http://dx.doi.org/10.62762/TETAI.2024.240529

RESEARCH ARTICLE

CT-DETR and ReID-Guided Multi-Target Tracking Algorithm in Complex Scenes

Ming Gao[1,*] **and Shixin Yang**[2]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
[2] School of Computer Science and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China

Abstract

In the era of rapid technological advancement, the demand for sophisticated Multi-Object Tracking (MOT) systems in applications such as intelligent surveillance and autonomous navigation has become increasingly critical. However, existing models often struggle with accuracy and efficiency in densely populated or dynamically complex environments. Addressing these challenges, we introduce a novel deep learning-based MOT model that incorporates the latest CT-DETR detection technology and an advanced ReID module for improved pedestrian tracking. Experimental results demonstrate the model's superior performance in accurately identifying and tracking multiple targets across varied scenarios, significantly outperforming existing benchmarks. This research not only marks a significant leap forward in the field of video surveillance technology but also lays a foundational framework for future advancements in intelligent system applications, underscoring the importance of innovation in deep learning methodologies for real-world challenges.

Keywords: multi-object tracking, deep learning, CT-DETR, pedestrian re-identification, intelligent surveillance systems.

1 Introduction

Recently, the widespread use of video data has driven the continuous demand for and ongoing development of Multiple Object Tracking (MOT) technology [1]. Multiple Object Tracking is a critical computer vision task aimed at identifying, localizing, and tracking multiple moving objects from video sequences, and inferring their relationships and behaviors. This technology finds extensive applications across various domains, including but not limited to intelligent surveillance, autonomous driving, robot navigation, medical image analysis, and sports analytics [2]. With the continuous improvement of sensor technology and computing power, coupled with the rapid advancement of deep learning methods, the importance and effectiveness of Multiple Object Tracking technology in practical applications are becoming increasingly prominent [3].

Pedestrian target tracking is of significant importance in today's digital age. This technology not only aids in real-time monitoring and tracking of pedestrian activities in intelligent surveillance systems but also finds applications in urban traffic management, intelligent transportation systems, and

Submitted: 23 November 2023
Accepted: 21 May 2024
Published: 29 May 2024

Vol. 1, **No.** 1, 2024.
10.62762/TETAI.2024.240529

***Corresponding author:**
✉ Ming Gao
aarongaoming@ieee.org

Citation
Gao, M., & Yang, S. (2024). CT-DETR and ReID-Guided Multi-Target Tracking Algorithm in Complex Scenes. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 1(1), 44–57.

© 2024 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

human-computer interaction fields [4]. In intelligent surveillance systems, pedestrian target tracking helps ensure public safety, monitor access to critical areas, and promptly identify abnormal behaviors. In urban traffic management, it can be utilized to monitor pedestrian movement trajectories and traffic flow, thereby optimizing traffic signal control and enhancing traffic efficiency and safety. Additionally, pedestrian target tracking assists autonomous vehicles in sensing pedestrian positions and dynamics, thus avoiding traffic accidents [4]. In the realm of human-computer interaction, pedestrian target tracking technology provides fundamental support for intelligent services and augmented reality applications, improving user experiences and driving the development and application of intelligent technologies. Hence, pedestrian target tracking technology holds broad prospects and significant societal importance in the digital era [5, 16].

With the advancement of single-object tracking technology, research on multiple-object tracking has gradually branched into two main directions: online tracking and batch tracking. Online tracking primarily consists of two key steps: target detection and Re-identification (ReID) [6]. Target detection involves identifying potential targets in each frame of an image, typically using object detection algorithms such as Convolutional Neural Networks (CNN) or other deep learning models. ReID, on the other hand, is about re-identifying targets across different time steps to ensure consistency in the tracking process. This process often involves steps like feature extraction, feature matching, and similarity measurement to accurately associate targets across different time steps. Batch tracking, meanwhile, adopts a different strategy. It does not update in real-time at each time step but optimizes tracking results by processing the entire video sequence. Batch tracking methods typically utilize spatial and temporal consistency among targets for global optimization, aiming to improve tracking accuracy and stability [7]. In batch tracking, target detection is initially required, similar to online tracking methods. However, instead of being used solely for tracking in the current frame, detected targets are integrated into a global optimization process to maintain consistency across the entire video sequence. This process often involves techniques such as data association and trajectory graph optimization.

However, in the past, most tracking networks have focused on using anchor-based networks for two-step detection and tracking methods to extract object features. However, recent research has found that **anchor-free** detection networks offer superior tracking capabilities by directly predicting object centroids and bounding boxes without relying on predefined anchors [7], enabling more flexible handling of overlapping and densely packed targets. During the tracking process, when targets overlap, anchor-based detection networks may encounter difficulties in clearly separating different targets. This ambiguity can lead to tracking algorithms erroneously considering overlapping targets as a single target or having difficulty accurately distinguishing them when targets intersect. Therefore, although anchor-based detection networks can provide excellent tracking performance in many cases, addressing this challenge, anchor-free frameworks that jointly optimize detection and re-identification in a single network—such as FairMOT [13]—have demonstrated that fair treatment of both subtasks substantially reduces identity switches and improves robustness in overlapping target scenarios, motivating the anchor-free design adopted in this paper.

To address these issues, this paper integrates the CT-DETR (Center and Tracked-DETR) framework with JDE (Joint Detection and Embedding) into a single network for object detection and tracking. Targets are detected by identifying their centroids and bounding boxes, followed by target association using ReID (Re-identification) techniques. By leveraging feature-based similarity metrics to link existing trajectories, the proposed method significantly improves tracking accuracy.

- This paper proposes an anchor-free detection framework that no longer relies on predefined anchor points but directly detects the centroids and bounding boxes of targets. This new detection framework is more flexible in handling overlapping targets and can more accurately segment different targets, bringing new ideas and methods to the field of multiple object tracking.
- The ByteTrack model, which is a video-based multi-object tracking model, is optimized in this paper. By integrating the CT-DETR framework and JDE model, ByteTrack is further improved to enhance its tracking accuracy and efficiency, making it more suitable for multi-object tracking tasks in real-world scenarios.
- Extensive experiments have been conducted to validate the effectiveness of the proposed methods. The experimental results confirm their

effectiveness and practicality, providing valuable references for both research and real-world applications in the field of multiple object tracking.

The structure of this paper comprises several key sections: related work, methodology, experimental results, and conclusion. In the related work section, we investigated existing literature and methodologies in pedestrian target tracking. The methodology section outlines our proposed approach, including the implementation of the tracking algorithm and any other innovations. Subsequently, the experimental section introduces the results and evaluation of our approach. Finally, the conclusion summarizes the research findings and discusses implications for future research and applications.

2 Related Work

2.1 The research of object detection

In the realm of multi-object tracking, target detection techniques play a pivotal role. These techniques serve as the foundation for identifying and localizing objects within a scene, a prerequisite for subsequent tracking processes [28, 31]. Notably, the emergence of deep learning architectures has significantly advanced the state-of-the-art in target detection. Among these architectures, YOLO (You Only Look Once) models [29, 30] have gained considerable attention for their real-time performance and accuracy. Noteworthy variants such as YOLOv5, YOLOv8, and YOLOv9 have been introduced, each refining the original architecture to achieve better performance in terms of speed, accuracy, or both.

Beyond YOLO models, recent advancements have been made in transformer-based object detection and tracking methods. One notable approach is DETR (DEtection TRansformer) [32], which leverages transformer architectures to directly predict object bounding boxes and categories in an end-to-end manner. Variants such as Deformable DETR, PnP-DETR, and RT-DETR have further extended the capabilities of transformer-based approaches, addressing shortcomings such as scalability, robustness to occlusions, and handling deformable objects [33].

Each of these algorithms has its strengths and limitations. YOLO models excel in real-time performance and simplicity; however, their detection outputs feed directly into downstream ReID modules, where imperfect detections or noisy training labels can substantially degrade re-identification quality. Robust ReID under such conditions has been addressed through dedicated noise-tolerant learning paradigms [8], underscoring the tight coupling between upstream detection reliability and downstream appearance matching performance in end-to-end tracking systems. Transformer-based approaches offer end-to-end learning capabilities and global context modeling, but they may suffer from increased computational complexity and training data requirements. Understanding these trade-offs is crucial for selecting the most suitable algorithm for specific tracking tasks.

2.2 Research on multi-object tracking

In the current landscape of multi-object tracking algorithms, several notable approaches have emerged, each with its strengths and limitations. SORT (Simple Online and Realtime Tracking) is a popular method known for its simplicity and efficiency [26]. It operates on a tracking-by-detection framework, associating object detections across frames based on motion models and appearance features. SORT is renowned for its real-time performance and effectiveness in tracking multiple objects simultaneously. However, its performance may degrade in scenarios with heavy occlusions or complex interactions between objects due to its reliance on simple motion and appearance models. DeepSORT [10] (Deep Simple Online and Realtime Tracking) builds upon SORT by integrating deep learning techniques for improved object appearance modeling and feature extraction. Its practical applicability has been demonstrated in people tracking deployments across real-world surveillance scenarios [27]. By leveraging deep neural networks, DeepSORT achieves better robustness to occlusions and appearance variations, resulting in more accurate and reliable tracking performance compared to traditional SORT. However, DeepSORT may suffer from higher computational costs and resource requirements due to the complexity of deep learning models. MOTDT (Multiple Object Tracking with Deep Learning and Tracklet) is another notable algorithm that combines deep learning methods with tracklets, short track segments, to improve tracking performance. MOTDT excels in handling complex scenarios with occlusions and crowded environments by leveraging deep feature representations and tracklet association techniques. However, its performance may be affected by issues such as track fragmentation and identity switches in highly dynamic scenes [14].

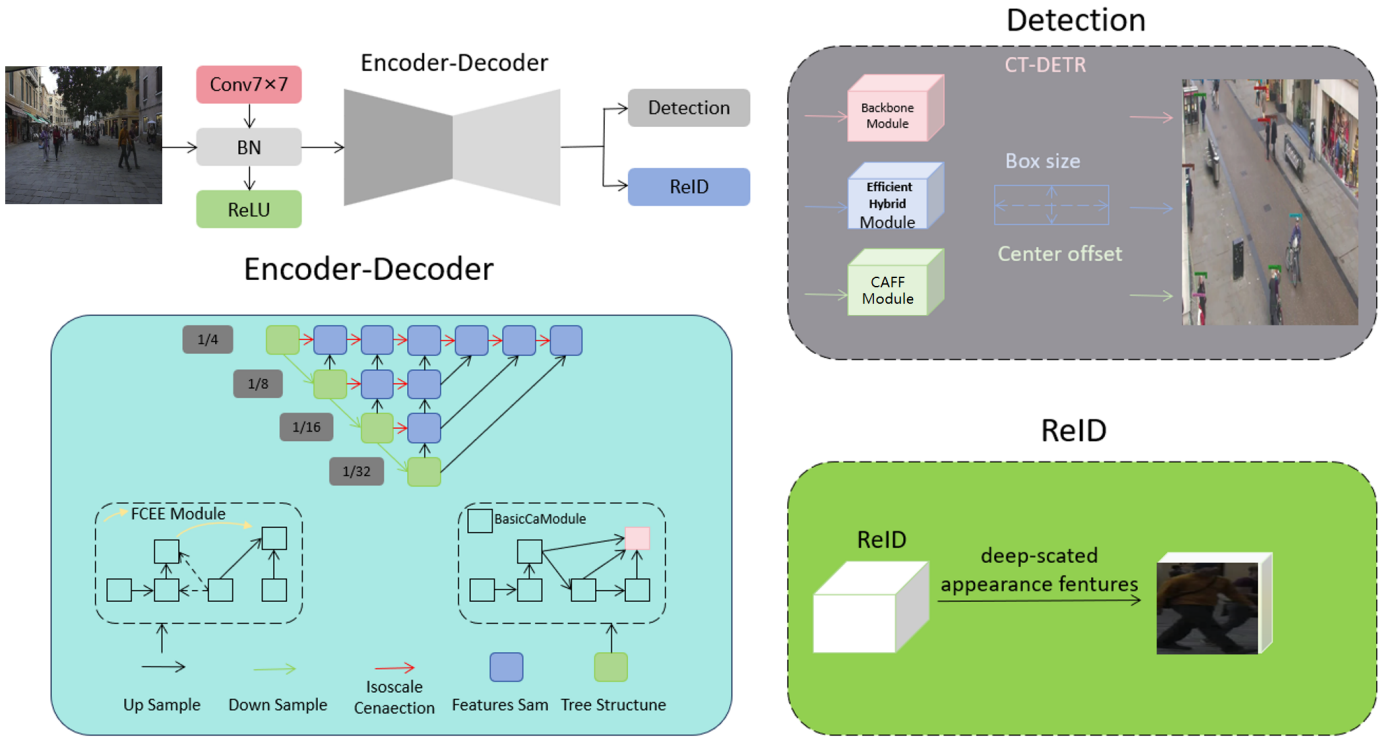


Figure 1. Overall structure diagram of the proposed model.

JDE (Joint Detection and Embedding) integrates object detection and feature embedding into a unified framework for multi-object tracking. By jointly optimizing detection and embedding tasks, JDE achieves superior performance in associating object detections and maintaining object identities over time. This approach benefits from end-to-end learning and feature embedding, resulting in robust tracking performance across various challenging scenarios. However, JDE may suffer from increased computational complexity and training data requirements compared to simpler tracking methods.

Bytetrack is a recent addition to the multi-object tracking landscape, leveraging advanced techniques such as deep learning, attention mechanisms, and graph neural networks. Bytetrack aims to address the limitations of existing methods by incorporating contextual information and global dependencies into the tracking process, leading to improved robustness and accuracy. However, its effectiveness in real-world scenarios and computational efficiency relative to traditional methods remain areas of ongoing research and evaluation [15].

Overall, while each of these algorithms offers distinct advantages in multi-object tracking, they also exhibit specific limitations that need to be addressed for more comprehensive and effective tracking solutions [20–22]. The progression from simpler methods like SORT to more advanced approaches like Bytetrack reflects the continuous evolution of multi-object tracking in response to complex real-world challenges. Among the most persistent difficulties is occlusion, which disrupts appearance consistency across frames and demands robust re-identification strategies to maintain target identities under partial or full occlusion [7, 9, 23].

3 Methodology

Our overall method framework is illustrated in Figure 1 and consists primarily of two components: the detection module and the re-identification (ReID) module [11]. The ReID module is designed to handle diverse appearance challenges: beyond standard pose and illumination variation, appearance-invariant feature learning becomes especially critical in extreme scenarios such as clothing change across camera views [12], motivating the use of deep component reconstruction strategies to capture identity-discriminative representations that remain stable under such conditions. Firstly, the detection module employs the CT-DETR model for pedestrian target detection. The CT-DETR model, based on Transformer architecture, accurately detects pedestrian targets in images and videos with high efficiency, making it suitable for handling large-scale

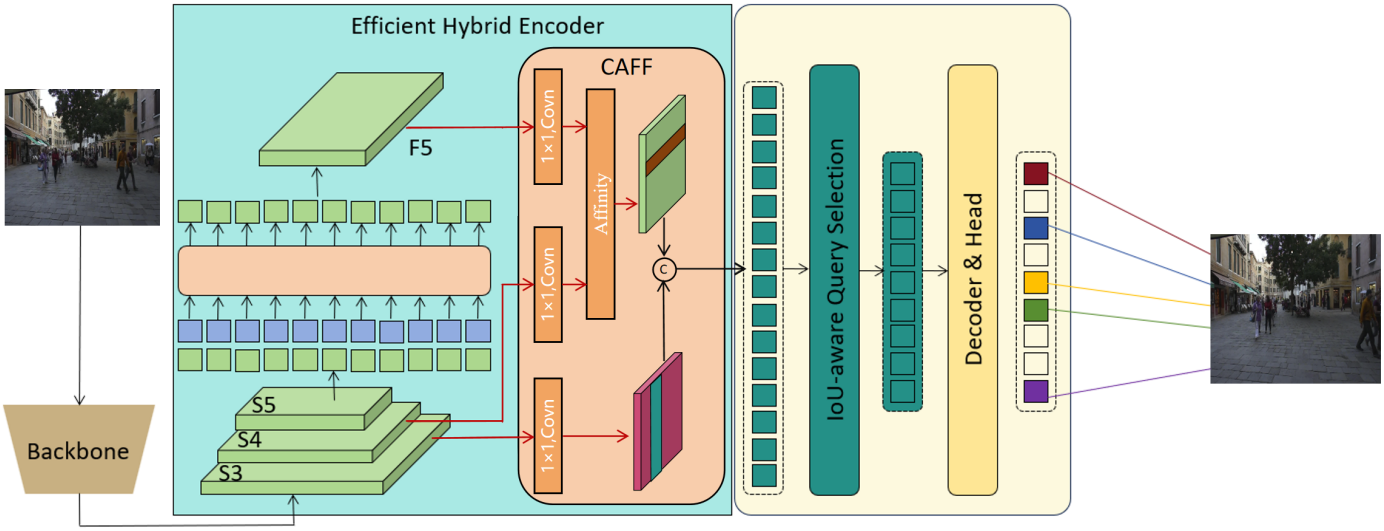


Figure 2. Schematic diagram of the CT-DETR model architecture.

real-time surveillance scenarios.

Secondly, the ReID module addresses the pedestrian target re-identification task. In this module, we adopt a deep learning approach that comprehensively utilizes both semantic [25] and spatial information, along with a fusion technique for shallow and deep features. Through these methods, our network can capture pedestrian target features more comprehensively, thereby enhancing re-identification accuracy and robustness.

In summary, our approach combines the detection and reID modules, utilizing advanced models and techniques to achieve efficient detection and accurate re-identification of pedestrian targets. This provides strong support for applications in real-time surveillance systems and intelligent transportation systems, among others.

3.1 CT-DETR Model

The complete architecture of the CT-DETR model comprises a backbone network, a hybrid encoder, and a transformer decoder with auxiliary prediction heads. The backbone network serves as the foundational component, responsible for extracting feature representations of input images, utilizing a pre-trained EfficientNet. The hybrid encoder, on the other hand, undertakes the task of feature encoding, converting the feature maps extracted by the backbone network into more expressive and semantically understandable feature representations. It consists of a series of Transformer encoders, encoding feature vectors into high-dimensional representations through self-attention mechanisms and fully connected layers. The transformer decoder with auxiliary prediction heads is a crucial component responsible for translating the encoded feature maps into the results of target detection. It consists of multiple layers of Transformer decoders, each layer containing self-attention mechanisms and cross-attention mechanisms. The auxiliary prediction heads provide additional supervisory signals, contributing to the convergence and stability of model training. By integrating these components, the CT-DETR model achieves efficient detection and re-identification of targets in complex scenes. The network structure is illustrated in Figure 2, where initially, features from the last three stages of the backbone network S3, S4, S5 are utilized as the input to the encoder. The efficient hybrid encoder transforms multi-scale features into a sequence of image features through Scale-Invariant Feature Interaction (AIFI) and Cross-Attention Feature Fusion Module (CAFF).

3.1.1 Efficient Hybrid Encoder

The proposed encoder comprises two modules: the Attention-based Intra-scale Feature Interaction (AIFI) module and the Convolutional Neural Network-based Cross-Attention Feature Fusion (CAFF) [17]. Building upon variant D, AIFI further reduces computational redundancy by confining intra-scale interactions solely to S5. We contend that applying self-attention operations to higher-level features with richer semantic concepts enables capturing relationships among conceptual entities in images, thereby facilitating subsequent modules in object detection and recognition.

Specifically, the fusion block consists of N RepBlocks,

and the outputs of the two paths are fused via element-wise addition. We can describe this process as follows:

$$\begin{aligned} \mathbf{Q} = \mathbf{K} = \mathbf{V} &= Flatten(S_5) \\ F_5 &= Reshape(Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V})) \\ Output &= CAFF(\{S_3, S_4, F_5\}) \end{aligned} \tag{1}$$

where $Attn$ stands for multi-head self-attention, while $Reshape$ is used to restore the shape of features to match that of S5, serving as the inverse operation of $Flatten$.

3.1.2 IoU-aware Query Selection

We propose the method of IoU-aware Query Selection, wherein during training, the model is constrained to generate high classification scores for features with high IoU scores, and low classification scores for features with low IoU scores. This approach helps the model to more effectively utilize IoU information in the object detection task, thereby improving the accuracy of object localization and classification. Specifically, through IoU-aware Query Selection, the model can focus more on features closely related to the target, while appropriately downweighting features with lower relevance to the target, thus effectively enhancing the performance of object detection. This mechanism introduces a finer IoU-aware adjustment during training, enabling the model to intelligently handle object detection tasks in various scenarios, thereby improving its robustness and generalization capability. We redefine the optimization objective of the detector as follows:

$$\begin{aligned} \mathcal{L}(\hat{y}, y) &= \mathcal{L}_{box}(\hat{b}, b) + \mathcal{L}_{cls}(\hat{c}, \hat{\boldsymbol{b}}, y, \boldsymbol{b}) \\ &= \mathcal{L}_{box}(\hat{b}, b) + \mathcal{L}_{cls}(\hat{c}, c, IoU) \end{aligned} \tag{2}$$

where $\hat{y} = \{\hat{c}, \hat{b}\}$ and $y = \{c, b\}$ represent predicted and true values, c stands for class and b for bounding box. We introduce the IoU score into the classification branch's objective function, akin to VFL, to ensure consistency in positive sample classification and localization.

3.2 Self-Attention

In our CT-DETR model, the Transformer plays a crucial role. This architecture enables the model to effectively capture long-distance dependencies within the image [18, 19, 24], whether they are spatial or temporal. By simultaneously calculating attention scores for each position in the image relative to every other position,

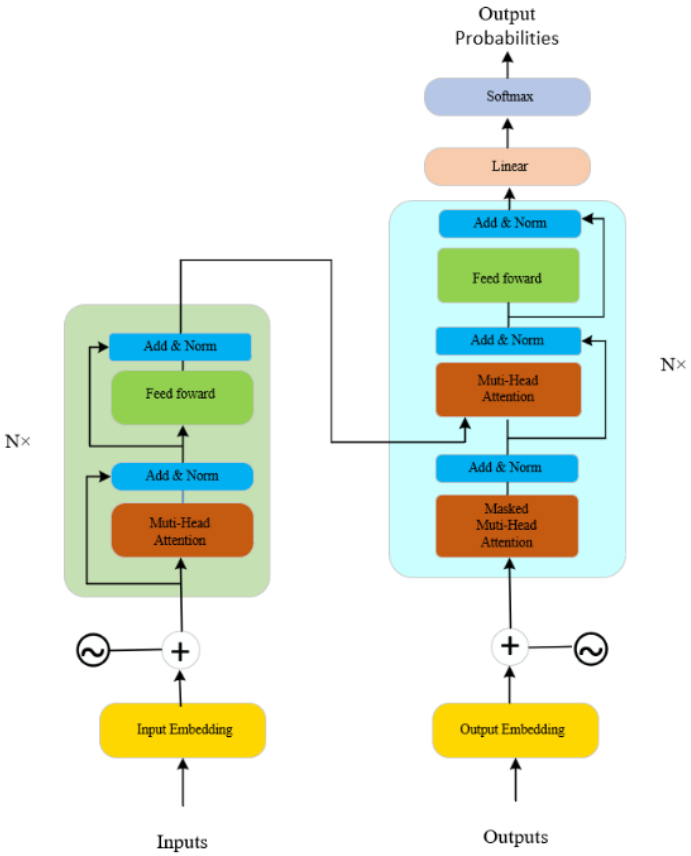


Figure 3. Schematic diagram of the Transformer model structure.

the Transformer endows the backbone network of CT-DETR with a powerful global understanding capability. This not only increases the model's flexibility in processing input sequences of varying lengths but also significantly boosts the efficiency of parallel computations, thereby accelerating training speed. Most importantly, the introduction of the Transformer significantly enhances the model's ability to recognize key features within images, providing strong support for achieving high-precision object detection and identification. The network architecture diagram of the Transformer is shown in Figure 3.

$$Q = W^Q X \tag{3}$$

where Q is the matrix of queries, X is the input representation (e.g., word embeddings), and W^Q is the weight matrix for queries.

$$K = W^K X \tag{4}$$

where K is the matrix of keys, X is the input representation, and W^K is the weight matrix for keys.

$$V = W^V X \tag{5}$$

where V is the matrix of values, X is the input representation, and W^V is the weight matrix for values.

$$A = \frac{QK^T}{\sqrt{d_k}} \tag{6}$$

where A represents the attention scores matrix, QK^T is the dot product of queries and keys transpose, and $\sqrt{d_k}$ is the scaling factor, with d_k being the dimension of the key vectors.

$$A' = \text{softmax}(A) \tag{7}$$

where A' is the softmax-normalized attention scores, ensuring all the attention scores are positive and sum up to 1.

$$O = A'V \tag{8}$$

where O is the output matrix of the self-attention mechanism, and V is the matrix of values.

$$\text{Output} = W^O O \tag{9}$$

where Output is the final output of the self-attention layer, O is the output matrix from the self-attention mechanism, and W^O is the output weight matrix.

3.3 Loss Function Design

The CT-DETR model is optimized with a composite multi-task loss that jointly supervises three objectives: center point localization, bounding box regression, and IoU-aware classification.

Center Point Heatmap Loss. The "C" (Center) component of CT-DETR introduces an auxiliary center point prediction head that localizes object centroids via Gaussian heatmap representations. The ground-truth heatmap is constructed by placing a Gaussian kernel at each annotated centroid:

$$Y_{xy} = \exp\left(-\frac{\left(x - P_{d(x)}\right)^2 + \left(y - P_{d(y)}\right)^2}{2\sigma^2}\right) \tag{10}$$

A modified focal loss is applied to handle the severe foreground-background imbalance in the heatmap:

$$L_{\text{heatmap}} = -\frac{1}{N}\sum_{xy}\begin{cases}(1-\hat{Y}_{xy})^{\alpha}\ln(\hat{Y}_{xy}), & Y_{xy}=1\\ (1-Y_{xy})^{\beta}(\hat{Y}_{xy})^{\alpha}\ln(1-\hat{Y}_{xy}), & Y_{xy}\neq 1\end{cases} \tag{11}$$

Bounding Box Regression Loss. The offset and size of each detected object are supervised with an ℓ_1 loss:

$$L_{\text{box}} = \sum_{i=1}^{N}\left\|O^i - \hat{O}^i\right\|_1 + \left\|S^i - \hat{S}^i\right\|_1 \tag{12}$$

Total Loss. Combining the above with the IoU-aware classification loss defined in Equation (2), the full training objective is:

$$L_{\text{total}} = L_{\text{heatmap}} + \lambda_1 L_{\text{box}} + \lambda_2 L_{\text{cls}} \tag{13}$$

where L_{cls} corresponds to the IoU-aware classification loss $\mathcal{L}_{\text{cls}}$ defined in Eq. (2), and λ_1, λ_2 are scalar weighting coefficients balancing the three loss terms.

3.4 Pedestrian reidentification

Pedestrian re-identification (ReID) is a specific computer vision task aimed at recognizing and tracking the appearance of the same pedestrian across a network of cameras. Effective ReID demands not only accurate modeling of target appearance features but also principled selection of candidate associations — learning to focus on true targets while suppressing distractors and irrelevant background objects [20, 21]. This capability is crucial for fields such as video surveillance, human traffic analysis, and social security. Pedestrian Re-ID systems work by extracting and comparing the appearance features of pedestrians from different video frames or different camera angles. These appearance features include, but are not limited to, clothing color, style, body posture, and gait. The detailed information in this study is shown in the equations.

$$L_{\text{identity}} = -\sum_{i=1}^{N}\sum_{k=1}^{K} L^i(k)\log(p(k)) \tag{14}$$

where $L(k)$ is the expected probability distribution for the identity of the k^{th} target, where the number of categories is denoted by k. $p(k)$ represents the actual probability distribution for the identity of the k^{th} object.



Figure 4. Representative samples from the dataset.

4 Experiments

4.1 Dataset

Our experiment makes use of two prominent datasets for evaluating object tracking algorithms: 2DMOT2015 and OTB-100. The 2DMOT2015 dataset [34] serves as a comprehensive benchmark for multiple object tracking, offering a diverse array of challenging sequences derived from real-world scenarios. The effectiveness of this benchmark has been validated by deep learning-based real-time tracking methods that demonstrate consistent and measurable performance improvements across its standardized test sequences [35]. These sequences are annotated with ground truth data, including object positions and identities, enabling rigorous evaluation of tracking algorithms' performance under various conditions such as occlusions, scale variations, and object interactions. On the other hand, the OTB-100 dataset, also known as Object Tracking Benchmark-100, is specifically designed for single object tracking tasks [36]. It encompasses a collection of 100 video sequences, each presenting unique challenges like appearance changes, background clutter, and motion blur. Similar to the 2DMOT2015 dataset, the OTB-100 dataset provides ground truth bounding box annotations for each sequence, facilitating a thorough assessment of tracking algorithms' accuracy and robustness. These datasets collectively serve as invaluable resources for assessing the effectiveness and robustness of our proposed tracking algorithm across a wide range of tracking scenarios and challenges. Figure 4 displays examples from two datasets, and we extracted four frames for illustration. These examples highlight the diversity and complexity of the tracking scenarios considered in our evaluation.

4.2 Experimental Environment

In this experiment, we have meticulously documented the configuration of the experimental environment to ensure the reproducibility of the results. The specific parameters of the experimental environment are shown in Table 1.

4.2.1 Model training

Table 2 shows the parameter settings for CT-DETR, where each row lists a specific parameter used during the model's training and evaluation along with its corresponding value. The detailed parameters include a learning rate set to 0.01, indicating the step size at each iteration while moving toward the minimum of the loss function. The optimizer used is AdamW, a popular optimization method for training deep learning models. The batch size is set to 16, referring to the number of training examples used in one iteration. Weight decay is set to none, training epochs to 300, model parameters to 4,385,523, and the number of layers to 252. The image size is set to 640, seed values to 0, and early stopping is enabled (True).

Table 1. Experimental environment demonstrated.

Parameter	Configuration
CPU	Intel Core i7-12700KF
GPU	NVIDIA GeForce RTX 4090 (24 GB)
CUDA version	CUDA 11.8
Python version	Python 3.8.16
Deep learning framework	Pytorch 1.8.1
Operating system	Ubuntu 22.04.2

Table 2. Model parameter settings.

Parameter	Value
Learning Rate	0.01
Optimizer	AdamW
Batch Size	16
Weight Decay	None
Training Epochs	300
Model Parameters	4,385,523
Number of Layers	252
Image Size	640
Seeds	0
Early Stop	True

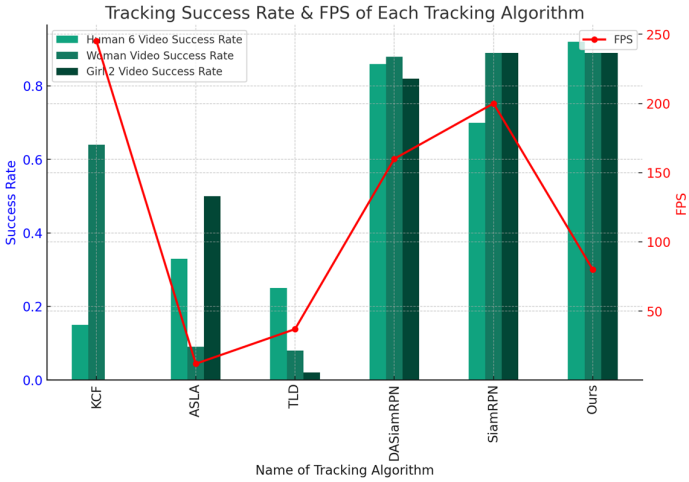


Figure 5. Model performance on the OTB-100 dataset.

4.2.2 *Evaluation Metrics*

In this experiment, we employed numerous evaluation metrics to thoroughly analyze the model's performance, as shown in the following formulas.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

where $Precision$ is the proportion of true positive predictions in all positive predictions, TP is the number of true positives, and FP is the number of false positives.

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

where $Recall$ measures the proportion of actual positives correctly identified, FN is the number of false negatives.

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

where $F1\text{-}Score$ is the harmonic mean of precision and recall, offering a balance between them.

$$TP = \text{Number of true positive predictions} \tag{18}$$



Figure 6. Visualization of ReID Results. Five frames extracted from Scene 258 are presented for demonstration.

where TP directly counts instances correctly identified as positive.

$$mAP = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k=1}^{n_q} P(k) \times rel(k) \tag{19}$$

where mAP (mean Average Precision) averages the precision scores at different recall levels, N is the number of queries, Q is the total number of classes, n_q is the number of relevant documents for the q^{th} query, $P(k)$ is the precision at cutoff k, and $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, 0 otherwise.

Table 3. Performance evaluation metrics.

Measure	Better	Description
MOTA	higher	Multiple Object Tracking Accuracy.
MOTP	higher	Multiple Object Tracking Precision.
IDF1	higher	The ratio of correctly identified detections.
MT	higher	Mostly tracked targets.
ML	lower	Mostly lost targets.
FP	lower	The total number of false positives.
FN	lower	The total number of false negatives.
ID Sw.	lower	The total number of identity switches.

4.3 Experimental Results and Analysis

In the evaluation of our pedestrian counting algorithm, we utilized the 2DMOT2015 benchmark for quantitative analysis of the algorithm's performance. The 2DMOT2015 is a recognized evaluation platform specifically designed for the fair comparison and assessment of multiple pedestrian tracking algorithms. This platform offers a series of detailed test video data encompassing various dynamic and complex urban environments. These settings include diverse lighting conditions, levels of crowd density, and camera movements, thereby providing rich testing scenarios for pedestrian detection and tracking algorithms. Each video contains precise ground truth data, which serves to verify the accuracy and consistency of each



Figure 7. Visualization of pedestrian flow tracking and counting results by the model.

Table 4. Tracking success rate list of each tracking algorithm in several pedestrian videos of OTB-100.

Name of Tracking Algorithm	Human 6 Video Success Rate	Woman Video Success Rate	Girl 2 Video Success Rate	FPS
KCF[37]	0.15	0.64	0.00	245
ASLA [38]	0.33	0.09	0.50	12
TLD [39]	0.25	0.08	0.02	37
DASiamRPN[40]	0.86	0.88	0.82	160
SiamIST [41]	0.70	**0.89**	**0.89**	200
Ours	**0.92**	**0.89**	**0.89**	80

pedestrian target identified by the algorithm. The specific description is shown in Table 3.

The model uses the OTB-100 (Object Tracking Benchmark-100) dataset to test accuracy. Since the model targets pedestrian tracking, videos containing pedestrians such as "Human 6 Video", "Woman Video", and "Girl 2 Video" were selected from the OTB-100 benchmark for evaluation. As shown in Table 4, the success rates and frames per second (FPS) of different tracking algorithms on various types of videos, including Human 6 Video, Woman Video, and Girl 2 Video. These algorithms have undergone testing on the OTB-100 dataset, which is a commonly used benchmark for evaluating tracking algorithm performance. In the Human 6 Video, our algorithm outperforms others with a success rate of 0.92, significantly higher than KCF (0.15), ASLA (0.33), TLD (0.25), and SiamIST (0.70), and also surpassing DASiamRPN (0.86). Similarly, on Woman Video and Girl 2 Video, our algorithm achieves excellent results with success rates of 0.89 each, matching SiamIST and DASiamRPN on Woman Video (0.88) and Girl 2 Video (0.82) respectively, while notably superior to KCF, ASLA, and TLD across all sequences. In terms of speed, although our algorithm's FPS is 80, lower than KCF (245), DASiamRPN (160), and SiamIST (200), it achieves significant advantages in tracking accuracy compared to these algorithms. Therefore, our algorithm demonstrates outstanding comprehensive performance in pedestrian tracking tasks. Figure 5 visualizes the table, providing a more intuitive representation of the performance of the

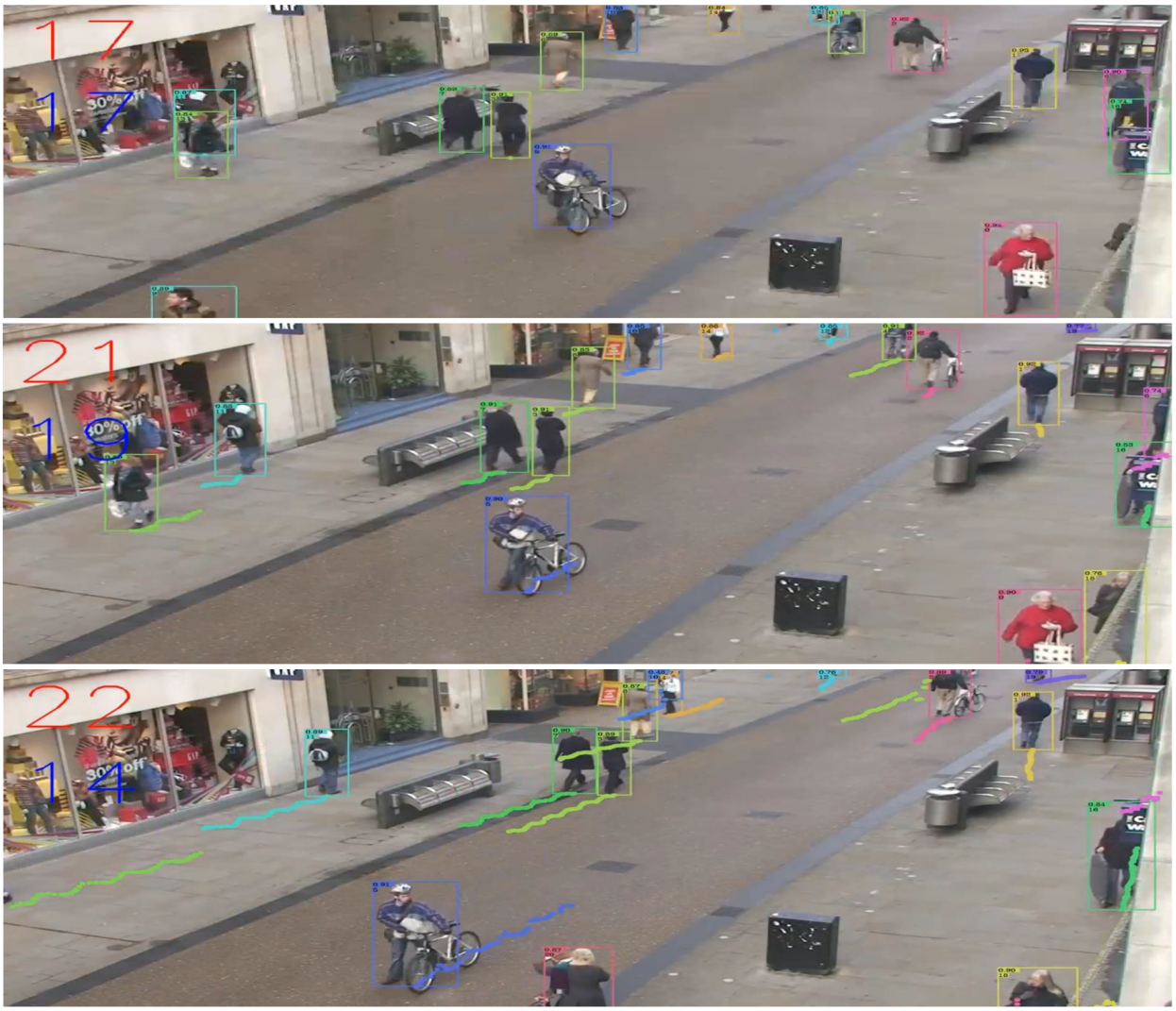


Figure 8. Visualization of pedestrian flow trajectory paths from tracking results.

model.

4.4 Discussion

The counting evaluation is conducted on five representative sequences selected from the 2DMOT2015 benchmark, covering diverse crowd densities and environmental conditions. Table 5 presents the evaluation results of pedestrian counting for these five video sequences. During the evaluation process, we counted the cases of missed detections (FN) and false positives (FP) based on manually observed counting results and calculated the Precision, Recall, and F1-Score based on these data to comprehensively evaluate the performance of the counting algorithm. In addition, the pedestrian count results for each video sequence (Seq1 to Seq5) are close to the actual count (GT), demonstrating the algorithm's high accuracy. The number of false positives (FP) and missed detections (FN) are relatively balanced across the sequences, with no extreme cases. The number of accurately detected pedestrians (TP) is close to the total count, further confirming the effectiveness of the algorithm. Overall, the Precision and Recall for all five sequences are above 93%, with Precision ranging from 96.09% to 98.26% and Recall from 93.98% to 96.49%, reflecting the algorithm's stability and high accuracy across different scenarios. The F1-Score, as a harmonic mean of Precision and Recall, provides a balanced metric for evaluation, with all sequences achieving an F1-Score above 95%. This indicates a good balance between accuracy and recall rate maintained by the algorithm. Summarizing the results of all sequences, the overall Precision reached 97.09%, Recall was 95.16%, and the F1-Score was 96.11%, further proving the high efficiency and accuracy of this pedestrian counting algorithm across different video sequences. Through this evaluation, we can consider the algorithm to

Table 5. The evaluation results of the algorithm in test videos.

Sequence	Count Result	GT	FP	FN	TP	Precision	Recall	F1-Score
Seq1	550	565	19	34	531	96.55%	93.98%	95.25%
Seq2	409	410	16	17	393	96.09%	95.85%	95.97%
Seq3	569	570	19	20	550	96.66%	96.49%	96.58%
Seq4	921	953	16	48	905	98.26%	94.96%	96.58%
Seq5	667	691	14	38	653	97.90%	94.50%	96.17%
TOTAL	3116	3189	84	157	3032	97.09%(Avg.)	95.16%(Avg.)	96.11%(Avg.)

be reliable and effective for pedestrian counting in practical applications, capable of meeting the accuracy requirements of most scenarios.

In this experiment, the ReID (Re-Identification) module played a crucial role as it was able to accurately identify and segment the flow of people in videos. In Figure 6, we presented a case study that demonstrates how the ReID module effectively recognizes specific pedestrian targets from complex scenes and continuously tracks these targets across different frames. This capability relies not only on the module's deep learning and understanding of pedestrian features but also on its sensitive capture of pedestrian behaviors and position changes in dynamic environments. By combining semantic and spatial information, as well as a fusion technique of both deep and shallow features, the ReID module can provide a more comprehensive representation of target features. This integrated feature extraction method not only improves the accuracy of identification but also enhances the system's robustness to variations in the target under different conditions (such as varying lighting conditions, occlusions, and the ability to distinguish between similar targets).

As shown in Figure 7, we conducted tests using our publicly released test videos, from which we accurately tracked pedestrian targets and performed count statistics. This demonstrated the efficiency and accuracy of our model. Through analyzing the publicly available test videos, we not only accurately tracked each pedestrian target but also conducted effective count statistics, proving our model's powerful capability in processing real-time video surveillance data. This success is attributed to the strong performance of the CT-DETR model, which is based on the Transformer architecture, providing a robust detection foundation for our multi-object tracking system. CT-DETR effectively processes every frame of the video, accurately locating pedestrian targets even in highly crowded or dynamically changing scenes. This high efficiency and precision in target detection are key to achieving accurate tracking and count statistics.

Further, Figure 8 demonstrates the model's capability by visualizing the tracking trajectories, highlighting the paths taken by individual pedestrians within the monitored area and showcasing the model's ability to continuously track each person even in complex scenarios.

5 Conclusion

In this study, we introduce a deep learning-based Multi-Object Tracking (MOT) model aimed at enhancing pedestrian tracking accuracy and efficiency. By combining CT-DETR detection technology with a ReID (Re-Identification) module, our model accurately identifies and tracks multiple pedestrian targets in various scenarios. Tested on public datasets, our model demonstrates efficient detection and tracking, particularly in urban settings and crowded environments. However, challenges remain, such as improving resilience in extreme conditions and optimizing real-time processing capabilities. Additionally, our model's tracking performance declines in low-light conditions and highly congested areas. Moving forward, our objective is to address these limitations by integrating advanced image processing technologies and optimizing algorithms to enhance resilience and reduce computational demands. We believe this research contributes to advancing multi-object tracking technology and provides insights for the practical deployment of video surveillance and intelligent systems.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., & Kim, T. K. (2021). Multiple object tracking: A literature review. *Artificial intelligence, 293*, 103448. [CrossRef]

[2] Ciaparrone, G., Sánchez, F. L., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2020). Deep learning in video multi-object tracking: A survey. *Neurocomputing, 381*, 61-88. [CrossRef]

[3] Yin, J., Wang, W., Meng, Q., Yang, R., & Shen, J. (2020). A unified object motion and affinity model for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6768-6777).

[4] Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016, October). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision* (pp. 17-35). Cham: Springer International Publishing. [CrossRef]

[5] Benfold, B., & Reid, I. (2011, June). Stable multi-target tracking in real-time surveillance video. In *CVPR 2011* (pp. 3457-3464). IEEE. [CrossRef]

[6] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence, 44*(6), 2872-2893. [CrossRef]

[7] Wang, Z., Zheng, L., Liu, Y., Li, Y., & Wang, S. (2020, August). Towards real-time multi-object tracking. In *European conference on computer vision* (pp. 107-122). Cham: Springer International Publishing. [CrossRef]

[8] Liu, M., Wang, F., Wang, X., Wang, Y., & Roy-Chowdhury, A. K. (2024). A two-stage noise-tolerant paradigm for label corrupted person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(7), 4944-4956. [CrossRef]

[9] Ning, E., Wang, C., Zhang, H., Ning, X., & Tiwari, P. (2023). Occluded person re-identification with deep learning: a survey and perspectives. *Expert Systems with Applications*, 122419. [CrossRef]

[10] Wojke, N., Bewley, A., & Paulus, D. (2017, September). Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)* (pp. 3645-3649). IEEE. [CrossRef]

[11] Wu, D., Ye, M., Lin, G., Gao, X., & Shen, J. (2021). Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE transactions on information forensics and security, 17*, 115-126. [CrossRef]

[12] Cui, Z., Zhou, J., Peng, Y., Zhang, S., & Wang, Y. (2023). Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE transactions on circuits and systems for video technology, 33*(8), 4415-4428. [CrossRef]

[13] Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision, 129*(11), 3069-3087. [CrossRef]

[14] Stadler, D., & Beyerer, J. (2021). Improving multiple pedestrian tracking by track management and occlusion handling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10958-10967).

[15] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., ... & Wang, X. (2022, October). Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision* (pp. 1-21). Cham: Springer Nature Switzerland. [CrossRef]

[16] Sun, Z., Chen, J., Chao, L., Ruan, W., & Mukherjee, M. (2020). A survey of multiple pedestrian tracking based on tracking-by-detection framework. *IEEE Transactions on Circuits and Systems for Video Technology, 31*(5), 1819-1833. [CrossRef]

[17] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ... & Chen, J. (2023). DETRs Beat YOLOs on Real-time Object Detection. *arXiv preprint arXiv:2304.08069*. [CrossRef]

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[19] Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., & Gao, J. (2021). Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems, 34*, 30008-30022.

[20] Mayer, C., Danelljan, M., Paudel, D. P., & Van Gool, L. (2021). Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13444-13454). [CrossRef]

[21] Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., & Yu, F. (2021). Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 164-173).

[22] Zheng, L., Tang, M., Chen, Y., Zhu, G., Wang, J., & Lu, H. (2021). Improving multiple object tracking with single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2453-2462).

[23] Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., &

Fu, C. (2023). Towards real-world visual tracking with temporal contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(12), 15834-15849. [CrossRef]

[24] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR), 54*(10s), 1-41. [CrossRef]

[25] Jin, X., Lan, C., Zeng, W., Wei, G., & Chen, Z. (2020, April). Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 11173-11180). [CrossRef]

[26] Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016, September). Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)* (pp. 3464-3468). Ieee. [CrossRef]

[27] Azhar, M. I. H., Zaman, F. H. K., Tahir, N. M., & Hashim, H. (2020, August). People tracking system using DeepSORT. In *2020 10th IEEE international conference on control system, computing and engineering (ICCSCE)* (pp. 137-141). IEEE. [CrossRef]

[28] Fan, L., Wang, Z., Cail, B., Tao, C., Zhang, Z., Wang, Y., ... & Zhang, F. (2016, August). A survey on multiple object tracking algorithm. In *2016 IEEE international conference on information and automation (ICIA)* (pp. 1855-1862). IEEE. [CrossRef]

[29] Tan, L., Dong, X., Ma, Y., & Yu, C. (2018, October). A multiple object tracking algorithm based on YOLO detection. In *2018 11th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)* (pp. 1-5). IEEE. [CrossRef]

[30] Kshirsagar, V., Bhalerao, R. H., & Chaturvedi, M. (2023). Modified yolo module for efficient object tracking in a video. *IEEE Latin America Transactions, 21*(3), 389-398. [CrossRef]

[31] Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*. [CrossRef]

[32] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Cham: Springer International Publishing. [CrossRef]

[33] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*. [CrossRef]

[34] Leal-Taixé, L., Milan, A., Reid, I., Roth, S., & Schindler, K. (2015). Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*. [CrossRef]

[35] Bergmann, P., Meinhardt, T., & Leal-Taixe, L. (2019). Tracking without bells and whistles. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 941-951).

[36] Wu, Y., Lim, J., & Yang, M. H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1834-1848. [CrossRef]

[37] Zhao, F., Hui, K., Wang, T., Zhang, Z., & Chen, Y. (2021). A KCF-based incremental target tracking method with constant update speed. *IEEE Access, 9*, 73544-73560. [CrossRef]

[38] Jia, X., Lu, H., & Yang, M. H. (2012, June). Visual tracking via adaptive structural local sparse appearance model. In *2012 IEEE Conference on computer vision and pattern recognition* (pp. 1822-1829). IEEE. [CrossRef]

[39] Yang, X., Zhu, S., Xia, S., & Zhou, D. (2020). A new TLD target tracking method based on improved correlation filter and adaptive scale. *The Visual Computer, 36*(9), 1783-1795. [CrossRef]

[40] Duan, Y., Wu, W., Liu, L., Liu, S., Liang, P., & Zhang, Y. (2022, December). DTTrack: Target Tracking Algorithm Combining DaSiamRPN Tracker and Transformer Tracker. In *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence* (pp. 1-5). [CrossRef]

[41] Qian, K., Zhang, S. J., Ma, H. Y., & Sun, W. J. (2023). SiamIST: Infrared small target tracking based on an improved SiamRPN. *Infrared Physics & Technology, 134*, 104920. [CrossRef]



Ming Gao received the M.Ed. in 2009. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China. His research interests include neural network, computer vision, image processing, and hyperspectral image processing. (email: aarongaoming@ieee.org)



Shixin Yang received the B.Eng. degree from Beijing Technology and Business University, China. Since 2019, he has been researching on artificial intelligence.His main research interests include artificial intelligence and Protein structure prediction. (email: 2330601026@st.btbu.edu.cn)