applied sciences

MDPI

*Article*

# Multi-Pedestrian Tracking Based on KC-YOLO Detection and Identity Validity Discrimination Module

**Jingwen Li [1,2], Wei Wu [1], Dan Zhang [3], Dayong Fan [3], Jianwu Jiang [1,2,*], Yanling Lu [1,2], Ertao Gao [1,2] and Tao Yue [1,2]**

[1] College of Geomatics and Geoformation, Guilin University of Technology, Guilin 541004, China
[2] Ecological Spatiotemporal Big Data Perception Service Laboratory, Guilin 541004, China
[3] Guilin Agricultural Science Research Center, Guilin 541004, China
* Correspondence: fengbuxi@glut.edu.cn

**Abstract:** Multiple-object tracking (MOT) is a fundamental task in computer vision and is widely applied across various domains. However, its algorithms remain somewhat immature in practical applications. To address the challenges presented by complex scenarios featuring instances of missed detections, false alarms, and frequent target switching leading to tracking failures, we propose an approach to multi-object tracking utilizing KC-YOLO detection and an identity validity discrimination module. We have constructed the KC-YOLO detection model as the detector for the tracking task, optimized the selection of detection frames, and implemented adaptive feature refinement to effectively address issues such as incomplete pedestrian features caused by occlusion. Furthermore, we have introduced an identity validity discrimination module in the data association component of the tracker. This module leverages the occlusion ratio coefficient, denoted by "k", to assess the validity of pedestrian identities in low-scoring detection frames following cascade matching. This approach not only enhances pedestrian tracking accuracy but also ensures the integrity of pedestrian identities. In experiments on the MOT16, MOT17, and MOT20 datasets, MOTA reached 75.9%, 78.5%, and 70.1%, and IDF1 reached 74.8%, 77.8%, and 72.4%. The experimental results demonstrate the superiority of the methodology. This research outcome has potential applications in security monitoring, including public safety and fire prevention, for tracking critical targets.

**Keywords:** KC-YOLO; object detection; identity validity discriminator; multi-pedestrian tracking



**Citation:** Li, J.; Wu, W.; Zhang, D.; Fan, D.; Jiang, J.; Lu, Y.; Gao, E.; Yue, T. Multi-Pedestrian Tracking Based on KC-YOLO Detection and Identity Validity Discrimination Module. *Appl. Sci.* **2023**, *13*, 12228. https://doi.org/10.3390/app132212228

Academic Editors: Zahid Mehmood Jehangiri, Mohsin Shahzad and Uzair Khan

Received: 13 October 2023
Revised: 7 November 2023
Accepted: 8 November 2023
Published: 10 November 2023



## 1. Introduction

Multi-pedestrian tracking (MPT) serves as a foundational task within the realm of computer vision and finds applications in numerous computer vision domains [1]. MPT involves estimating the trajectories of multiple objects of interest within video sequences, holding pivotal significance in video analytics systems for domains like surveillance security [2], automated driving, intelligent transportation [3], behavioral recognition [4], human–computer interaction, and intelligent agriculture [5,6]. While extensive research has been conducted in this field, a definitive method that can consistently perform exceptionally well in addressing the challenges posed by complex scenes with frequent occlusions in surveillance videos remains elusive [7]. The current focus for enhancing the accuracy of multi-pedestrian tracking primarily involves optimizing pedestrian detector performance, refining the extraction of representative pedestrian features, and improving data association matching algorithms [8].

For the optimization of pedestrian detector performance, Zhang [9], in his study, introduced a small-target pedestrian inspection model incorporating residual networks and feature pyramids, which dispenses with unnecessary, redundant computations in the model and solves the gradient problem in a neural network by using residual blocks with a discarded layer instead of the standard residual block, thus significantly improving the

accuracy and anti-jamming ability of small-target pedestrian detection. Liu [10] introduced an enhanced detection-and-tracking framework with a semantic matching strategy based on deep learning. Integrating scene-aware affinity detection, this framework proves to be highly effective in alleviating challenges related to occlusion and similar appearances. Zhang [11] introduced an innovative approach, FairMOT, which combines CenterNet and directly embeds the Re-ID module, whose training process utilizes the cross-entropy loss function, aiding in obtaining more accurate target features. This amalgamation achieves higher precision in capturing target features, all while considering the trade-off between speed and accuracy in the multi-target tracking model. Zhang et al. [12] proposed a multi-pedestrian tracking algorithm using the Tracking-by-Detection framework. It addresses the diversity of human postures, appearance similarities, and occlusion in real-time road traffic scenes. The algorithm effectively leverages both pedestrian depth appearance features and motion features to establish correlations among the tracking targets, thus realizing the multi-objective target tracking of pedestrians. Zhou et al. [13] proposed an improved MOT approach for occlusion scenarios, combining attention mechanisms and occlusion sensing as a solution. Jia [14] designed and developed a network for learning separate representations for processing occlusion re-identification guided by semantic preference object queries in a converter without strict character image alignment or any additional supervision. To better eliminate occlusion interference, they devised a Contrast Feature Learning approach to better separate hidden features from recognition features. Bewley et al. [15] proposed the Simple Online Real-Time Tracking (SORT) method, which fuses positional and motion information in a similarity matrix for target ID association and achieved good results in short-range matching. Bewley et al. [16] proposed DeepSORT based on SORT, which adds an offline pedestrian re-identification network and achieves better results in long-distance matching by merging appearance and motion information. Zhang [17] proposed the BYTE data association method, which introduces low-confidence detection frames into data association matching and utilizes these low-confidence similarities between the detection frames and the tracking trajectories to mine out heavily occluded targets, thus maintaining the continuity of the tracking trajectories.

While significant progress has been made in enhancing detector performance, extracting more representative features, and improving data association and matching algorithms, most tracking tasks still face common challenges in complex scenarios, such as occlusion, omissions, and distractions [18]. As a result, the robustness of existing methods is in need of improvement [19].

Based on the above problems, in order to solve complex surveillance video scenes with multiple targets tracked simultaneously, we propose a method for the simultaneous tracking of multiple pedestrians based on KC-YOLO detection and an identity validity discrimination module (IVDM). We have made improvements in both the detector and the tracker. The Convolution Block Attention Module (CBAM) [20] is introduced into the detector, utilizing attention weights to allow for a more focused and refined representation of the target features, which improves the ability of the detector to capture the effective feature information of the target, which has been decisive in improving the overall precision and accuracy of the test procedure. In the tracking process, to address the issue of tracking failure due to the short-term occlusion of the target, this method constructs an IVDM after cascade matching. The target occlusion coefficient $k$ is calculated to discriminate whether the target identity in the low-scoring detection frame after target detection and cascade matching is valid or not and to decide whether to update its appearance features so as not to generate redundant identity data, thereby ensuring the purity of the tracked pedestrian's identity and improving the overall performance of the tracking task.

To summarize, this paper's primary contributions can be outlined as follows:

- An efficient, robust, and practical multi-pedestrian tracking method based on KC-YOLO deep detection and identity validity discrimination is proposed. This method provides an effective solution for multi-pedestrian tracking tasks in complex surveillance videos. Experimental results demonstrate its high utility, making it suitable for

the long-term tracking of critical targets in various scenarios, such as public safety and firefighting.

- An improved pedestrian object detector based on YOLOv5, tailored for complex environments, has been designed. This detector employs the K-means++ clustering method to select optimal detection frames and introduces the CBAM for adaptive feature refinement. The KC-YOLO network is introduced for extracting target depth features.
- A pedestrian identity validity model has been developed. To address challenges such as targets reappearing after occlusion and rapid identity switches, this model assesses the identity validity of newly generated targets. Different processing strategies are applied to targets with identity validity, enhancing the tracking accuracy while ensuring the purity of pedestrian target identities.

This paper is structured such that Section 2 introduces the summary of the work related to the proposed method in this study. The multi-Pedestrian Tracking Method Based on IVDM is discussed in Section 3. The experimental data and analyses the experimental results are highlighted in Section 4, and in Section 5 we summarize this study. Lastly, we discuss this study and provide an outlook for future research in Section 6.

## 2. Related Work

### 2.1. Target Detection Methods

Target detection serves as the foundational component in the domain of multi-target tracking. The role of the detector is to furnish the tracker with the positional information of objects within the image, typically yielding the detection frame of the object. Presently, target detection algorithms achieving high accuracy are frequently implemented on the bedrock of Deep Convolutional Neural Networks (CNNs) [21]. Unlike traditional methods, deep-learning-based object recognition utilizes CNNs to autonomously capture recognizable object features. This automatic extraction process allows the model to learn complex patterns and representations from the input data, thereby improving the recognition accuracy and efficiency. In addition, hierarchical learning using CNNs allows the model to recognize features at different levels of abstraction, resulting in more justifiable and precise feature extraction. It has diverged based on detection principles, segregating into two types of methodologies [7].

Two-stage target detection involves generating candidate regions and subjecting them to a two-fold classification process. Region proposals utilizing a CNN (R-CNN) [22] input fixed-size images into a neural network to facilitate training and object feature extraction. While it attains higher detection precision compared to traditional object-detection methods, it does suffer from computational intensity and tardiness in object detection. Extending from the R-CNN algorithm, Fast R-CNN and Faster R-CNN emerge. Although these methods enhance detection accuracy compared to traditional approaches, the bifurcation between candidate region generation and classification engenders sluggish algorithmic operation, hampering real-time target detection realization. Efforts to enhance real-time capabilities still grapple with the challenge of duplicated computation. Additionally, R-CNN is hamstrung by its fixed input image size. To mitigate the pre-input image-scaling computational burden, the Spatial Pyramid Pool Network (SPPNet) was conceived, albeit only partially reducing superfluous computations. Among the R-CNN family, Faster R-CNN currently stands out with the swiftest and closest-to-real-time detection performance. This efficiency is pivotal for applications demanding rapid and accurate object detection, yet it remains encumbered in meeting the demands of intricate target detection scenarios.

One-stage detection algorithms eschew the region proposal phase and promptly produce class probabilities and the positional coordinates of objects. Representative algorithms include the YOLO family [23], the Single-Shot Multi-Frame Detector (SSD), and RetinaNet. YOLO's framework implements the detection process by allowing the model to directly predict the bounding box and class probability of each cell, which distinguishes this method from two-stage object-detection models. By doing so, YOLO achieves a more efficient and faster differentiation and correlation process, making it particularly suitable for real-time

applications such as video analysis and object tracking. Notwithstanding its strengths, YOLO demonstrates suboptimal detection accuracy for smaller objects. SSD capitalizes on feature maps of varying dimensions for object detection, rectifying YOLO's shortcomings in smaller-object detection. The contemporary YOLO family of algorithms collectively refines the detection accuracy without compromising on high detection speed.

However, in intricate traffic environments, existing object-detection algorithms still cannot simultaneously ensure real-time performance and capture as many feature points as possible.

### 2.2. Attention Mechanisms

The attention mechanism is a mechanism that mimics human attentional processes, and it is widely used in deep learning [24]. This mechanism enables the rapid extraction of key information from the environment and allows the observer to scrutinize the details of the object. After the attention mechanism, different regions will have their own weights so that the system can focus on the important information. This mechanism was initially introduced in the sphere of computer vision and is now utilized across various domains, such as natural language processing, speech recognition, and recommendation systems. The versatility of attention mechanisms lies in their ability to enhance model performance by focusing on relevant information while reducing the computational burden associated with processing unnecessary or redundant data. Therefore, they constitute a crucial component of advanced machine-learning models in diverse domains [25]. Based on their different scopes of action, attention mechanisms have been classified into three categories.

The spatial attention mechanism originates from the rationale that certain regions within input images are extraneous to recognition or segmentation tasks. The mechanism processes only regions pertinent to the task, preserving task-relevant regions while suppressing extraneous ones. An exemplary embodiment, the Spatial Transformation Network (STN) by Google DeepMind, learns preprocessing operations from input data that align with the specific task [13].

In the detection task, the input images pass through both the spatial and channel dimensions, one after the other. The network provides a significantly more comprehensive understanding of the underlying information based on the inter-channel dependencies [19]. The prominent channel attention model, SENet, compresses the input feature map spatially while preserving its channel dimension. SENet devises channel weights, adapts them during training, and then utilizes them to amplify crucial channel information while dampening insignificant channel data. Consequently, the network's feature extraction efficiency is notably enhanced [6].

Hybrid attention mechanisms amalgamate spatial and channel methods. However, certain models inadequately address the inherent interplay between features, rendering them unable to concurrently process both spatial and channel features. In this domain, representative models include the CBAM and dual-attention networks.

### 2.3. Multi-Objective Tracking Methods

Multi-objective tracking (MOT) can be classified into a detection-based tracking framework [26] and a joint detection-and-tracking-based framework, depending on the method. The detection-based tracking framework is a common approach to tracking multiple targets; it relies on target detection as the first step in locating and identifying targets in each frame and crops the objects according to the enclosing frame to obtain all of the targets in the image. Then, it is transformed into a target association problem between neighboring frames, and a similarity matrix is constructed based on IOU, appearance, etc., and solved by methods such as the Hungarian algorithm. As the performance of target detection has improved by leaps and bounds, the field of MOT has revolved around detection-based tracking frameworks for quite some time [27]. Representative methods are SORT and DeepSORT. SORT is an algorithm for tracking objects in a video sequence in real time, which is similar to many modern tracking methods [23]. Consider the case where two targets are

occluded. The trajectories of the matched targets cannot be matched for detection, and the targets temporarily disappear. When a target that disappeared briefly reappears later, the target will regain its ID number to stop changing. To enhance the SORT algorithm, the researchers added cascade matching and state estimation to it. In recent years, several joint detection tracking approaches [11] have been introduced to jointly enhance detection and a few other components. The joint tracker provides equivalent performance with minimal computational cost. However, any inconsistencies or inaccuracies in any of the components can propagate errors, which can degrade the overall tracking performance, due to the fact that there are too many components, causing this type of method to not perform very well. Therefore, the detection-based tracking framework remains the most suitable multi-target tracking method in terms of tracking accuracy.

### *2.4. Current Issues in Multi-Pedestrian Tracking*

In today's day and age, multi-pedestrian tracking still presents many challenges. For example, the following are three of the more common challenges:

- Robustness: In complex scenarios characterized by rapidly changing lighting conditions, frequent occlusions, and dynamic blurring, the robustness of multi-pedestrian tracking algorithms tends to be compromised. To tackle this, we constructed the KC-YOLO detection model as the detector in our research. This model optimizes the selection of detection frames and implements adaptive feature refinement, thereby enhancing the robustness and accuracy of the detection algorithm.
- Long-term tracking: Tracking targets in long temporal sequences presents several challenges, as it requires addressing cross-frame target re-identification and scene updates. To tackle this, we employ cascade matching for target re-identification, which effectively reduces instances of target loss caused by occlusion and scene updates [28].
- Algorithm efficiency: In real-world applications, multi-pedestrian tracking algorithms often need to process a large volume of data in real time. Overcoming these challenges is crucial for improving multi-pedestrian tracking methods and ensuring their effectiveness in diverse and complex practical environments. In our research work, we introduced an identity validity discrimination module into the tracking algorithm. This module is designed to assess and remove erroneous data resulting from incomplete or unclear features, reducing the unnecessary data-processing workload.

## 3. Multi-Pedestrian Tracking Method Based on IVDM

In the context of multi-pedestrian tracking, detection and tracking tasks are both independent and closely related to each other [29]. We adopt the KC-YOLO detection model to detect pedestrians in complex traffic environments, where the apparent information may be incomplete and unclear. Then, we introduce the IVDM as part of the improved approach to realizing the tracking of multiple pedestrian targets. The integrated detector–tracker structure is shown in Figure 1.

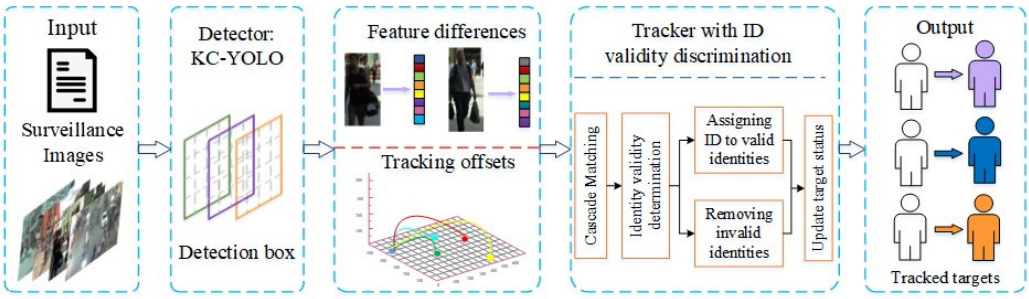


**Figure 1.** Integrated detector–tracker structure.

We utilize the KC-YOLO network as our detector to extract the adaptive deep features of the targets. We combine this with trajectory matching based on Kalman filtering

predictions. For tracking, we employ DeepSORT, which incorporates pedestrian identity validity discrimination. This combination allows us to perform accurate and efficient multi-pedestrian tracking.

### *3.1. Construction of KC-YOLO Detection Model*

We propose a model called KC-YOLO, which is applied to complex scenes in surveillance video and uses the YOLOv5 [30] detection model as the base algorithm.

The core steps of the KC-YOLO model are as follows:

- Determine the optimal anchor frame that is compatible with the input pedestrian image;
- Extract the deep features of the pedestrian image through the KC-YOLO network. Use the attention mechanism to highlight its salient information and achieve adaptive feature refinement.

We introduce the CBAM into the backbone and neck parts of the detection network for the following reasons: the backbone part is the key part for extracting pedestrian features, while the neck part fuses the features and sends them to the head for prediction, and the introduction of the CBAM here can improve the feature extraction ability of the network more effectively. The structure of the improved KC-YOLO network is shown in Figure 2.

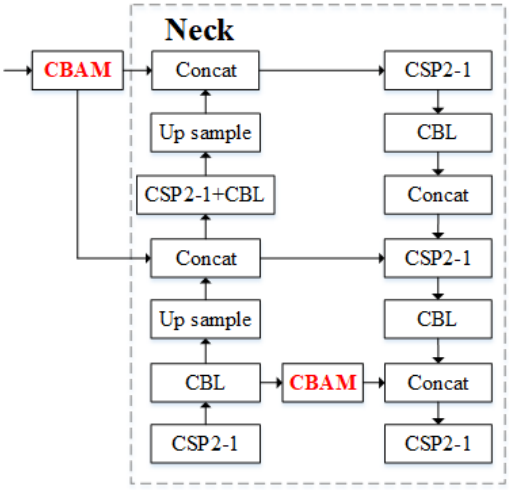


**Figure 2.** KC-YOLO network model structure (Concat is primarily responsible for combining addition and residual convolution operations; CBL is a convolutional block; Up sample means that usampling operations are performed; CSP2_1 divides the input feature map into two parts).

Concat is primarily responsible for combining addition and residual convolution operations. Through feature fusion, it allows the detection network to simultaneously utilize the extracted shallow and deep features. The main purpose of the upsample structure is to perform upsampling operations; CBL is a convolutional block. Within CSP2_1, the input feature map is divided into two parts. One part is processed through a subnetwork, while the other part undergoes further processing directly. These two sets of feature maps are then concatenated and used as input for the next layer. By combining the features processed by the subnetwork with those processed directly, a series of convolution operations are performed. This approach effectively integrates low-level detail features with high-level abstract features, thereby improving the feature extraction efficiency.

#### 3.1.1. Optimal Pedestrian Detection Frame Determination

In the context of pedestrian detection, YOLOv5 defaults to using k-means clustering to generate anchor frames. However, before performing k-means clustering, it is crucial to initialize $k$ cluster centers, as the convergence can be significantly affected by uninitialized cluster centers. To address this issue, we employ the k-means++ clustering method [31]. Here is how it works:

- Initially, a random sample point is selected from the dataset as the first initial cluster center.
- Then, the shortest distance between each sample point and the currently existing cluster centers is calculated.
- Finally, each sample point is chosen as the next cluster center with a probability proportional to the shortest distance. The sample point with the highest probability is selected as the next cluster center.

This approach provides a more reliable initialization method, improving the stability and convergence of the clustering process, which, in turn, optimizes the selection of detection frames. The formula for the calculation is as follows:

$$P(\mathrm{x}) = \frac{D(X_i)^1}{\sum_{i=1}^{n} D(X_i)^1} \tag{1}$$

where $C_i$ represents the first initial cluster center; $D(X)$ denotes the shortest distance between each sample point and the currently existing cluster centers; and $P(X)$ represents the probability of each sample point being selected as the next cluster center.

#### 3.1.2. Deep Feature Extraction

Deep features extracted by convolutional neural networks can provide an effective description of the high-level semantic information of an image, and the CBAM is an attention mechanism module used to enhance the performance of convolutional neural networks with significant results. In order to improve the feature extraction capability of the detection network [32], we introduce the CBAM [20] into the detection model.

The CBAM depicted in Figure 3 comprises both the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). The CAM is employed to enhance the weights of important features while reducing the weights of irrelevant features. It begins by subjecting the input feature map to max pooling and average pooling along the channel dimension. The output results are then fused through an MLP network, and subsequently, weight coefficients are obtained by applying the Sigmoid activation function.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1(W_0(F^c_{\max}))\right) \tag{2}$$

where $\sigma$ is the Sigmoid activation function; $w_0 \in R^{C_r \times C}$ and $w_1 \in R^{C \times C_r}$ are the weights, and $r$ is the contraction rate; $MLP$ stands for a neural network; $M_c(F)$ is obtained by performing element-wise summation and applying the Sigmoid activation operation on the shared fully connected layer; and $F^c_{avg}$ and $F^c_{\max}$ are the two features obtained by pooling the extracted features.

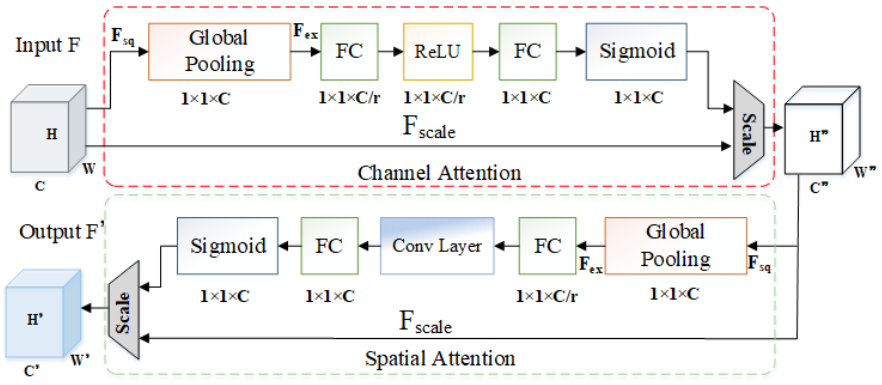


**Figure 3.** The structure of CBAM (Global Pooling is the global maximum pooling layer; FC is Rectified Linear Unit; Sigmoid is an S-type activation function).

The SAM focuses on the intrinsic relationships within the spatial dimensions of the input feature map. It takes the output from the CAM and performs max pooling and average pooling along the channel direction. The results obtained are then processed through a convolutional layer with a kernel size of $7 \times 7$. Finally, the SAM's feature map is obtained by applying the Sigmoid activation function. The calculation is performed with the following equation:

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F); MaxPool(F)])\right) = \sigma\left(f^{7\times7}\left(\left[F^c_{avg}; F^c_{\max}\right]\right)\right) \quad (3)$$

where $f^{7\times7}$ denotes the convolution kernel size; $M_s(F)$ is obtained by the logistic activation function.

### *3.2. Multi-Pedestrian Tracking Methods Based on IVDM*

Pedestrian tracking not only provides trajectory information but also provides valuable information for behavioral analysis. However, in crowded scenes, a large number of targets may be occluded, resulting in missing and blurred features, which seriously affects the function of detection-based tracking methods [16]. When the video surveillance fields of view do not overlap and the pedestrians are heavily occluded, the "1-n" pedestrian identity phenomenon results. Existing tracking algorithms still lack a flexible approach to dealing with heavily occluded targets and thus perform poorly in complex scenarios where heavy occlusion occurs frequently [33]. To address the above situation, based on the improved detection method in the previous section, we introduce pedestrian identity validity judgment into the pedestrian tracking process, which performs "occlusion perception-occlusion ratio *k* calculation-pedestrian identity validity discrimination" on unmatched targets between different frames (Figure 4).

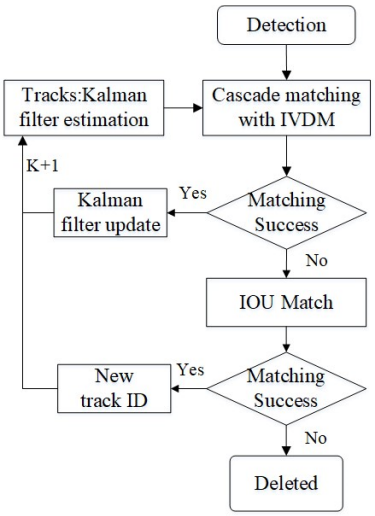


**Figure 4.** Multi-pedestrian tracking method based on IVDM.

The module performs a series of calculation and discrimination operations on unmatched targets between different frames, such as "occlusion perception-calculation of occlusion ratio *k*-pedestrian identity validity discrimination", which determines the degree of occlusion of pedestrians detected by the surveillance video based on the magnitude of the coefficient k of the proportion of occlusion of pedestrians in the frame and categorizes the occluded pedestrians into valid $ID_Y$ and invalid $ID_N$ through *k*. In essence, the above process is used to discern whether or not the identity of the detected pedestrian has validity. The most successful associations in pedestrian tracking often occur in the cascade matching section. Therefore, we have incorporated the IVDM into the cascade matching process, as depicted in Figure 5.

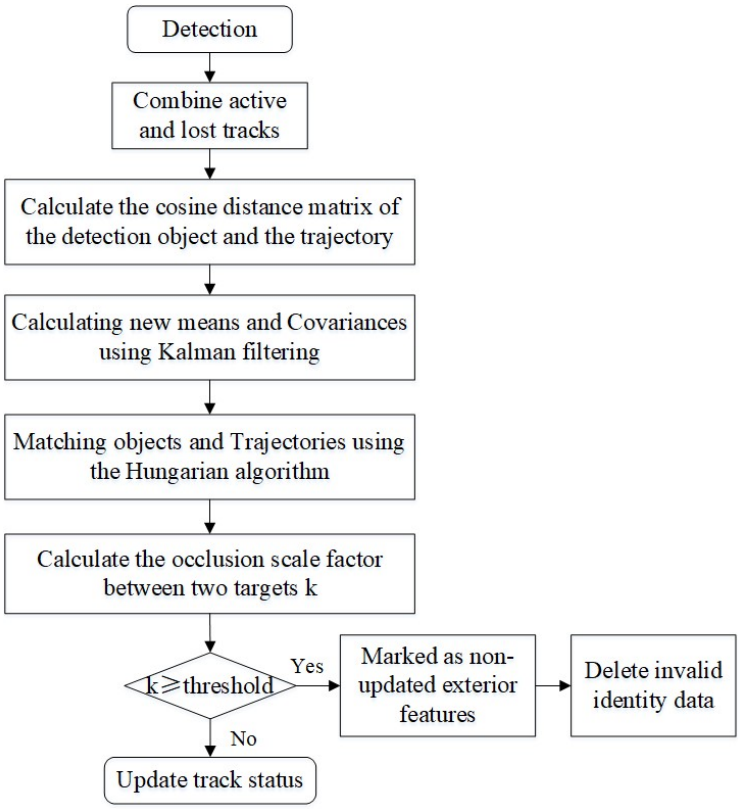


**Figure 5.** Cascade matching with IVDM.

#### 3.2.1. Occlusion-Aware Detection

For occlusion-aware detection, traditional Intersection Over Union (IOU) cross-ratio algorithms calculate the overlap ratio and filter targets that satisfy the requirements by setting a threshold [23].

Figure 6a,b show that the IOU algorithm is effective in discriminating pedestrians when they have similar body size ratios. However, real applications mostly involve complex scenes, and the size of the pedestrian detection frame produces a very large error due to the different distances of the camera from the ground. The IOU algorithm has very little utility in this case (Figure 6c), which is why it cannot be used as a calculation standard to show the occlusion of pedestrians and small targets in real applications [13]. Therefore, we propose the identity validity discriminant coefficient $k$, which calculates the ratio of the extent of the occluded portion of an occluded pedestrian to its detection frame and can more accurately discriminate the degree of the pedestrian's occlusion.


**Figure 6.** Target occlusion diagram: (**a**) similarly proportioned pedestrian non-influential screening; (**b**) similarly proportioned pedestrian-impacted screening; (**c**) disparately proportioned pedestrian-impacted screening.

3.2.2. Determination of the Shading Scale Factor $k$

In order to express the derivation of the occlusion ratio coefficient more intuitively, we define the coordinates of the detection frame. As shown in Figure 7, $(x_{a_1}, y_{a_1})$ denotes the coordinates of the upper-left corner of the blocked pedestrian detection frame; $(x_{a_2}, y_{a_2})$ denotes the coordinates of the upper-right corner of the blocked pedestrian detection frame; $(x_{b_1}, y_{b_1})$ denotes the coordinates of the upper-left corner of the blocked pedestrian detection frame; $(x_{b_2}, y_{b_2})$ denotes the coordinates of the upper-right corner of the blocked pedestrian detection frame; $(x_1, y_1)$ is the upper-left corner of the blocking section; and $(x_2, y_2)$ is the upper-right corner of the blocking section, calculated with the following equation:

$$\begin{cases} x_1 = \max(x_{a_1}, x_{b_1}), y_1 = \max(y_{a_1}, y_{b_1}) \\ x_2 = \max(x_{a_2}, x_{b_2}), y_2 = \max(y_{a_2}, y_{b_2}) \\ S = (x_{a_2} - x_{a_1} + 1.0).(y_{a_2} - y_{a_1} + 1.0) \\ S_0 = \max(x_2 - x_1 + 1.0).\max(y_2 - y_1 + 1.0) \end{cases} \tag{4}$$

$S$ denotes the range of the occluded target frame; $S_0$ denotes the range of the occluded region.

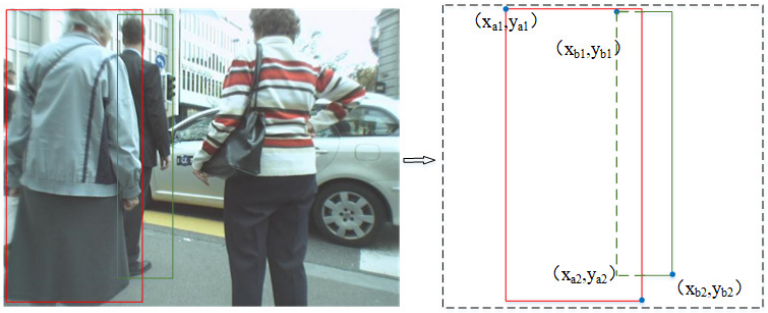


**Figure 7.** Occlusion frame coordinate plot.

Based on the obtained data for each attribute of the pedestrian detection frame, the unmatched target occlusion ratio coefficient $k$ after cascade matching is derived from the ratio of the occluded area range. $S_0$ and the occluded target frame range $S$ are calculated with the following equation:

$$k = \frac{S_0}{S} \tag{5}$$

3.2.3. Identity Validity Determination Module

When two pedestrians form an occlusion, in general, if the center of mass of one target is detected within the detection frame coordinates of the other target, the identity of the pedestrian is determined to be invalid due to the effect of the occlusion, and then the identity validity = 0; otherwise, identity validity = 1. However, when occlusion is generated, the above method will not be able to accurately determine the degree of occlusion of the pedestrian if the center of mass of the pedestrian is not within the coordinates of the other pedestrian detection frames, and then the specific degree of occlusion of the pedestrian needs to be calculated. The degree of occlusion of the pedestrian is determined if the target occlusion ratio coefficient k is greater than a threshold value, and then identity validity = 1; otherwise, identity validity = 0.

In this study, the original model was tested on consecutive frames from the MOT16, MOT17, and MOT20 datasets. Based on the experimental responses in Figure 8, when $k > 0.535$, the proportion of tracking failures due to occlusion increases significantly.

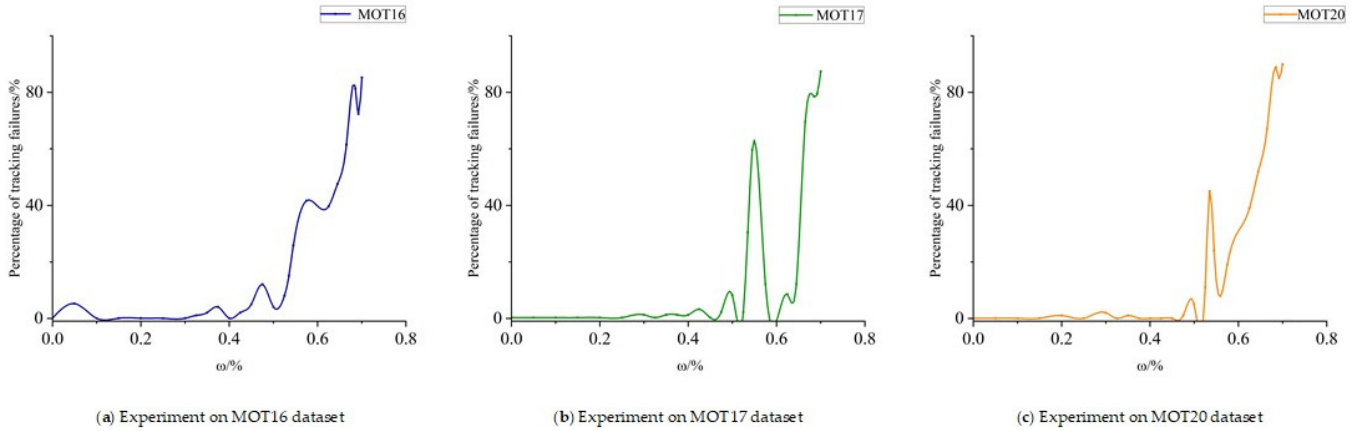


**Figure 8.** Percentage of tracking failures for different levels of occlusion: (**a**) Experiment on MOT16 dataset; (**b**) Experiment on MOT17 dataset; (**c**) Experiment on MOT20 dataset.

The identity validity is binarized by the occlusion ratio coefficient $k$ as the identity validity score of the corresponding pedestrian, where 1 indicates that the target identity is invalid and 0 indicates that the pedestrian identity is valid, and the relationship of the identity validity score calculation is calculated with the following equation:

$$e_i = \begin{cases} 1, & k \geq \omega \\ 0, & else \end{cases} \tag{6}$$

## 4. Experiments and Analyses

Here, we statistically summarize the results of the experiments and analyze them in depth, leading to well-reasoned conclusions.

### *4.1. Experimental Environment*

In this study, we used Pytorch [34] for code writing, and we conducted the experiments on a server configured with Intel(R) Xeon(R) CPU E5-2680 V4 @ 2.40GHz (Intel, made in Malaysia) and NVIDIA GeForce RTX 3090 GPUs (Msi, made in China).

### *4.2. Experimental Dataset and Evaluation Index*

#### 4.2.1. Experimental Data

We opted for the MOT series datasets, CrowdHuman dataset, and MIX datasets, commonly utilized in pedestrian tracking tasks, to conduct our experiments. This choice enhances the credibility of our proposed method's effectiveness. Below is an introduction to the three datasets:

- MOT series datasets: These are datasets on the Open Data Lab platform and are mainly targeted at pedestrian tracking tasks in dense scenes.
- CrowdHuman dataset [35]: It is for pedestrian detection. Unlike other mainstream human detection datasets, the pedestrian targets in the CrowdHuman dataset are much denser, more crowded, and even have serious overlaps. According to the data provided in the citation, the CrowdHuman dataset has an average of 22.64 figures per image, which is far more than other human detection datasets.
- MIX datasets: They are diverse and comprehensive, covering different types of pedestrian detection and tracking scenarios. This comprehensiveness allows researchers to test the robustness and effectiveness of algorithms in a variety of real-world situations. Using these datasets, researchers can conduct multimodal data studies, explore commonalities and differences between different datasets, and lay the foundation for improving multi-target tracking and pedestrian detection algorithms.

4.2.2. Evaluation Metrics

- Pedestrian Detection Evaluation Metrics: These are quantitative measures utilized to assess the performance and accuracy of algorithms and models designed to detect pedestrians in images or videos. They offer in-depth insights into a system's ability to recognize pedestrians within a given dataset. Common evaluation metrics for pedestrian detection include precision, recall, and *mAP*. Precision is the ratio of true positives to the total number of predicted positives, where true positives are the instances where the prediction is correct. Recall calculates the ratio of instances where the prediction is correct to the total number of actual positives. *mAP* is the sum of the average precision values for all classes divided by the number of classes. In other words, it represents the average of the average precisions for all classes in the dataset.

  The mathematical expressions for the above evaluation indicators are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Re}call = \frac{TP}{TP + FN} \tag{8}$$

$$mAP = \frac{1}{R}\int_0^1 P(R)dR \tag{9}$$

where *TP* represents the true positives, *FP* represents the false positives, and *FN* represents the false negatives.

- Pedestrian Tracking Evaluation Metrics: In order to test our proposed multi-target pedestrian tracking method, we use five criteria as evaluation metrics: the multi-objective tracking accuracy (*MOTA*) [36], which is commonly expressed as a percentage, ranging from 0% to 100%, where a higher score indicates the superior performance of the tracking algorithm; the ratio of the average of the number of correctly recognized ground-truth detections to the number of computed detections ($IDF_1$) [37]; the Majority of Tracked (*MT*); Major Lost Targets (*ML*); and Identity Switches *(IDS*).

  The mathematical expressions are as follows:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDS_{w_t})}{\sum_t GT_t} \times 100\% \tag{10}$$

$$IDF_1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \times 100\% \tag{11}$$

where $t$ represents the index of each frame of the video, and $GT$ is the number of real labeled targets in the image. $IDS_w$ denotes the total number of *ID* switches occurring in the tracked target in frame $t$.

*4.3. Analysis of Experimental Results*

4.3.1. Comparison Experiments

We conducted the following comparative experiments on different multi-pedestrian tracking algorithms using the MOT16, MOT17, and MOT20 datasets. Table 1 shows some common algorithms for the target tracking task and the experimental results of our introduced algorithm.

**Table 1.** Experimental results on the MOT16 dataset.

| Method | MOTA/% | IDF1/% | MT/% | ML/% | IDs |
|---|---|---|---|---|---|
| SORT [15] | 59.8 | 53.8 | 25.4 | 22.7 | 1423 |
| JDE [38] | 64.4 | 55.8 | 35.4 | 20.0 | 1544 |
| CNNMTT [39] | 65.2 | 62.2 | 32.4 | 21.3 | 946 |
| CTrackV1 [40] | 67.6 | 57.2 | 32.9 | 23.1 | 5529 |
| FairMOT [11] | 73.7 | 72.4 | 44.7 | 15.9 | 1074 |
| DeepSORT [16] | 74.8 | 73.6 | 45.2 | 15.4 | 1022 |
| Our Method | 75.9 | 74.8 | 42.5 | 18.3 | 816 |

Table 1 shows that our multi-pedestrian tracking method has an absolute advantage over many methods. In terms of evaluation metrics, *MOTA* is improved by 1.1%, and *IDF1* is improved by 1.2%, with enhanced robustness compared to the original tracking method. It is worth noting that the IDs of our method are significantly lower than those of DeepSORT; presumably, improved models may improve the predictive power of the tracking method compared to the original model, making the tracking results more accurate and significantly improving the problem of ID hopping. The smaller number of *IDs* makes the tracking results of the model more practical in real applications. After a series of comparisons, it leads to the conclusion that our tracking algorithm has significant advantages in all aspects of performance.

In Table 2, we can see that our multi-pedestrian tracking method significantly improves the experimental metrics on the MOT17 dataset, with improvements of 2.1% and 4.4% for *MOTA* and *IDF1*. We believe that the proposed IDVM is better at presenting false and erroneous identity data and thus can be applied to complex surveillance video scenarios with frequent occlusions.

**Table 2.** Experimental results on the MOT17 dataset.

| Method | MOTA/% | IDF1/% | MT/% | ML/% | IDs |
|---|---|---|---|---|---|
| SST [41] | 52.4 | 49.5 | 21.4 | 30.7 | 8431 |
| TubeTK [42] | 63.0 | 68.6 | 31.2 | 24.2 | 4137 |
| CenterTrack [33] | 67.8 | 64.7 | 34.6 | 24.6 | 2583 |
| FairMOT [11] | 73.1 | 72.7 | 41.1 | 19.0 | 2964 |
| TransMOT [38] | 75.1 | 74.6 | 40.8 | 22.6 | 2340 |
| ByteTack [17] | 77.4 | 76.1 | 39.9 | 20.2 | 2236 |
| DeepSORT [16] | 76.4 | 73.4 | 39.1 | 21.0 | 1898 |
| Our Method | 78.5 | 77.8 | 38.6 | 19.9 | 1586 |

In Table 3, we compare the original DeepSORT method with the multi-pedestrian tracking method that we propose based on DeepSORT on the MOT20 dataset for comparative tests. The *MOTA* and *IDF1* indices of our proposed method are improved by 3.3% and 4.5%, respectively. Therefore, the improved method significantly improves the power to extract the apparent features of the pedestrian target, which leads to more accurate feature extraction and makes the overall performance of this tracking method significantly better. In addition to this, the other two metrics are also significantly improved, which illustrates that the robustness of the tracker to act on the same target during the pedestrian tracking process has been improved. Meanwhile, the pedestrian IVDM not only reduces the appearance feature contamination problem but also improves the tracking robustness for whether or not to update the appearance features after discrimination.

**Table 3.** Experimental results on the MOT20 dataset.

| Method | MOTA/% | IDF1/% | MT/% | ML/% | IDs |
|---|---|---|---|---|---|
| DeepSORT [16] | 66.8 | 67.9 | 68.7 | 8.4 | 2269 |
| Our Method | 70.1 | 72.4 | 69.2 | 8.7 | 1689 |

4.3.2. Pedestrian Detection Algorithm Ablation Experiments

In order to validate our proposed improvement strategy for YOLOv5, ablation experiments were carried out on datasets such as CrowdHuman to judge the effect of each enhancement point. The results of effectiveness experiments for each component of YOLOv5 are as follows.

As shown in Table 4, the KC-YOLO model's accuracy value is improved by 6%, and AP is improved by 4%. The improved model greatly improves the ability to extract intra-pedestrian detection deformations and appearance features, thus capturing more accurate features.

**Table 4.** YOLOv5 ablation experiment.

| k-Means++ | CBAM | Precision | Recall | AP |
|---|---|---|---|---|
| × | × | 0.85 | 0.78 | 0.85 |
| × | √ | 0.89 | 0.83 | 0.88 |
| √ | × | 0.88 | 0.84 | 0.85 |
| √ | √ | 0.91 | 0.85 | 0.89 |

In Figure 9, YOLOv5+K-means+++CBAM is the KC-YOLO model proposed in this study. At the beginning of training, the values of *AP* and accuracy reach more than 0.8, which is mainly due to the pre-trained model when training YOLOv5. After using the K-means++ clustering method, both *AP* and accuracy are inevitably improved compared to YOLOv5, while the *Loss* value has a small decrease and gradually converges, which indicates that our improvements to the model are positively oriented and the effects are evident. After embedding the CBAM, with increasing epochs, *AP* and accuracy increase significantly, and the *Loss* becomes smaller and converges gradually, which indicates that the improved model is more desirable.

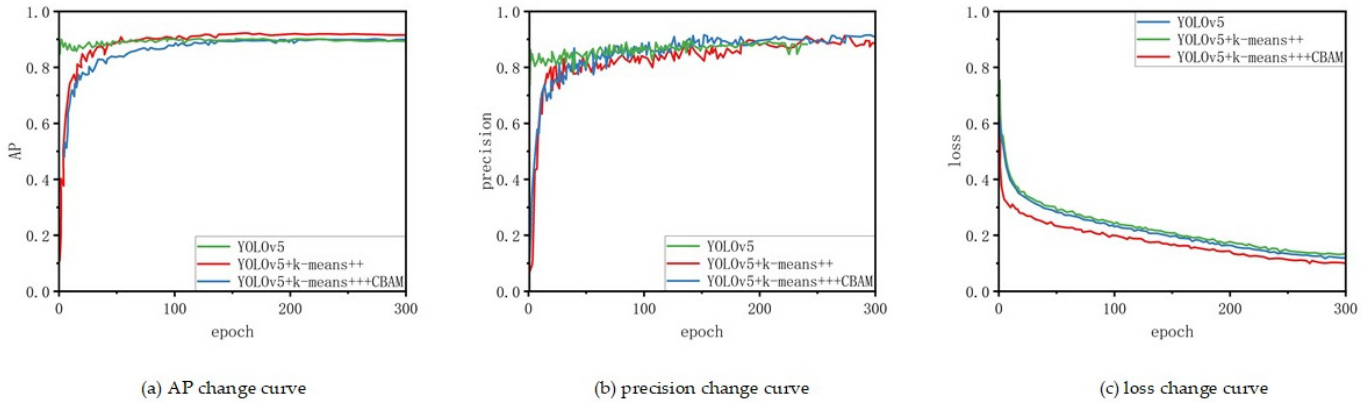


**Figure 9.** Pedestrian ablation test results: (**a**) AP change curve; (**b**) precision change curve; (**c**) loss change curve.

4.3.3. Pedestrian Tracking Algorithm Ablation Experiments

To check whether the IVDM in our proposed multi-pedestrian tracking method is a positive improvement, we conducted ablation experiments on the IVDM.

The results of ablation experiments with the IVDM are shown in Table 5. They clearly show that the addition of the IVDM improves the *MOTA* of the multi-pedestrian tracking method by 3.0% and the *IDF1* by 2.8%, and the *IDs* are also significantly reduced. The role of the IVDM is mainly to eliminate false and erroneous identity data due to occlusion and to maintain the purity of the original tracking target while reducing the generation of redundant data. After the introduction of the IVDM, the overall performance of the task is improved.

**Table 5.** Experimental ablation study of IVDM on the MOT20 dataset.

| Method | MOTA/% | IDF1/% | MT/% | ML/% | IDs |
|---|---|---|---|---|---|
| Without IVDM | 67.1 | 69.6 | 67.6 | 9.1 | 2638 |
| With IVDM | 70.1 | 72.4 | 69.2 | 8.7 | 1689 |

To illustrate more intuitively the advantages of the improvements made to the tracking method in this study, we present visual comparisons of the tracking experiments conducted on the MOT16 training dataset. Figure 10 compares the demonstrations of the multi-pedestrian tracking method with and without the IVDM.

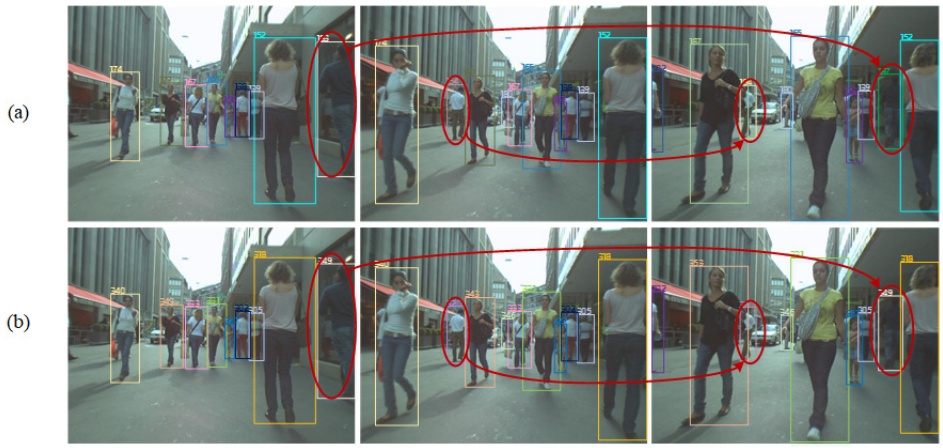


**Figure 10.** Visualization of multi-pedestrian tracking results: (**a**) without IVDM; (**b**) with IVDM.

As shown in Figure 10, when using the multi-pedestrian tracking without adding the IVDM, the target pedestrians with ID number 154 and ID number 183 changed their ID numbers due to brief occlusion, which also means that the tracking failed and generated redundant and incorrect ID numbers at the same time. However, the problem of tracking failures due to transient occlusion is well solved after adding the IVDM.

## 5. Discussion

This study introduces a method for the multi-object tracking of pedestrians across multiple cameras in complex scenes, and it exhibits a higher tracking accuracy compared to existing methods in practical applications. However, like any research, our work has certain limitations that need to be considered. One major limitation is the potential influence of environmental factors on the accuracy of our model. For instance, the spacing between cameras could impact the accuracy of our tracking algorithm. Additionally, further research on the algorithm using different datasets can enhance its robustness and generalizability.

Furthermore, in order to enable rapid and accurate tracking of critical targets in applications such as public safety and fire protection systems, our next step will involve considering the design of a more lightweight model to reduce storage and computational requirements. These studies will contribute to expanding the applicability of our approach and assist in the development of more efficient and powerful pedestrian tracking algorithms.

## 6. Conclusions

In this research, we have developed a multi-pedestrian tracking method based on deep detection and identity validity assessment, specifically designed for complex surveillance video scenarios where issues like target occlusion are frequent.

We have constructed the KC-YOLO network as the detector, which employs the k-means++ clustering method to select the optimal target detection frames. Additionally, we have integrated a convolutional attention mechanism into the target detection algorithm, utilizing attention weights for adaptive feature refinement. This effectively suppresses

secondary features to highlight crucial target characteristics, enhancing the robustness of target detection in complex scenes, where target features may become less distinct due to occlusion. The robustness of the detector has been verified through experiments.

In the target tracker, we have introduced the IVDM, which performs occlusion-aware processing on pedestrian targets after feature extraction by the detector. In cases where target identities are compromised due to occlusion-induced errors, we use the occlusion coefficient "k" to assess the validity of the identity. Based on the output of this module, we determine whether pedestrian targets possess valid identities, influencing the decision to update the appearance features of the current dynamic target.

Here are the experimental results on the MOT16 dataset: MOTA is 75.9%, and IDF1 is 74.8%. Compared to SORT, there is a 20.1% increase in MOTA and a 21.0% increase in IDF1. In comparison to CNNMTT, MOTA has improved by 10.7%, and IDF1 has seen a 12.6% improvement. When contrasted with the prototype DeepSORT method, MOTA has increased by 1.1%, and IDF1 has increased by 1.2%. The most noteworthy aspect is the substantial reduction in IDS, maintaining a high level of tracking continuity. For the MOT17 dataset, MOTA is 78.5%, and IDF1 is 77.8%. For the MOT20 dataset, the results show a MOTA of 70.1% and an IDF1 of 72.4%. When contrasted with the prototype DeepSORT method, the MOTA and IDF1 indices of our proposed method are improved by 3.3% and 4.5%. These experiments confirm that our research outperforms several advanced MOT algorithms across nearly all metrics. This study provides a stable and efficient approach to multi-pedestrian tracking in complex scenarios, significantly reducing the number of ID switches to ensure the continuity of tracking trajectories. This approach is particularly well suited for public safety and fire protection departments, enabling the continuous tracking of critical targets in crowded scenes with severe occlusion.

**Author Contributions:** Conceptualization, W.W. and J.L.; methodology and validation, J.J., D.F. and D.Z.; formal analysis, Y.L., E.G. and T.Y.; investigation, W.W. and D.Z.; writing—original draft preparation, J.L. and Y.L.; writing—review and editing, T.Y. and E.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, grant number 41961063; Guilin Technology Application and Promotion Project, 2022, grant 20220138-2; and Guilin Key R&D Project, 2022, grant 20220109.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/xuanwang-91/Framework-for-Pedestrian-Detection-Tracking-and-Re-identifcation.git, accessed on 10 February 2023.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

| | |
|---|---|
| MPT | Multi-pedestrian tracking |
| SORT | Simple Online Real-Time Tracking |
| CBAM | Convolution Block Attention Module |
| CNNs | Deep Convolutional Neural Networks |
| R-CNN | Region proposals utilizing CNN |
| SPPNet | Spatial Pyramid Pool Network |
| SSD | Single-Shot Multi-Frame Detector |
| STN | Spatial Transformation Network |
| MOT | Multi-objective tracking |
| IOU | Intersection Over Union |
| MOTA | Multi-objective tracking accuracy |
| ML | Major Lost Targets |
| MT | Majority of Tracked |

| | |
|---|---|
| IDS | Identity Switches |
| IVDM | Identity validity discrimination module |

## References


1. Xiao, C.; Luo, Z. Improving multiple pedestrian tracking in crowded scenes with hierarchical association. *Entropy* **2023**, *25*, 380. [CrossRef] [PubMed]
2. Pouyan, S.; Charmi, M.; Azarpeyvand, A.; Hassanpoor, H. Propounding first artificial intelligence approach for predicting robbery behavior potential in an indoor security camera. *IEEE Access* **2023**, *11*, 60471–60489. [CrossRef]
3. Zhang, Q. Multi-object trajectory extraction based on YOLOv3-DeepSort for pedestrian-vehicle interaction behavior analysis at non-signalized intersections. *Multimed. Tools Appl.* **2023**, *82*, 15223–15245. [CrossRef]
4. Geng, P.; Xie, H.; Shi, H.; Chen, R.; Tong, Y. Pedestrian Fall Event Detection in Complex Scenes Based on Attention-Guided Neural Network. *Math. Probl. Eng.* **2022**, *2022*, 4110246. [CrossRef]
5. Lin, Y.; Hu, W.; Zheng, Z.; Xiong, J. Citrus Identification and Counting Algorithm Based on Improved YOLOv5s and DeepSort. *Agronomy* **2023**, *13*, 1674. [CrossRef]
6. Osman, Y.; Dennis, R.; Elgazzar, K. Yield Estimation and Visualization Solution for Precision Agriculture. *Sensors* **2021**, *21*, 6657. [CrossRef]
7. Yang, J.; Ge, H.; Yang, J.; Tong, Y.; Su, S. Online pedestrian multiple-object tracking with prediction refinement and track classification. *Neural Process. Lett.* **2022**, *54*, 4893–4919. [CrossRef]
8. Li, X.; Hu, W.; Shen, C.; Zhang, Z.; Dick, A.; Hengel, A.V.D. A survey of appearance models in visual object tracking. *ACM Trans. Intell. Syst. Technol.* **2013**, *4*, 1–48. [CrossRef]
9. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. Trackformer: Multi-object tracking with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8844–8854. [CrossRef]
10. Liu, C.J.; Lin, T.N. DET: Depth-enhanced tracker to mitigate severe occlusion and homogeneous appearance problems for indoor multiple-object tracking. *IEEE Access* **2022**, *10*, 8287–8304. [CrossRef]
11. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
12. Yi, Z.; Shen, Y.; Zhao, Q. Multi-Person tracking algorithm based on data association. *Optik* **2019**, *194*, 163124. [CrossRef]
13. Zhou, X.; Chan, S.; Qiu, C.; Jiang, X.; Tang, T. Multi-Target Tracking Based on a Combined Attention Mechanism and Occlusion Sensing in a Behavior-Analysis System. *Sensors* **2023**, *23*, 2956. [CrossRef] [PubMed]
14. Jia, M.; Cheng, X.; Lu, S.; Zhang, J. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Trans. Multimed.* **2022**, *25*, 1294–1305. [CrossRef]
15. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468. [CrossRef]
16. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649. [CrossRef]
17. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. Bytetrack: Multi-object tracking by associating every detection box. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 1–21.
18. Ning, C.; Menglu, L.; Hao, Y.; Xueping, S.; Yunhong, L. Survey of pedestrian detection with occlusion. *Complex Intell. Syst.* **2021**, *7*, 577–587. [CrossRef]
19. Wang, Z.; Li, Z.; Leng, J.; Li, M.; Bai, L. Multiple pedestrian tracking with graph attention map on urban road scene. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 8567–8579. [CrossRef]
20. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018. [CrossRef]
21. Hu, X.; Xu, X.; Xiao, Y.; Chen, H.; He, S.; Qin, J.; Heng, P.A. SINet: A scale-insensitive convolutional neural network for fast vehicle detection. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 1010–1019. [CrossRef]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
23. Yan, J.; Du, S.; Wang, Y. Multi-Pedestrian Tracking in Crowded Scenes by Modeling Movement Behavior and Optimizing Kalman Filter. *IEEE Access* **2022**, *10*, 118512–118521. [CrossRef]
24. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154. [CrossRef]
25. Zhou, Q.; Zhong, B.; Zhang, Y.; Li, J.; Fu, Y. Deep alignment network based multi-person tracking with occlusion and motion reasoning. *IEEE Trans. Multimed.* **2018**, *21*, 1183–1194. [CrossRef]
26. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. Strongsort: Make deepsort great again. *IEEE Trans. Multimedia* **2023**, 1–14. [CrossRef]

27. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
28. Li, H.; Liu, Y.; Wang, C.; Zhang, S.; Cui, X. Tracking algorithm of multiple pedestrians based on particle filters in video sequences. *Comput. Intell. Neurosci.* **2016**, *2016*, 8163878. [CrossRef] [PubMed]
29. Mykhaylo, A. People-tracking-by-detection and people-detection-by-tracking. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, USA, 23–28 June 2008. [CrossRef]
30. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788. [CrossRef]
31. Hämäläinen, J.; Kärkkäinen, T.; Rossi, T. Improving scalable K-means++. *Algorithms* **2020**, *14*, 6. [CrossRef]
32. Li, Z.H.; Chen, J.; Bi, J. Multiple object tracking with appearance feature prediction and similarity fusion. *IEEE Access* **2023**, *11*, 52492–52500. [CrossRef]
33. Chen, K.; Song, X.; Zhai, X.; Zhang, B.; Hou, B.; Wang, Y. An integrated deep learning framework for occluded pedestrian tracking. *IEEE Access* **2019**, *7*, 26060–26072. [CrossRef]
34. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, 32. [CrossRef]
35. Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; Sun, J. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv* **2018**, arXiv:1805.00123. [CrossRef]
36. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
37. Hua, G.; Jégou, H. (Eds.) *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16 2016, Proceedings, Part II*; Springer: Berlin/Heidelberg, Germany, 2016. [CrossRef]
38. Wang, Z.; Zheng, L.; Liu, Y.; Wang, S. Towards Real-Time Multi-Object Tracking. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 107–122. [CrossRef]
39. Mahmoudi, N.; Ahadi, S.M.; Rahmati, M. Multi-target tracking using CNN-based features: CNNMTT. *Multimed. Tools Appl.* **2019**, *78*, 7077–7096. [CrossRef]
40. Peng, J.; Wang, C.; Wan, F.; Wu, Y.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Fu, Y. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part IV 16. Springer International Publishing: Cham, Switzerland, 2020; pp. 145–161. [CrossRef]
41. Sun, S.; Akhtar, N.; Song, H.; Mian, A.S.; Shah, M. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 104–119. [CrossRef]
42. Pang, B.; Li, Y.; Zhang, Y.; Li, M.; Lu, C. Tubetk: Adopting tubes to track multi-object in a one-step training model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6308–6318. [CrossRef]