all three datasets with negligible additional computation, while maintains simple, online and real-time characteristics.

4.3.1 MOT17

The dataset consists of 7 sequences validation set and 7 sequences test set. The tracking performance on MOT17 is compared in Table 1. In particular, our DIMATrack outperforms the previous top-performing trackers across most metrics (80.7% MOTA, 79.0% IDF1, and 31.2 FPS) with minimal additional computational requirements. It is noteworthy that our method is primarily tailored to tackle the difficulties associated with object clustering and intricate motion patterns. Nevertheless, even when applied to the MOT17 dataset, which represents a more general and easier scenario of linear motion patterns, our method consistently exhibits enhanced tracking performance.

4.3.2 MOT20

In the MOT20 validation, our method exhibits superior performance, as depicted in Table 2 coupled with high inference speed. Notably, our approach outperforms state-of-the-art methods across all metrics (77.1% MOTA, 78.2% IDF1, and 15.3 FPS). These results underscore the effectiveness, robustness, and generalization of our proposed method in effectively capturing weak cues amidst scenarios involving clustering, heavy occlusion, and dense objects.

4.3.3 DanceTrack

In contrast to the preceding state-of-the-art heuristic trackers, our tracker demonstrates notably superior performance, boasting a 92.5% MOTA score and 79.9% IDF1 score. Importantly, this achievement is attained with equivalent association inputs and nearly identical computational complexity, as detailed in Table 3. These findings serve as compelling evidence that the integration and consideration of various weak cues, including width and height state parameters, offer an effective and efficient means of resolving ambiguous and erroneous matches that might elude traditional strong cue-based approaches.

Table 2. Comparison results on MOT20 dataset.

Tracker	MOTA (↑)	IDF1 (↑)	FP (↓)	FN (↓)	FPS (↑)
TrackFormer [18]	68.6%	65.7%	**20348**	140373	-
FairMOT [38]	61.8%	67.3%	103440	98901	13.2
MOTRv2 [39]	76.2%	72.2%	-	-	-
CSTrack [16]	66.6%	68.6%	25404	144358	4.5
SimpleTrack [12]	72.6%	70.2%	25515	114463	-
GSDT [28]	67.1%	67.5%	31913	135409	0.9
TransTrack [14]	65.0%	59.4%	28566	151377	7.2
TransCenter [32]	67.7%	58.9%	54967	108376	8.4
RelationTrack [35]	67.2%	70.5%	61134	104597	4.4
Hybrid-SORT [33]	76.7%	76.2%	-	-	-
StrongSORT++ [9]	73.8%	77.0%	16632	117920	1.4
GHOST [22]	73.7%	75.2%	-	-	-
UCMCTrack [34]	75.7%	77.4%	-	-	-
OC-SORT [5]	75.5%	75.9%	18100	108000	-
DIMATrack (Ours)	**77.1%**	**78.2%**	49615	**90245**	**15.3**

4.3.4 Visualization Results

In addition to the above quantitative results, we visualize DIMATrack tracking performance on two MOT benchmarks, showcasing its effectiveness in real-world scenarios. In Figure 4, the results of MOT17-02 sequence highlight our method's ability to accurately assign identities, even when pedestrians cross paths. The results of MOT17-06 sequence illustrate our robust performance under significant scale variations. In Figure 5, the results of two MOT20 sequences demonstrate our capability to maintain correct identities and bounding boxes in highly crowded scenes and fast dynamic motions. Overall, DIMATrack effectively handles identity assignment and bounding boxes localization for heavily occluded and overlapping objects, while also performing well in scenarios with diverse poses.

4.4. Ablation Study

4.4.1 DIMA-IoU

We posit that incorporating information about the height and width states can enhance data association. To this end, we proposed the utilization of the average of height-aware and width-aware IoUs as substitutes for conventional IoU. As demonstrated in Table 4, our suggested approach of averaging the height and width states yields superior benefits for data association compared to employing the conventional IoU, width-aware and height-aware data association methods individually. Employing height-aware and width-aware separately shows lower performance on the MOTA metric but better performance on the IDF1 metric on the MOT17 dataset compared to conventional IoU. This discrepancy arises because MOTA is primarily influenced by detection performance. However, Using DIMA-IoU, we observe higher performance on both MOTA and IDF1 metrics. This is attributed to the height state undergoing relatively short and continuous changes during actions such as squatting or standing up, which can be effectively modeled by the Kalman Filter. Conversely, the width state presents challenges for precise estimation by the Kalman Filter during pose changes, limb movements, or posing. However, by leveraging the average of these states, significant improvements in MOT data association are achieved.

When applying the height-aware and width-aware IoUs individually, it could potentially impact data association. For example, HAIoU might face challenges when the target object is occluded by objects with greater height. Similarly, WAIoU could be affected during rapid motion, making it difficult to obtain an accurate measurement. Therefore, we leverage the strengths of both HAIoU and WAIoU through their average, thus advancing the performance. As observed in Table 4, our DIMA-IoU outperforms the individual height-aware and width-aware IoUs.

4.4.2 Generality Across Other Trackers

We apply our DIMA-IoU to several representative tracking-by-detection trackers, namely SORT [2], DeepSORT [30], and ByteTrack [37]. Among these, SORT and ByteTrack rely solely on spatial information, while DeepSORT jointly utilizes both spatial and appearance information. The results are presented in Table 5, where significant improvements can be observed on both MOT17 and MOT20 datasets across all three trackers. For instance, our design, utilizing the average of height and width aware IoUs, improves SORT's MOTA by 23.8% and 8.2% on MOT17 and MOT20, respectively, and it boosts DeepSORT by 1.9% and 1.2%, respectively. These results provide compelling ev-

Table 3. Comparison results on DanceTrack dataset.

Tracker	MOTA (↑)	IDF1 (↑)	DetA (↑)	AssA (↑)
MOTR [36]	79.7%	51.5%	73.5%	40.2%
FairMOT [38]	82.2%	40.8%	66.7%	23.8%
CenterTrack [40]	86.8%	35.7%	78.1%	22.6%
UCMCTrack [34]	88.9%	65.0%	-	51.3%
GHOST [22]	91.3%	57.7%	81.1%	39.8%
MeMOTR [10]	89.9%	71.2%	80.5%	58.4%
ColTrack [15]	92.2%	77.3%	-	66.9%
OC-SORT [5]	92.0%	54.6%	84.4%	40.4%
StrongSORT++ [9]	91.1%	55.2%	80.7%	38.8%
MOTRv2 [39]	92.1%	76.0%	83.7%	64.4
Hybrid-SORT [33]	91.8%	67.4%	-	-
DIMATrack (Ours)	**92.5%**	**79.9%**	**84.4%**	**65.1%**

MOT17-02

MOT17-06

Figure 4. Visualization results of DIMATrack on MOT17 dataset. Two video sequences are selected to illustrate the results of sampled frames in chronological order. The bounding boxes and identities are marked, with the same box color representing the same identity.

idence that our insight of introducing weak cues such as height state and width state as compensation for strong cues is effective and generalizes well across different trackers and scenarios. Moreover, our method can be readily applied to existing trackers in a plug-and-play and training-free manner for enhanced performance.

5. Conclusion

In this paper, we present a novel online tracker, namely DIMATrack, to address the inherent limitations of current state-of-the-art MOT methods. Operating within the tracking-by-detection paradigm, DIMATrack leverages the Kalman Filter for robust trajectory prediction. It also introduces a simple yet powerful data association metric, DIMA-IoU, specifically designed to manage challenging scenarios characterized by high object overlap and diverse poses. By integrating both height and width information, DIMATrack improves association accuracy beyond that of conventional IoU.

Extensive experiments validate DIMATrack's superior generalization capability across a variety of trackers and scenarios. Employing standard CLEAR MOT metrics, DIMATrack not only outperforms existing state-of-the-art trackers but also simplifies the data association process, making it faster and more efficient. The simplicity, online operational capability, and robust generalization potential of DIMATrack render it an excellent solution for a wide range of MOT applications, especially those requiring real-time processing with limited computational resources.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (62376287), International Science and Technology Innovation Joint Base of Machine

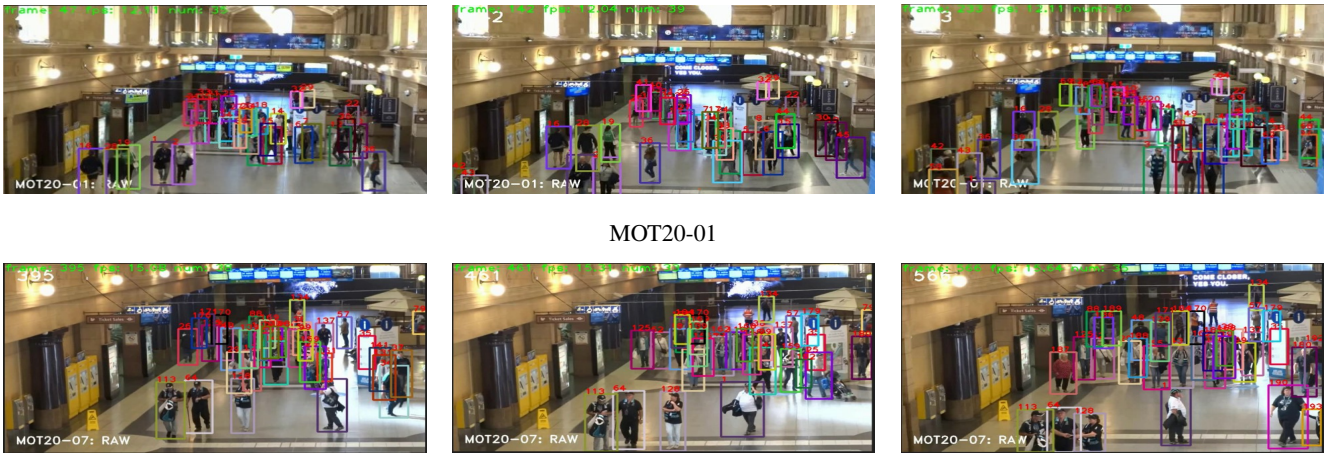

Figure 5. Visualization results of DIMATrack on MOT20 dataset

Table 4. Results of different IoUs on MOT17 validation set.

IoU type	MOTA (↑)	IDF1 (↑)
Conventional IoU	80.3%	77.3%
WAIoU	79.7%	77.9%
HAIoU	79.9%	78.7%
DIMA-IoU	**80.7%**	**79.0%**

Table 5. Results of applying DIMA-IoU to different tracking-by-detection trackers on two benchmarks in MOTA metric.

Tracker	DIMA-IoU	MOT17	MOT20
SORT [2]	-	33.4%	42.7%
	✓	**57.2%**	**50.9%**
DeepSORT [30]	-	78.0%	71.8%
	✓	**79.9%**	**73.0%**
ByteTrack [37]	-	80.3%	**77.8%**
	✓	**80.7%**	77.1%

Vision and Medical Image Processing in Hunan Province (2021CB1013), and the High Performance Computing Center of Central South University.

References

[1] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5

[2] 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)*, pages 3464–3468. IEEE, 2016. 1, 2, 4, 6, 7, 9

[3] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[4] R. G. Brown and P. Y. Hwang. *Introduction to Random Signals and Applied Kalman Filtering with Matlab Exercises*. John Wiley & Sons, 1997. 2, 3

[5] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 1, 6, 7, 8

[6] L. Chen, H. Ai, Z. Zhuang, and C. Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *2018 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2018. 3

[7] P. Dendorfer, A. Osep, A. Milan, K. Schindler, D. Cremers, I. Reid, S. Roth, and L. Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129:845–881, 2021. 5

[8] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 5

[9] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 25:8725–8737, 2023. 2, 6, 7, 8

[10] R. Gao and L. Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9901–9910, 2023. 6, 8

[11] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[12] J. Li, Y. Ding, H.-L. Wei, Y. Zhang, and W. Lin. Simpletrack: Rethinking and improving the jde approach for multi-object tracking. *Sensors*, 22(15):5863, 2022. 2, 6, 7

[13] C. Liang, Z. Zhang, X. Zhou, B. Li, and W. Hu. One more check: making "fake background" be tracked again. In *Pro-*

ceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 1546–1554, 2022. 2

[14] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, and W. Hu. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Transactions on Image Processing*, 31:3182–3196, 2022. 1, 7

[15] Y. Liu, J. Wu, and Y. Fu. Collaborative tracking learning for frame-rate-insensitive multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2023. 6, 8

[16] Z. Lu, V. Rathod, R. Votel, and J. Huang. Retinatrack: Online single stage joint detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2020. 2, 6, 7

[17] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, pages 4321–4329, 2019. 3

[18] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022. 3, 6, 7

[19] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 164–173, 2021. 3, 6

[20] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020. 2

[21] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[22] J. Seidenschwarz, G. Brasó, V. C. Serrano, I. Elezi, and L. Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13813–13823, 2023. 1, 6, 7, 8

[23] P. Sun, J. Cao, Y. Jiang, Z. Yuan, S. Bai, K. Kitani, and P. Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 5

[24] P. Sun, Y. Jiang, E. Xie, W. Shao, Z. Yuan, C. Wang, and P. Luo. What makes for end-to-end object detection? In *International Conference on Machine Learning*, pages 9934–9944. PMLR, 2021. 2

[25] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021. 2

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 1–11, 2017. 3

[27] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. 1, 5

[28] Y. Wang, K. Kitani, and X. Weng. Joint object detection and multi-object tracking with graph neural networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13708–13715. IEEE, 2021. 7

[29] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang. Towards real-time multi-object tracking. In *European Conference on Computer Vision*, pages 107–122. Springer, 2020. 2, 3

[30] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 2, 3, 4, 7, 9

[31] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12352–12361, 2021. 5

[32] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus, and X. Alameda-Pineda. Transcenter: Transformers with dense representations for multiple-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7820–7835, 2022. 7

[33] M. Yang, G. Han, B. Yan, W. Zhang, J. Qi, H. Lu, and D. Wang. Hybrid-sort: Weak cues matter for online multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6504–6512, 2024. 4, 6, 7, 8

[34] K. Yi, K. Luo, X. Luo, J. Huang, H. Wu, R. Hu, and W. Hao. Ucmctrack: Multi-object tracking with uniform camera motion compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6702–6710, 2024. 6, 7, 8

[35] E. Yu, Z. Li, S. Han, and H. Wang. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*, 2022. 6, 7

[36] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022. 3, 6, 8

[37] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 1, 2, 3, 4, 5, 6, 7, 9

[38] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 2, 6, 7, 8

[39] Y. Zhang, T. Wang, and X. Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 22056–22065, 2023. 6, 7, 8

[40] X. Zhou, V. Koltun, and P. Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 2, 6, 8

[41] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2