Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.038154
Article

Multi-Target Tracking of Person Based on Deep Learning

Xujun Li*, Guodong Fang, Liming Rao and Tengze Zhang

Physics and Optoelectronic Engineering College, Xiangtan University, Xiangtan, 411105, China
*Corresponding Author: Xujun Li. Email: xjli@xtu.edu.cn
Received: 29 November 2022; Accepted: 17 February 2023; Published: 28 July 2023

Abstract: To improve the tracking accuracy of persons in the surveillance video, we proposed an algorithm for multi-target tracking persons based on deep learning. In this paper, we used You Only Look Once v5 (YOLOv5) to obtain person targets of each frame in the video and used Simple Online and Realtime Tracking with a Deep Association Metric (DeepSORT) to do cascade matching and Intersection Over Union (IOU) matching of person targets between different frames. To solve the IDSwitch problem caused by the low feature extraction ability of the Re-Identification (ReID) network in the process of cascade matching, we introduced Spatial Relation-aware Global Attention (RGA-S) and Channel Relation-aware Global Attention (RGA-C) attention mechanisms into the network structure. The pre-training weights are loaded for Transfer Learning training on the dataset CUHK03. To enhance the discrimination performance of the network, we proposed a new loss function design method, which introduces the Hard-Negative-Mining way into the benchmark triplet loss. To improve the classification accuracy of the network, we introduced a Label-Smoothing regularization method to the cross-entropy loss. To facilitate the model's convergence stability and convergence speed at the early training stage and to prevent the model from oscillating around the global optimum due to excessive learning rate at the later stage of training, this paper proposed a learning rate regulation method combining Linear-Warmup and exponential decay. The experimental results on CUHK03 show that the mean Average Precision (mAP) of the improved ReID network is 76.5%. The Top 1 is 42.5%, the Top 5 is 65.4%, and the Top 10 is 74.3% in Cumulative Matching Characteristics (CMC); Compared with the original algorithm, the tracking accuracy of the optimized DeepSORT tracking algorithm is improved by 2.5%, the tracking precision is improved by 3.8%. The number of identity switching is reduced by 25%. The algorithm effectively alleviates the IDSwitch problem, improves the tracking accuracy of persons, and has a high practical value.

Keywords: YOLOv5; DeepSORT; deep learning; attention mechanism; person re-identification; multi-target tracking

1 Introduction

Multi-Object Tracking (MOT) uses the contextual semantic information of video or image sequences to model the appearance characteristics and motion state of the target, to predict the motion state of the target, and to calibrate the position of the target [1]. In the past decade, Tracking By Detection (TBD) has attracted more attention with the rapid development of target detection algorithms. It has become the mainstream framework in the field of multi-target tracking. The multi-target tracking method under this framework can decompose video multi-target tracking tasks into two independent sub-tasks: object detection and data association. Firstly, the trained target detector is used to detect each frame in the video. Then the detection results from different video frames belonging to the same target are associated with forming the target's trajectory [2].

The traditional object detection algorithm uses the artificial construction of target features and then uses a classification algorithm to classify and judge whether the target exists. Typical algorithms such as Haar-like Features (Haar) + Adaptive Boosting (AdaBoost), Histograms of Oriented Gradients (HOG) + Support Vector Machine (SVM), and Deformable Parts Model (DPM) require sliding window operation in the image, which has low detection efficiency, high resource consumption, and low robustness of artificially designed features. The generalization effect could be better, quickly leading to the false detection of person targets and missing detection phenomenon. With the continuous development of deep learning and Graphics Processing Unit (GPU) parallel computing technologies, object detection has gradually changed from traditional methods to methods based on deep learning, which can be divided into the Two-Stage algorithm and the One-Stage algorithm [3]. The Two-Stage algorithm first generates the candidate regions during detection, then classifies and calibrates based on the candidate regions with a relatively high accuracy, representing Regional Convolutional Neural Network (R-CNN) series models [4,5]. When detecting with the One-Stage algorithm, there is no need to generate candidate regions, and the categories and boundaries of the targets are directly regressed. The representative models include Single Shot Multibox Detector (SSD), Focal Loss for Dense Object Detection (RetinaNet), and You Only Look Once (YOLO) series [6–8]. More advanced methods have been proposed in object detection and tracking in recent years. A shadow-background-noise 3D spatial decomposition (SBN-3D-SD) model is proposed to enhance shadows for higher detection accuracies. It boosts the shadow detection accuracy of Faster R-CNN, RetinaNet, and YOLOv3 [9]. The Optical Remote Sensing Imagery detector (ORSIm detector), integrating diverse channel features extraction, feature learning, fast image pyramid matching, and boosting strategy, is proposed to meet the demand for effectively and efficiently handling image deformations, remarkably accurate scaling and rotation [10]. The Deep Interactive U-Net (DI-U-Net) architecture is proposed to tradeoff network depth and feature spatial resolution while learning feature context representation and interaction to distinguish from the background [11]. To address the problems of severe occlusions, few pixels per head, and significant variations in a person's head sizes due to broad sports areas, Khan et al. proposed a deep learning-based method, which works as a head detector and takes into consideration the scale variations of heads in videos [12]. To alleviate the problems of significant variations in scale poses, and appearances, Khan et al. proposed an end-to-end scale consistent head detection framework that can handle a broad range of scales [13]. To deal with the problems of complex backgrounds, scale variations, nonuniform distributions, and occlusions, Basalamah et al. also proposed a scale-driven convolutional neural network (SD-CNN) model, which is based on the assumption that heads are the dominant and visible features regardless of the density of crowds [14].

Data Association is also a crucial stage in the process of multi-target tracking. Traditional data association algorithms include Nearest Neighbor Data Association (NNDA) and Joint Probabilistic Data Association (JPDA). SINGGR first proposed NNDA in 1971. The basic idea of this algorithm

is to regard the association gate as a search subspace. Only the detection points that fall within the scope of the association gate and are closest to the center of the gate are selected. The remaining detection points are regarded as false or other target detection results. The advantage of the Nearest-neighbor association algorithm is that the algorithm complexity is low, and it is easy to implement. It is suitable for sparse target tracking in a low-cluttered environment. When the targets are relatively dense, marks are likely to be lost. The Probabilistic Data Association (PDA) is a classic suboptimal Bayesian method proposed by Jaffer et al. Its basic idea is that compelling echoes may come from the target, and each echo has a different probability of coming from the target. The current prior information is used to update and filter the target. PDA can effectively track a single target, but it is easy to produce mistracking in an environment with dense marks, such as the scene where the target is blocked or overlaps. Therefore, BAR-SHA-LOM et al. extended the method and proposed a data association algorithm for multi-target tracking, JPDA. JPDA considers all echoes falling into the tracking wave gate and believes that the typical echo is not only from one target but may belong to different targets. This algorithm is suitable for multi-target tracking in a cluttered environment, but it introduces the probability of a joint event; thus, it needs enormous computation. The current popular MOT system usually adopts the data association method based on the TBD tracking framework. The core idea is to use the target detected by the detector as the input of the prediction algorithm (Kalman filter, Particle filter, etc.) to predict the trajectory state of the next frame. Then the algorithm matched the detected target in the next frame with the expected target trajectory state (Hungarian algorithm) to achieve the purpose of tracking. In the Simple Online and Real-time Tracking (SORT) algorithm, Gong et al. directly use the Hungarian matching algorithm to solve the data association between the Kalman predicted state and the new state of the target. The advantages of this algorithm are its simplicity, feasibility to implement, and high real-time performance. The disadvantage is that it needs to use the target's appearance features, leading to frequent IDSwitch problems. Based on the SORT algorithm, Gong et al. proposed the DeepSORT algorithm, which combined the target's movement and characteristic appearance information to conduct data association, significantly alleviating the IDSwitch problem [15].

This paper studies the DeepSORT algorithm [16–18] under the TBD multi-target tracking framework. It uses the YOLOv5 [19] to obtain the person target in each video frame and then performs cascade matching and IOU matching on the person targets between different pictures through DeepSORT. To solve the IDSwitch problem caused by the poor ability of the ReID network to extract person appearance features in the process of cascade matching, we introduced RGA-S and RGA-C attention mechanisms [20–22] into the network structure. Then we loaded pre-training weights to perform Transfer Learning training on the dataset CUHK03 to improve the feature extraction ability of the ReID network. We proposed a new loss function design method and introduced the Hard-Negative-Mining [23] method into the benchmark triple loss to enhance the discriminative performance of the network. We introduced the Label-Smoothing regularization method [24,25] into the cross-entropy loss to improve the classification accuracy of the network. In the training stage of the ReID network, we proposed a learning rate optimization method combining Linear-Warmup and exponential decay [26] so that the network can quickly converge in the early stage of training and can better converge to the optimal value in the later stage of training.

2 Proposed Methodology

After getting the person target of each frame through YOLOv5, the DeepSORT tracker will match and associate the detected person target. For the person target seen for the first time, DeepSORT initializes the target state to tentative, adopts IOU matching, and determines whether it is the

same target by calculating the intersection ratio between the target frames of the front and rear pictures. When the person target is successfully matched for three consecutive frames through IOU matching, the algorithm will update the status to confirm. For the person target in the established state, DeepSORT uses cascade matching, which includes appearance feature information matching and motion state information matching. The matching of appearance feature information inputs the person target detected by the current frame into the ReID network to obtain a set of corresponding feature vectors. The cost matrix is constructed by calculating the cosine distance [27] between the feature vector of the current frame target and the feature vector of the previous frame target. When matching the motion state information, the Kalman filter [28] is used to obtain the prediction target of the current frame, and the Mahalanobis distance between the current frame's prediction target and the previous frame's target is calculated to update the cost matrix. Finally, DeepSORT adopts the Hungarian matching algorithm based on the cost matrix to achieve the matching and association of person objects. Fig. 1 shows the algorithm framework of DeepSORT. $F_{\min}$ represents the minimum number of frames to confirm the track. $F_{\max}$ means the maximum number of frames the track can survive after interruption.

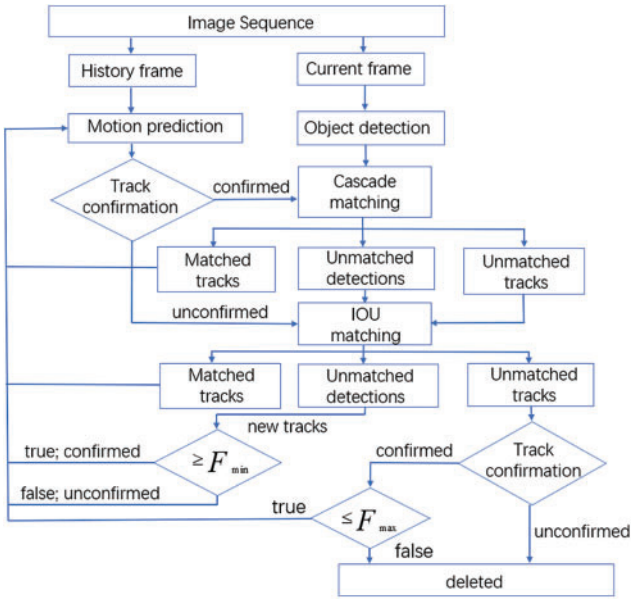


Figure 1: The algorithm framework of DeepSORT

2.1 Cascade Matching and IOU Matching

For the person target in the tentative state, DeepSORT realizes the matching and association between the front and rear frame targets through IOU matching:

$$R_{IOU} = \frac{A \cap B}{A \cup B} \tag{1}$$

where R_{IOU} is the ratio of the intersection and union area of the two detection frames, A is the detection frame of the current frame, and B is the detection frame of the previous frame.

For the person target in the confirmed state, the matching and association of the appearance feature information and the motion state information between the marks is completed through cascade matching. When matching the appearance feature information of the target, the person target frame is

input into the ReID network, and the feature vector reflecting the appearance feature of the target is extracted. By calculating the cosine distance between different feature vectors and constructing a cost matrix based on the cosine distance, the matching of appearance feature information between person objects is achieved:

$$d\left(\vec{a},\vec{b}\right)=1-\frac{\vec{a}\cdot\vec{b}}{\left\|\vec{a}\right\|\times\left\|\vec{b}\right\|} \tag{2}$$

$d\left(\vec{a},\vec{b}\right)$ is the cosine distance of two feature vectors; $\vec{a}$ and $\vec{b}$ are two eigenvectors with the same dimension.

After the cosine distance constructs the cost matrix, the algorithm must match the motion state information of the target. The Kalman filter algorithm predicts the person of the current frame. The Mahalanobis distance between the target of the recent picture and the object of the previous frame updates the cost matrix. The Kalman filter prediction stage is as follows:

$$\hat{X}_k^- = A\hat{X}_{k-1} \tag{3}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{4}$$

where $\hat{X}_k^-$ is the state vector of the prior estimation of the current frame, $\hat{X}_{k-1}$ is the state vector of the posterior estimate of the previous picture, A is the state transition matrix, P_k^- is the covariance matrix of the current image before assessment, P_{k-1} is the covariance matrix of the posterior analysis of the previous frame, and Q is the noise matrix.

The Kalman filter update stage is as follows:

$$K_k = \frac{P_k^- C^T}{CP_k^- C^T + R} \tag{5}$$

$$\hat{X}_k = \hat{X}_k^- + K_k\left(Y_k - C\hat{X}_k^-\right) \tag{6}$$

$$P_k = (I - K_k C)\,P_k^- \tag{7}$$

where K_k is the Kalman gain, which is responsible for adjusting the weight of the observed value and the predicted value to render the variance of the optimal estimated value smaller; $\hat{X}_k$ is the state vector of the current frame posteriors estimation; P_k is the covariance matrix of the current frame posteriors estimation; C is the mapping matrix, which can map the state vector to the measurement space; R is the observation noise matrix; Y_k is the state vector of the current frame detection; I is the identity matrix.

We set a threshold based on the cost matrix and calculate the Mahalanobis distance between the state vector predicted by the Kalman filter and the state vector of the previous frame. For the Mahalanobis distance that exceeds the threshold, we update the corresponding element in the cost matrix to infinity; otherwise, we keep the cosine distance unchanged and complete the updating of the cost matrix. Finally, the Hungarian matching algorithm achieves the matching and association

between pedestrian objects. The Mahalanobis distance is as follows:

$$d(X, Y) = \sqrt{(X - Y)^T P^{-1} (X - Y)} \tag{8}$$

X and Y are state vectors with the same dimension; P^{-1} is the inverse of the covariance matrix of X and Y; $d(X, Y)$ is the Mahalanobis distance between two state vectors.

2.2 The ReID Network by Introducing RGA-S and RGA-C Attention Mechanisms

The attention mechanism increases the symbolic power of the network by reinforcing features of interest and suppressing unnecessary ones. For Convolutional Neural Networks (CNN), attention mechanisms are usually learned through local convolution, which tends to ignore hidden relationships between global information and features. If the realization of a feature's importance is wanted, it is necessary to consider its relevance to other elements; thus, the global information that reflects the hidden relationships between feature points is essential. The RGA attention mechanism learns the attention weight of feature points through the correlation between features within the scope of the global structure. It includes Spatial Relation-aware Global Attention (RGA-S) and Channel Relation-aware Global Attention (RGA-C). Because of the IDSwitch problem, this paper integrated the RGA attention mechanism into the original ReID network after four blocks of feature extraction to strengthen the parts of interest and suppress the irrelevant details. The improved network structure is shown in Table 1.

Table 1: Improved ReID network structure

Layer	Number of convolution	Output size
Conv	1	64 × 128 × 64
Max Pool	0	64 × 64 × 32
Block1	10	256 × 64 × 32
RGA-S + RGA-C	12	256 × 64 × 32
Block2	13	512 × 32 × 16
RGA-S + RGA-C	12	512 × 32 × 16
Block3	19	1024 × 16 × 8
RGA-S + RGA-C	12	1024 × 16 × 8
Block4	10	2048 × 16 × 8
RGA-S + RGA-C	12	2048 × 16 × 8

For the intermediate feature tensor $X \in \mathbb{R}^{C \times H \times W}$ of the given CNN layer, RGA-S uses the C-dimensional feature vector at each spatial position as a feature node. Then all spatial parts will form a map with $N = H \times W$ feature nodes ($X_i \in \mathbb{R}^C$ denotes N features nodes, where $i = 1, \ldots, N$). The affinity relationship from the feature node X_i to X_j is expressed as follows:

$$r_{i,j} = f_s\left(X_i, X_j\right) = \theta_s (X_i)^T \phi_s\left(X_j\right) \tag{9}$$

θ_s and ϕ_s are functions containing a 1 × 1 convolution layer, BN layer, and ReLU layer.

Similarly, we can get the affinity relation $r_{j,i} = f_s\left(X_j, X_i\right)$ from the feature node X_j to X_i, and finally, we can use the affinity matrix $R_s \in \mathbb{R}^{N \times N}$ to represent the relationship between all nodes. For the attention weight of the feature node X_i, in addition to the affinity relationship with other feature

nodes, we should also consider the connection with the feature node itself to make use of the global scope structure information and the original local information related to the feature node. However, these two kinds of information do not belong to the same feature domain, so the algorithm should combine them according to the channel:

$$\widetilde{Y}_i = [pool_C\ (\psi_s\ (X_i)),\ \varphi_s\ (r_i)] \tag{10}$$

$\widetilde{Y}_i$ represents the feature tensor of spatial relational perception; ψ_s and φ_s are functions containing a 1×1 convolution layer, BN layer, and ReLU layer; r_i represents the affinity between the feature node X_i and other feature nodes; $pool_C$ represents the averaged function that is pooled by channel dimensions.

Finally, we can obtain the weight value of the feature node in the spatial position by Formula (11):

$$a_i = Sigmoid\left(w_2 \mathrm{ReLU}\left(w_1 \widetilde{Y}_i\right)\right) \tag{11}$$

w_1 and w_2 are the weight parameters of the two 1×1 convolutional layers. Fig. 2 shows the structure diagram of RGA-S.

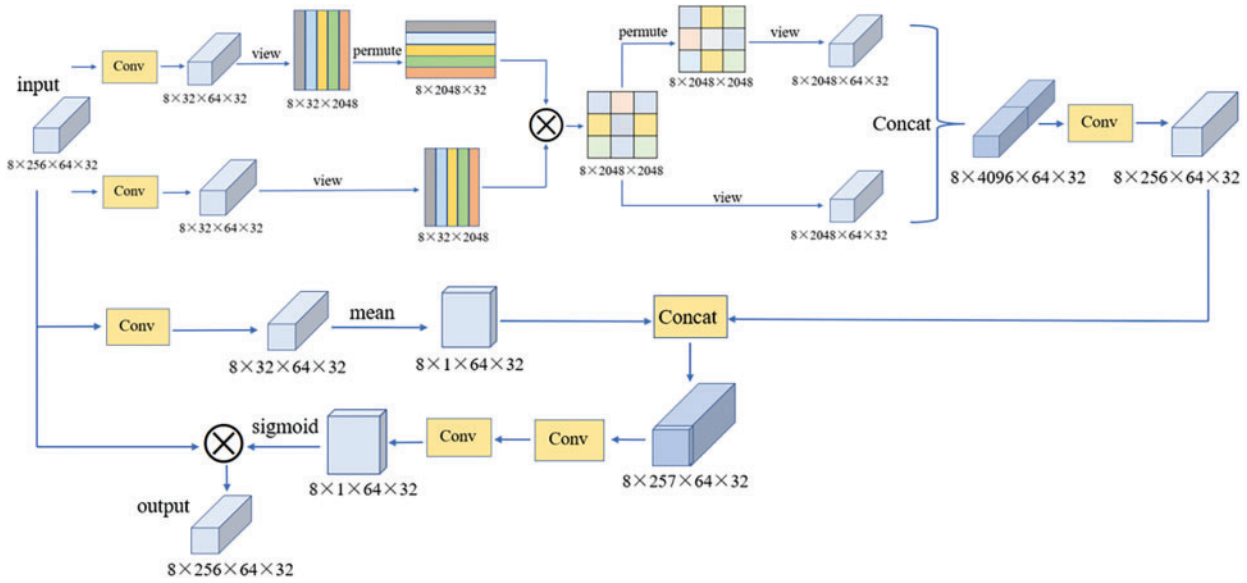


Figure 2: Structure of RGA-S

For the intermediate feature tensor $X \in \mathbb{R}^{C \times H \times W}$ of the CNN layer, RGA-C takes the $H \times W$-dimensional feature map on each channel as the feature node, and finally, all channels form a map with C feature nodes ($X_i \in \mathbb{R}^{H \times W}$ denotes C feature nodes, where $i = 1, \ldots, C$). Similar to the spatial relation, we can express the affinity relationship from the feature node X_i to X_j as follows:

$$r_{i,j} = f_C\left(X_i, X_j\right) = \theta_C\ (X_i)^T\ \phi_C\left(X_j\right) \tag{12}$$

θ_C and ϕ_C represent functions, including the 1×1 convolutional layer, BN layer, and ReLU layer. To obtain the weight of the feature node on the channel, similar to the derivation of spatial attention, in addition to considering the relationship with other feature nodes, the feature node itself should also be considered. Fig. 3 shows the structure diagram of RGA-C.

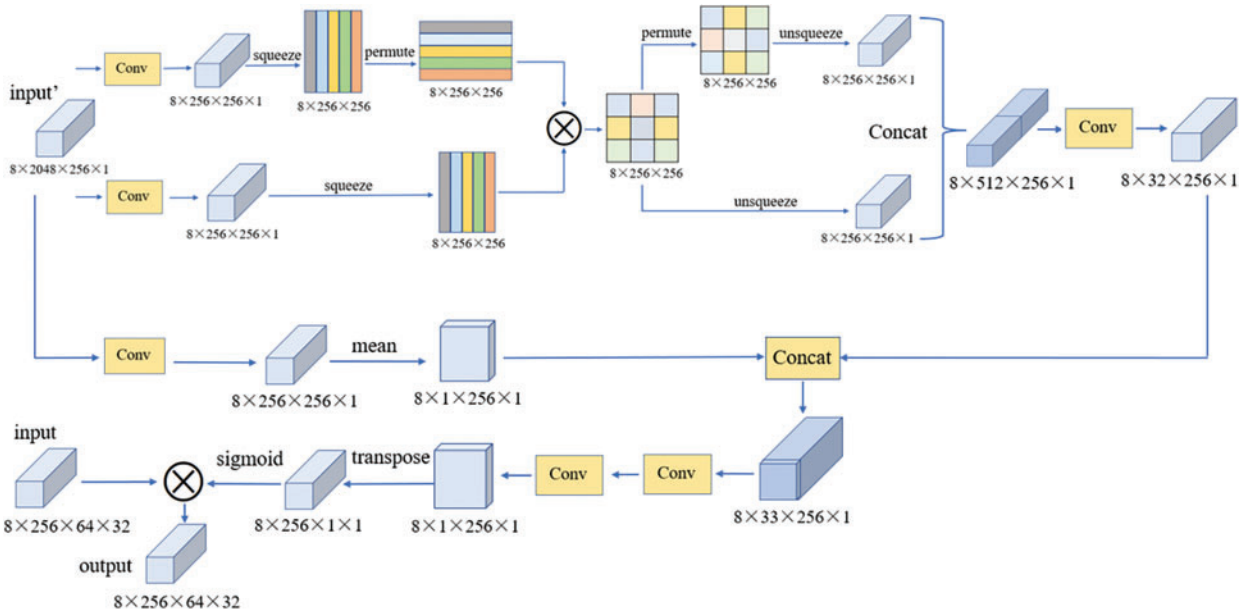


Figure 3: Structure of RGA-C

2.3 Triplet Loss by Introducing Hard-Negative Mining Method

Triplet Loss requires three pieces of data (which can be obtained from a batch), namely: current data (Anchor), similar data of the Anchor (Positive), and different categories of data from the Anchor (Negative). The three pieces of data are encoded by the ReID network, as shown in Fig. 4:

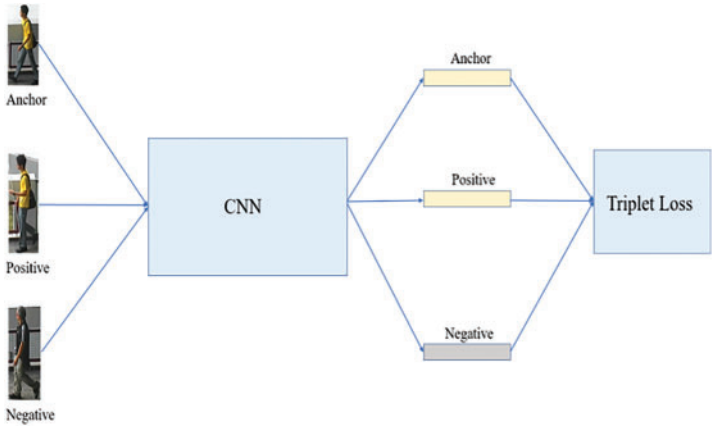


Figure 4: The encoding process of the ReID network

Among them, Triplet Loss makes the Anchor very close to the Positive and keeps the Anchor and the Negative as far away as possible, that is, to minimize the distance between feature vectors *dist* (*Anchor*, *Positive*) and to maximize the distance between feature vectors *dist* (*Anchor*, *Negative*), as shown in Fig. 5:

The formula for calculating Triplet Loss is as follows:

$$L(A,P,N) = \max\left(0, \alpha - \left(\| f(A) - f(N) \|^2 - \| f(A) - f(P) \|^2\right)\right) \tag{13}$$

$\| f(A)-f(N) \|^2$ represents the distance between Anchor and Negative; $\| f(A)-f(P) \|^2$ represents the distance between Anchor and Positive; α, a hyperparameter, stands for the distance gap. According to Formula (13), the loss is zero only when the difference between the distance from Anchor to Negative and the distance from Anchor to Positive is α. Otherwise, the value is a number greater than zero.

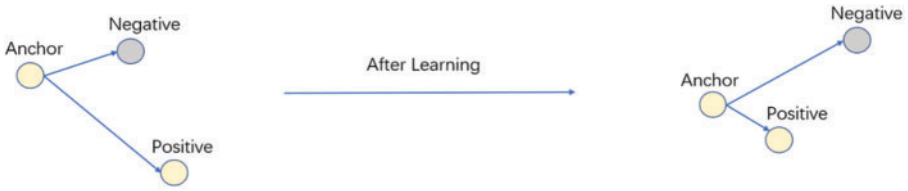
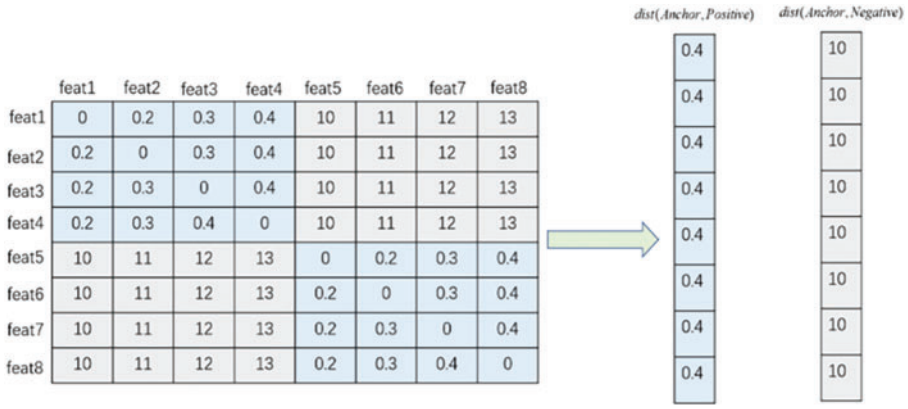


Figure 5: Effect of triplet loss

In the training stage, eight images of the row person are input. After passing through the ReID network, eight feature vectors are generated: feat1-feat8 (For the eight images of the row person, the first four images are the same person, and the last four images are another person. Therefore, feat1-feat4 corresponds to four feature vectors of the first person. The feat5-feat8 corresponds to the four feature vectors of the second person). Then we can construct an 8×8 2-dimensional cost matrix based on the Euclidean distance or the cosine distance. The elements in the cost matrix represent the distance between the two feature vectors. Finally, the triplet loss introduced by the Hard-Negative Mining method is used to measure the gap between the triplet loss and the actual value. Some difficult-to-divide negative samples are added to the loss function to enhance the network's learning ability. In other words, let the distance *dist* (*Anchor*, *Positive*) between the eigenvectors be as large as possible and let the distance *dist* (*Anchor*, *Negative*) be as small as possible. As shown in Fig. 6, the blue area represents the distance between feature vectors of the same category. The grey area represents the distance between feature vectors of different types.



	feat1	feat2	feat3	feat4	feat5	feat6	feat7	feat8
feat1	0	0.2	0.3	0.4	10	11	12	13
feat2	0.2	0	0.3	0.4	10	11	12	13
feat3	0.2	0.3	0	0.4	10	11	12	13
feat4	0.2	0.3	0.4	0	10	11	12	13
feat5	10	11	12	13	0	0.2	0.3	0.4
feat6	10	11	12	13	0.2	0	0.3	0.4
feat7	10	11	12	13	0.2	0.3	0	0.4
feat8	10	11	12	13	0.2	0.3	0.4	0

dist(Anchor, Positive)	dist(Anchor, Negative)
0.4	10
0.4	10
0.4	10
0.4	10
0.4	10
0.4	10
0.4	10
0.4	10

Figure 6: Distance calculation of the Hard-Negative-Mining method

2.4 Cross-Entropy Loss by Label Smoothing Regularization Method

In common multi-classification problems, to make the probability distribution predicted by the network on the test set close to the actual distribution, a common practice is to use one-hot to encode the proper label and then use the predicted probability to fit the real likelihood of one-hot but this poses some problems: the generalization ability of the model cannot be guaranteed, making the network confident causing over-fitting; total probability and zero probability encourage the gap between the

category and other categories to increase as much as possible, and according to the gradient bounded, this causes the model's excessive reliance on the predicted class. Introducing the cross-entropy loss of the label smoothing regularization method can alleviate these two problems:

$$T^* = (1-\varepsilon)\,T + \frac{\varepsilon}{N} \tag{14}$$

$$E = -\sum_{i=1}^{m}\left[(1-\varepsilon)\,t_i \ln(y_i) + \frac{\varepsilon \ln(y_i)}{N}\right] \tag{15}$$

T represents the real label value after one-hot encoding, N means the number of categories, ε indicates the hyperparameter whose value ranges from 0 to 1, and T^* represents the label value after smoothing. m is the dimension of the vector, y_i is the predicted value of the network, t_i means the actual label value, and E is the cross-entropy loss.

According to Formula (14), this regularization method makes the probability of the label with ε come from a uniform distribution. The probability of the label with $1-\varepsilon$ comes from the original distribution, equivalent to adding noise to the original brand. The model's predicted value will not be excessively concentrated on the category with a higher probability but also the type with a lower likelihood for a better generalization of the network.

2.5 Learning Rate Adjustment Method Combining Linear Warmup and Exponential Decay

In the early stage of model training, since the weight parameters are randomly initialized, the model may be unstable if a significant learning rate is selected. Therefore, the Linear Warmup method is used to adjust the learning rate in the early stage of training. At the beginning of a few epochs of training, the first use of preheating generates a small learning rate so that the model can slowly lean to stability. It helps to slow down the model in the initial stage of the mini-batch in advance of the overfitting. It can maintain a smooth distribution, helping preserve the model's robust stability. At the later stage of training, if a constant learning rate is used for training, the model will oscillate near the optimal solution, failing to reach the optimal solution of the lowest point of the loss function. Therefore, the exponential decay method is adopted to adjust the learning rate. Near the optimal solution, the gradient decreases gradually, and the corresponding learning rate drops, enabling the model to converge smoothly to the correct expected value. Fig. 7 shows the changing trend of the learning rate.

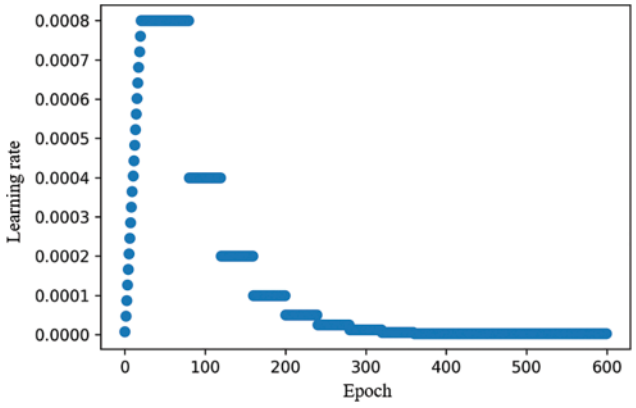


Figure 7: Change in learning rate

In the first 20 epochs, the model is in the Linear-Warmup stage, and the learning rate will increase linearly. In the 20–79 epochs, the learning rate remains at the introductory learning rate. Then every 40 epochs, the learning rate becomes half of the original in an exponentially decaying way. After 360 epochs have been iterated, the learning rate remains unchanged.

3 Experimental Results and Analysis

The hardware configuration of the experimental platform includes Intel(R) Core(TM) i7-10875H CPU @ 2.30 GHz, NVIDIA GeForce RTX 2060, etc. Software configuration includes Windows 10 operating system; Compute Unified Device Architecture (CUDA) 11.4; Pytorch 1.7.1; Tensorboard 2.4.1; Python 3.8.5, etc. To solve the IDSwitch problem caused by the poor ability of the ReID network to extract person appearance features in the process of cascade matching, we introduced RGA-S and RGA-C attention mechanisms into the network. We loaded pre-training weights to carry out Transfer Learning training on the dataset CUHK03. Then we compared the mAP and CMC evaluation indexes of the improved ReID network and the original structure on the CUHK03 verification set. Finally, we used actual street videos to test the DeepSORT algorithm optimized in this paper for person multi-object tracking and evaluate the algorithm's performance based on the experimental results.

3.1 Evaluation Metrics

The algorithm used mean Average Precision (mAP), Cumulative Matching Characteristics (CMC), and model size to evaluate the results. The mAP is used to assess the overall effect of the person re-identification algorithm, as shown in Formula (16). CMC represents a Top-k hit probability and is also used to determine the performance of the ReID algorithm.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{16}$$

where AP_i represents the average precision of a query sample; N represents the number of all categories; mAP represents the average value of AP in all query samples, reflecting the overall effect of the model on all samples.

The person multi-target tracking algorithm uses Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), IDSwitch times, and Frames Per Second (FPS) to evaluate the tracking effect. MOTA and MOTP are shown in Formulas (17) and (18).

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{17}$$

m_t represents the number of false negative samples (FN); fp_t is the number of false positive samples (FP); mme_t means the number of identity switches; g_t is the number of real targets.

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{18}$$

d_t^i represents the distance between the hypothetical box and the real box; c_t is the total number of matches at frame t.

3.2 Person ReID Training and Validation

The methods of normalized and Random Erasing [29] are mainly used for data enhancement. Normalization converts the data into a standard Gaussian distribution, as shown in Fig. 8. It

normalizes the person images channel by channel (mean becomes 0 and standard deviation becomes 1), which can speed up the convergence of the model. Random Erasing in the original image selects a random rectangular area. It replaces the pixel values of this area with arbitrary values to improve the robustness of the model, as shown in Fig. 9.

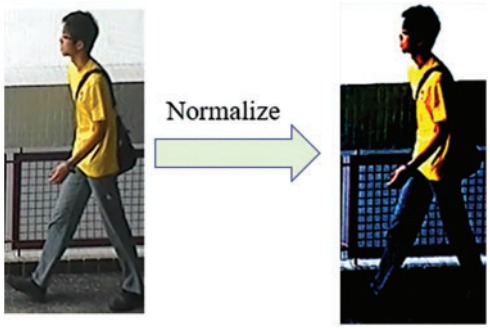


Figure 8: Normalization

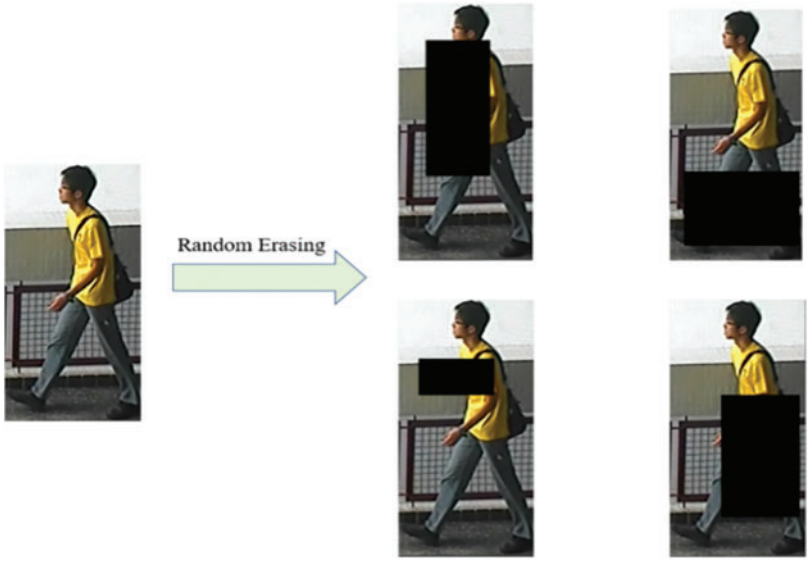


Figure 9: Random Erasing

The dataset contains 1467 persons, each with about ten images, for a total of 14097 images. The training set includes 767 persons, a total of 7365 images. The query set in the validation set contains 700 persons, each with two pictures, a total of 1400 images. The gallery set in the validation set includes 700 persons, with about eight pictures for an individual, a total of 5332 images. On the CUHK03 training set, we respectively trained the original ReID network and the improved ReID network. In the two training sessions, we set Epoch = 600 and Batchsize = 8 and used the combination of Linear-Warmup and exponential decay to adjust the learning rate. After 229800 iterations, the comparison results of loss value and Top-1 accuracy are shown in Figs. 10 and 11, respectively.

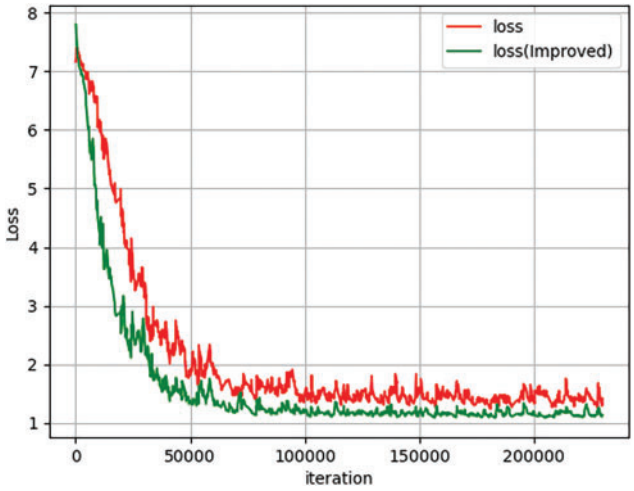


Figure 10: Loss value comparison

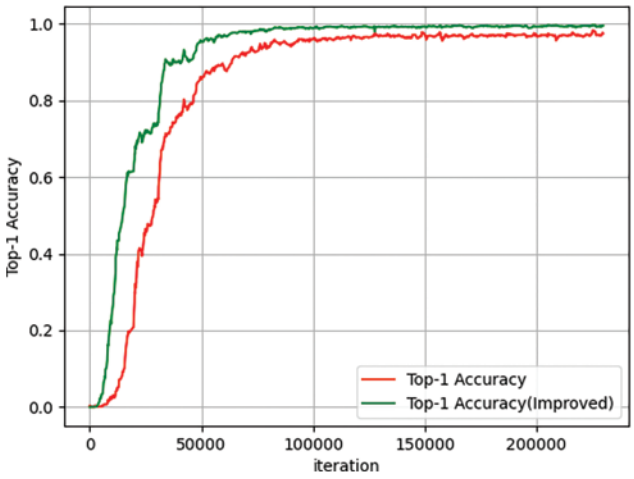


Figure 11: Top-1 accuracy comparison

The loss value includes triplet loss and classification loss. The improved ReID network consists of the RGA-S and the RGA-C attention mechanism, the introduction of the Hard-Negative-Mining method in the benchmark triplet loss, and a Label-Smoothing regularization method. The training results show that the two network models can converge well after 229,800 iteration training. The original ReID network Top 1 accuracy is stable at 94.5%, and the improved ReID network Top 1 accuracy is stable at 99.5%, with the accuracy increased by 5%. The comparison figure of loss function before and after using the learning rate regulation method is shown in Fig. 12.

In the verification stage, the query set's feature vectors and the gallery set's feature vectors are matched based on cosine distance. The mAP and CMC were used to evaluate the performance of the ReID network. Table 2 shows the ablation results. The results show that the mAP of the improved ReID network is improved by 7.5% compared with the original ReID network. CMC Top 1 increased by 9.3%, Top 5 by 9.4%, and Top 10 by 7%; The model size was 359 MB, and the number of parameters was increased by 15.4%.

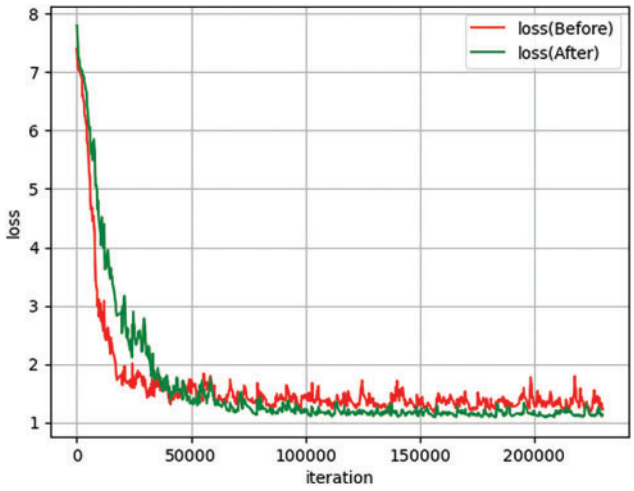


Figure 12: Loss function before and after using the learning rate regulation method

Table 2: Ablation experiment results of ReID network on the CUHK03 validation set

RGA-SC	Hard-Negative-Mining	Label Smoothing	mAP	CMC Top 1	CMC Top 5	CMC Top 10	Params (MB)
_	_	_	0.690	0.332	0.560	0.673	311
_	_	√	0.708	0.344	0.572	0.688	311
_	√	_	0.720	0.362	0.607	0.704	311
_	√	√	0.733	0.376	0.616	0.711	311
√	_	_	0.730	0.368	0.600	0.712	359
√	_	√	0.748	0.380	0.622	0.728	359
√	√	_	0.760	0.402	0.639	0.739	359
√	√	√	0.765	0.425	0.654	0.743	359

3.3 Person Multi-Target Tracking

Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) are adopted as tracking effect evaluation metrics in the multi-target tracking experiment. The larger the value, the better the effect is; Meanwhile, we also recorded the Frames Per Second (FPS) and IDSwitch times. The experimental results are shown in Table 3. The results show that the MOTA of the improved DeepSORT is 65.3%, which is improved by 2.5%. MOTP was 78.4%, with an increase of 3.8%; The IDSwitch occurs 21 times, which is reduced by 25% compared with the original algorithm. The FPS dropped by 23%.

Table 3: Comparison results of improved DeepSORT and the DeepSORT

Model	MOTA	MOTP	IDSwitch	FPS
YOLOv5 + DeepSORT	0.628	0.746	28	30
YOLOv5 + DeepSORT (improved)	0.653	0.784	21	23

We used daytime and nighttime street videos to test the improved DeepSORT and DeepSORT for person multi-object tracking. Fig. 13 shows the daytime comparison results.

Figure 13: Comparison of DeepSORT and improved DeepSORT tracking result (day). (a–c) DeepSORT tracking result; (d–f) improved DeepSORT tracking result

According to the test results during the day, under the condition that a new person constantly appears in the video, a person frequently disappears from the video, and the person continually blocks each other. For the original DeepSORT, from Figs. 13a to 13b, due to the occlusion, the ID = 7 becomes ID = 24, and the ID = 9 becomes ID = 23. From Figs. 13b to 13c, the IDs of the two persons restore the original value. For the improved DeepSORT, from Figs. 13d to 13f, the ID of each person stays the same, without the problem of IDSwitch, and can still maintain excellent tracking.

Fig. 14 shows the nighttime comparison results.

Figure 14: Comparison of DeepSORT and improved DeepSORT tracking result (night). (a–c) DeepSORT tracking result; (d–f) improved DeepSORT tracking result

For the original DeepSORT, we can see from Figs. 14a–14c, the ID = 2 becomes ID = 4, the ID = 4 becomes ID = 2, and the ID = 7 becomes ID = 16. The frequency of identity switching is relatively high. However, the improved DeepSORT significantly suppresses the IDSwitch problem.

For the improved DeepSORT, we introduce the RGA-S and RGA-C attention mechanisms to learn the attention weights of feature points within the scope of the global structure through the correlation between features. In the model reasoning stage, the output of the attention-weight feature map from the middle layer of the network is extracted and weighted with the original picture. Figs. 15 and 16 show the test results during the day and the night.

Figure 15: Visualization of attention mechanism (day)

Figure 16: Visualization of attention mechanism (night)

4 Conclusion

In this paper, we proposed a multi-target tracking algorithm for persons based on deep learning to improve the tracking accuracy in surveillance video. The experimental results have shown that the proposed algorithm has significantly improved performance compared with the original algorithm. The algorithm effectively alleviates the IDSwitch problem, improves the tracking accuracy of persons, and has a high practical value. The person re-identification dataset used in this paper was collected from the Chinese University of Hong Kong campus. The height and angle changes of camera shots are relatively simple, which has a particular impact on the robustness of the model. The MOTA and MOTP of the optimized DeepSORT improved significantly, but the number of model parameters also increased, resulting in a decrease in FPS. In the follow-up work, we will collect more data with different angle changes to expand the original dataset. At the same time, since the performance of the tracking algorithm depends on the accuracy of the target detection model, a more lightweight and efficient network structure is suggested to improve the performance of the detection algorithm.

Funding Statement: The authors received no specific funding for this study.

Availability of Data and Materials: The data used to support the findings of this study are included within the article.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. Li, Y. F. Cha, T. Z. Zhang, Z. Cui, W. M. Zuo *et al.,* "A survey of object tracking algorithms based on deep learning," *Journal of Image and Graphics*, vol. 24, no. 12, pp. 2057–2080, 2019.

[2] Q. Chu, "Research on video multi-object tracking algorithm based on deep learning Ph.D. Dissertation," *University of Science and Technology of China*, Hefei, China, 2019.

[3] D. G. Xu, L. Wang and F. Li, "A survey of typical object detection algorithms based on deep learning," *Computer Engineering and Applications*, vol. 57, no. 8, pp. 10–25, 2021.

[4] R. Girshick, "Fast R-CNN," in *IEEE Int. Conf. on Computer Vision*, Santiago, Chile, New York, pp. 1440–1448, 2015.

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

[6] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018. [Online]. Available: http://arXiv:1804.02767

[7] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 7263–7271, 2017.

[8] X. Wu, W. Li, D. Hong, R. Tao and Q. Du, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 91–124, 2022.

[9] X. W. Xu, X. L. Zhang, T. W. Zhang, Z. Y. Yang, J. Shi *et al.,* "Shadow-background-noise 3D spatial decomposition using sparse low-rank Gaussian properties for video-SAR moving target shadow enhancement," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, no. 4516105, pp. 1–5, 2022.

[10] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li *et al.,* "ORSIm Detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 5146–5158, 2019.

[11] X. Wu, D. Hong, Z. Huang and J. Chanussot, "Infrared small object detection using Deep Interactive U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, no. 6517805, pp. 1–5, 2022.

[12] S. D. Khan, H. Ullah, M. Ullah, N. Conci, F. A. Cheikh *et al.,* "Person head detection based deep model for people counting in sports videos," in *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, Taipei, Taiwan, pp. 1–8, 2019.

[13] S. D. Khan and S. Basalamah, "Scale and density invariant head detection deep model for crowd counting in pedestrian crowds," *The Visual Computer*, vol. 37, no. 8, pp. 2127–2137, 2021.

[14] S. Basalamah, S. D. Khan and H. Ullah, "Scale Driven convolutional neural network model for people counting and localization in crowd scenes," *IEEE Access*, vol. 7, pp. 71576–71584, 2019.

[15] X. Gong, Z. C. Le, H. Wang and Y. K. Wu, "A survey of data association techniques in multi-target tracking," *Computer Science*, vol. 47, no. 10, pp. 136–144, 2020.

[16] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE Int. Conf. on Image Processing*, Beijing, China, pp. 3645–3649, 2017.

[17] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking," in *Int. Conf. on Image Processing*, Phoenix, AZ, USA, pp. 3464–3468, 2016.

[18] X. L. Zhu, H. Cai, T. T. Kou, D. H. Du and J. X. Sun, "Person multi-target tracking algorithm," *Journal of Jilin University (Science Edition)*, vol. 59, no. 5, pp. 1161–1170, 2021.

[19] J. Yu and S. Luo, "A YOLOv5-based method for unauthorized building detection," *Computer Engineering and Applications*, vol. 57, no. 20, pp. 236–244, 2021.

[20] Z. Zhang, C. Lan, W. Zeng, X. Jin and Z. Chen, "Relation-aware global attention for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA, pp. 3186–3195, 2020.

[21] Z. Niu, G. Zhong and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[22] L. F. Chen, J. S. Le, H. Y. Wu, C. S. C. Zhu and S. C. Ye, "Person re-identification algorithm combining attention and local feature fusion," *Computer Engineering and Applications*, vol. 58, no. 14, pp. 282–290, 2022.

[23] D. Thapar, G. Jaswal, A. Nigam and V. Kanhangad, "PVSNet: Palm vein authentication siamese network trained using triplet loss and adaptive hard mining by learning enforced domain specific features," in *IEEE 5th Int. Conf. on Identity, Security, and Behavior Analysis (ISBA)*, Hyderabad, India, pp. 1–8, 2019.

[24] R. Müller, S. Kornblith and G. E. Hinton, "When does label smoothing help," in *Advances in Neural Information Processing Systems*, Vancouver, Canada, pp. 4694–4703, 2019.

[25] M. Lukasik, S. Bhojanapalli, A. K. Menon and S. Kumar, "Does label smoothing mitigate label noise," in *Int. Conf. on Machine Learning*, Vienna, Austria, pp. 6448–6458, 2020.

[26] J. R. Shi, D. Wang, F. H. Shang and H. Y. Zhang, "Advances in stochastic gradient descent algorithms," *Acta Automatica Sinica*, vol. 47, no. 9, pp. 2103–2119, 2021.

[27] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *IEEE Winter Conf. on Applications of Computer Vision*, Lake Tahoe, NV, USA, pp. 748–756, 2018.

[28] Z. Y. Qin, J. Huang, X. Yang, S. Y. Zheng and G. D. Fu, "Multi-target tracking based on attention mechanism and Kalman Filter," *Computer Systems & Applications*, vol. 30, no. 12, pp. 128–138, 2021.

[29] B. Xie, "Research on person re-identification algorithm based on deep learning M.S. Thesis," *Nanjing University of Posts and Telecommunications*, Nanjing, China, 2021.