Method for Top View Pedestrian Flow Detection Based on Small Target Tracking

Ming Li[1,2*], Hui Dong[3], Fei Zhang[2], Xiaoxiao Liu[2]
[1]School of Computer, Electronic and Information, Guangxi University, Nanning 530004, China
[2]Intelligent Manufacturing Department, Zaozhuang Vocational College, Zaozhuang 277000, China
[3]Medical School, Zaozhuang Vocational College, Zaozhuang 277000, China
E-mail: liming860311@163.com

Keywords: target tracking, vision transformer, deep sort, abortion detection

Received: April 17, 2024

In public spaces, monitoring pedestrian flow can effectively avoid the occurrence of crowding and stampede incidents, and can effectively improve public safety. To improve the accuracy and efficiency of small target tracking in pedestrian flow detection from a top-down perspective, this study integrates the Vision Transformer architecture and the Deep SORT tracking algorithm to improve the YOLOv5 model. This method aims to achieve efficient detection in complex pedestrian flow scenarios by enhancing the recognition ability and tracking accuracy of small targets. In the experiment, the improved YOLOv5-V-D model quickly converged after 61 iterations, achieving excellent operational efficiency with an average delay of only 7.2ms. Compared to CenterNet and RetinaNet, it has increased by 3.9ms and 6.4ms, respectively. Furthermore, the model demonstrated an exceptional capacity for accurately predicting pedestrian flow, with a prediction accuracy of 98.72%, which is significantly higher than the comparison model's range of 20.59% to 28.61%. In summary, the improved YOLOv5 model not only provided faster detection speed, but also significantly improved the accuracy of pedestrian flow detection. This advancement offers an efficacious solution for the monitoring of high-density crowds, establishing a robust foundation for the advancement of future real-time monitoring systems and significantly enhancing public safety.

Povzetek: Predlagana je metoda za zaznavanje pešcev z vidika od zgoraj, ki temelji na sledenju majhnih tarč. Izboljšuje zaznavanja pešcev s pomočjo izboljšanega modela YOLOv5, Vision Transformer in algoritma Deep SORT.

1 Introduction

The rapid development of Artificial Intelligence (AI) technology has led to the increasing application of real-time video surveillance systems in many fields. It is of great significance for maintaining public order, optimizing traffic management, and formulating emergency evacuation plans [1-2]. However, traditional monitoring systems are often limited by resolution and the influence of complex backgrounds when detecting pedestrian flow at a top view angle, making it difficult to accurately identify and track [3]. In response to this challenge, this study fully utilizes the advantages of YOLOv5 network and Vision Transformer (ViT) to construct a more adaptive Small Target Detection (STD) method to improve the accuracy of Pedestrian Flow Detection (PFD) from a Top-down Perspective (TP-PFD). This method is based on YOLOv5 and relies on its lightweight and efficient characteristics to ensure the speed and sensitivity of the model when processing real-time video streams [4]. To further improve the recognition ability of small targets, ViT has been introduced. This technology effectively captures global information in images through a Self Attention Mechanism (SAM), which helps to distinguish between pedestrian flow and background in complex scenes [5]. By integrating the fine-grained feature recognition advantages of ViT, the accuracy of STD can be enhanced. In addition, the enhanced data preprocessing module further enriches the model's adaptability to small target shapes. The innovation of this study lies in the combination of YOLOv5 and ViT, which enhances YOLOv5's shortcomings in STD and improves feature utilization. The proposal of this method aims to improve the robustness and accuracy of TP-PFD, thereby better serving application scenarios such as public safety monitoring, crowd management, and business analysis. It is hoped that through this paper, a new technological path can be provided for relevant fields, which will have a positive driving effect on PFD and analysis from a top-down perspective. As as result, while ensuring public safety, it can promote the development of smart city construction. The study is divided into four parts. The first part is a summary of the fields of Small Target Tracking (STT) and PFD. Part 2 is the implementation of the proposed improvement method. Part 3 is the validation and testing of the research method. Part 4 is a summary of the entire text.

2 Related works

STT detection is an important research direction in the field of computer vision that focuses on detecting and tracking smaller objects in images. In application scenarios such as surveillance videos, drone images, and satellite images, small targets become very challenging to detect and track due to their small pixel size, lack of detailed information, and susceptibility to environmental noise and occlusion. With the improvement of computing power and the advancement of deep learning technology, STT detection algorithms are moving towards more accurate, real-time, and robust directions. Shi et al. proposed a sea surface small target-feature detector based on dispersion relative entropy. This method was superior to existing single-feature detectors in suppressing clutter and improving detection performance, and could be comparable to three-feature detectors [6]. Yang et al. proposed an improved helmet detection algorithm based on YOLO V4 to address the issue of existing helmet detection algorithms being susceptible to occlusion. This algorithm significantly improved the detection accuracy of small and occluded targets, and optimized the convergence speed and regression accuracy of model training [7]. Zhi et al. proposed a framework called attention context region detection for precise recognition of small and medium-sized traffic signs in intelligent transportation systems. This framework utilized attention contextual features and combined target and environmental information through dot convolutional layers, achieving advanced levels in small traffic sign detection [8]. Bommes et al. proposed a ResNet-34 Convolutional Neural Network (CNN) based on unsupervised domain adaptation problem for the automatic detection of module faults in photovoltaic (PV) systems. It combined supervised comparison loss and K-means cluster classifier for anomaly detection of small target images. In nine combinations of four source and target datasets containing 2.92 million infrared images, its classification accuracy for normal and abnormal images reached 79.4% and 77.1%, and it could reliably detect unknown types of anomalies [9]. Qin et al. proposed a dense sampling and detail enhancement network to address the issue of insufficient performance of existing object detection algorithms in STD. It improved feature map resolution and expanded receptive fields through dense sampling modules. On the Minico2021 and VisDrone datasets, this method improved by approximately 4.6% and 4.2% compared to the advanced DetectoRS algorithm, respectively [10].

PFD refers to the use of various sensors or video devices to estimate the number of people in a specific area. This field focuses on how to accurately count each individual in a population, and is a key sub-field in computer vision and image processing. PFD technology has extensive applications in various fields such as business analysis, public safety, urban planning, and traffic management. The advancement of machine learning technology and the increase in computing resources are continuously improving the accuracy and efficiency of PFD. Minegishi et al. proposed a pedestrian flow simulator based on actual physical parameters to address the challenge of corridor fire evacuation. When the density exceeded 2.35 people/square meter, pedestrians exhibited stagnant behavior, with a direct increase in speed and spacing, while a specific flow rate increased linearly with density, and density was inversely proportional to speed [11]. Yang et al. proposed a deep learning detection method based on a single multi-frame detector for precise target recognition and localization in smart city applications. This algorithm optimized the network structure through VGG16, achieving a maximum mAP of 77% and an accuracy of 96.31% [12]. Song et al. proposed a progressive refinement network for pedestrian detection in complex occlusion situations. The proposed method effectively improved the accuracy and domain adaptability of occluded pedestrian detection [13]. Yang et al. proposed a deep learning detection method based on SSD to address the impact of crowded subway stations on large pedestrian traffic. This method had higher performance compared to other mainstream detection methods [14]. Zhang S et al. proposed an asymmetric multi-stage network to address the challenge of small-scale pedestrian detection. It utilized rectangular anchor frames and asymmetric convolutional kernels to address pedestrian body asymmetry, improved detection performance through three-stage gradual feature selection, and demonstrated excellent performance in benchmark testing [15]. The summary table of the related works is shown in Table 1.

Table 1: Related works summary table

Research	Major technology	Application scenario	The state-of-the-art gap
Shi et al	Feature detector of sea surface small target based on dispersion relative entropy	Marine surveillance video	Insufficient capture of small target details
Yang and Wang	Improved helmet detection algorithm based on YOLO V4	Security monitoring	The detection accuracy of occluded target is limited
Zhi et al	Attention context area detection framework	Intelligent transportation system	Small traffic sign detection
Bommes and	ResNet-34 network based on	Photovoltaic system	Accuracy of anomaly

Hoffmann	unsupervised domain adaptation problem	module fault detection	detection
Qin et al	Intensive sampling and detail enhancement network	Drone image	STD performance
Minegishi et al	Flow simulator based on physical parameters	Fire evacuation	Dynamic flow simulation
Yang et al	Deep learning detection method based on single frame detector	Smart city	Target recognition and location
Song et al	Progressive refinement network	Complex occlusion condition	Blocked pedestrian detection
Zhang et al	Asymmetric multistage network	Small scale pedestrian detection	The pedestrian is asymmetrical
Research method	YOLOv5 and ViT	Public space monitoring	Small target recognition and tracking

In summary, the current literature on STT and PFD methods has demonstrated significant advantages based on deep learning models. Nevertheless, research on PFD under overhead angles is still lacking. There are still challenges in dealing with extremely dense crowd scenarios, especially in improving the recognition rate of small targets while maintaining high detection speed. Therefore, this study proposes a TP-PFD method based on STT. It employs an enhanced YOLOv5 architecture, integrating the strengths of ViT. This integration leverages the lightweight nature of ViT to enhance the STD capabilities of the native network, ultimately aiming to achieve accurate TP-PFD.

3 Construction of STT algorithm for TP-PFD

To improve the accuracy of TP-PFD, this study constructs an improved algorithm by integrating the Token-to-Token ViTs (T2T-ViT) architecture into the YOLOv5 framework. This algorithm aims to optimize YOLOv5's recognition ability for small targets, and enhance the model's resolution and tracking accuracy for individuals in crowded scenes. The introduction of T2T-ViT aims to enhance the richness of feature expression through its SAM, thereby enhancing the robustness of the model to small targets in complex backgrounds. It is expected that through this algorithm improvement, the accuracy and stability of TP-PFD can be improved while maintaining real-time performance.

3.1 Construction of YOLOv5 Algorithm for PFD

In fields such as surveillance video analysis, intelligent transportation systems, and public safety, PFD is a key technology used to estimate the number of people in specific areas, monitor pedestrian density, and track individual movements [16]. Observing the crowd from top to bottom with a top-down angle reduces occlusion issues and helps with more accurate flow counting and behavior analysis. However, from a top-down perspective, the size of the human body in the image is usually small, and the interaction and occlusion between individuals can lead to a decrease in detection accuracy. Monitoring the dynamic lighting changes in the scene can also pose challenges for PFD. YOLOv5 is an extremely fast object detection model that can meet the needs of real-time monitoring, and it has good performance on STD [17]. Meanwhile, due to the excellent customization performance of YOLOv5, it can be targeted for expansion and optimization according to actual needs. Its relatively small number of parameters is also the reason for its excellent deployment difficulty [18]. Therefore, this study chooses YOLOv5 as a strong candidate model for TP-PFD. This model can combine speed and accuracy to provide reliable detection performance in challenging scenarios. Figure 1 shows the process of applying YOLOv5 model to TP-PFD in this study.

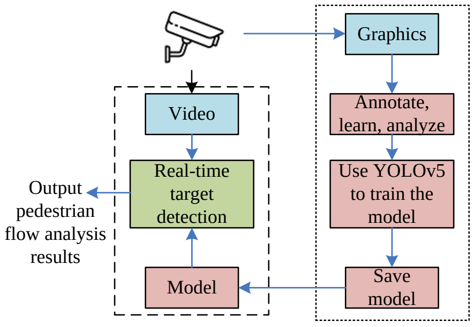

Graphics
Video
Annotate, learn, analyze
Real-time target detection
Use YOLOv5 to train the model
Output pedestrian flow analysis results
Model
Save model

Figure 1: Flow framework of YOLOv5 model under TP-PFD

The Transformer architecture is a deep learning model proposed in 2017. It was originally designed to solve sequence problems in natural language processing. Afterwards, Transformer quickly became the mainstream model in the field of natural language processing due to its efficiency and powerful performance, and gradually expanded to computer vision and other fields. One of the key innovations of Transformer is that it is entirely based on the "SAM". This mechanism allows the model to dynamically focus on other elements in the sequence while processing each element, thereby capturing their relationships. Its unique SAM allows the model to calculate attention scores at different positions in the sequence, allowing the model to focus on the relevant parts of the input sequence. This is very useful for understanding long-distance dependencies in sequences. The Transformer model can parallelize multiple SAMs, with each focusing on different sub-spaces of input, which improves the model's ability to capture different types of information [19]. ViT is the first attempt to directly apply Transformer to image classification, which divides images into multiple small pieces (tokens) and then processes them using a standard Transformer model. Although ViT performs well on large-scale datasets, its performance is weaker on small-scale images or datasets. This is partly because it generates significant information loss when dealing with these situations, especially in terms of its ability to capture fine-grained features. Figure 2 shows the network structure of ViT.

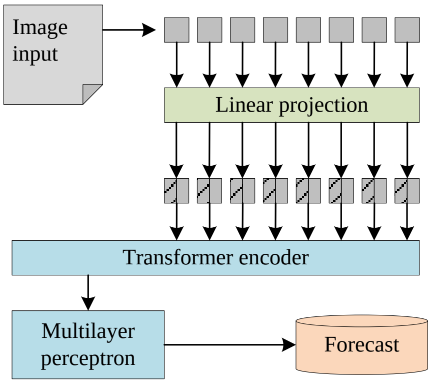


Figure 2: ViT network structure diagram

To overcome the limitations of ViT on small-scale images or datasets, an innovative T2T-ViT mechanism is introduced. This mechanism can effectively aggregate locally relevant feature information together. Specifically, T2T-ViT recursively merges adjacent image blocks (tokens) into higher-level tokens to construct a hierarchical representation. This method not only improves the model's ability to express small targets and complex textures, but also reduces the complexity and computational cost of the model. In traditional ViT design, an image is divided into a series of fixed size blocks, each of which is flattened and linearly projected into a token. Then these tokens are fed into the Transformer structure for processing. However, this method usually ignores local structural information between blocks, especially when the image details are relatively fine. To address this issue, T2T-ViT proposes a token-to-token conversion mechanism. In this conversion process, the model first performs Soft Split (SS) on the image, which means that the segmented blocks are allowed to have overlapping parts to retain more local information. Next, the model recursively merges adjacent tokens into new tokens, similar to the way features are gathered layer by layer in a neural network. After each merge, the model is able to obtain a more global and abstract image representation, while reducing the number of tokens required for processing. Through this approach, T2T-ViT can maintain and refine the image representation at each step, allowing the model to capture global information while also paying attention to local details of the image. This makes T2T-ViT more effective in processing delicate image features, especially in small target recognition and fine-grained classification tasks, showing better performance than ViT. Figure 3 shows the T2T module.

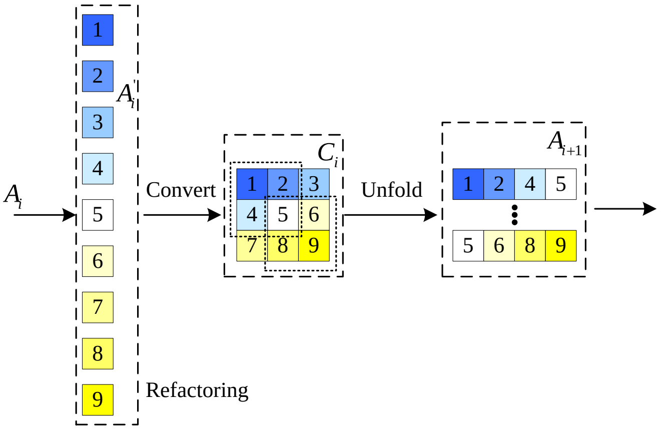


Figure 3: Diagram of the T2T module

The reconstruction process in Figure 3 can be specifically represented as shown in formula (1).

$$A_i^{'} = MLP\left(MSA\left(A_i\right)\right) \qquad (1)$$

In formula (1), $A_i^{'}$ represents the reconstructed output. MLP represents a multi-layer perception module. MSA represents the multi-head attention mechanism. A_i represents the initial input. The SS process can be described as formula (2).

$$C_i = Reshape\left(A_i^{'}\right) \tag{2}$$

In formula (2), C_i represents the conversion output, and $Reshape$ represents the total conversion operation of the module. The final output representation is formula (3).

$$A_{i+1} = SS\left(a_i\right) \tag{3}$$

In formula (3), A_{i+1} represents the output after SS and SS represents the SS operation. T2T-ViT also reduces the burden of self-attention computation in the Transformer model through this structured merging approach. Due to the number of merged tokens has decreased, the computational complexity of self-attention has also decreased. This design not only improves the model's expressive power, but also makes it more suitable for use in environments with limited computing resources. In summary, T2T-ViT enhances the model's ability to capture image details through its innovative token to token conversion process, and improves computational efficiency by reducing the number of tokens. Therefore, the application of the Transformer architecture in image related tasks becomes more powerful and efficient.

3.2 Improvement of PFD YOLOv5 algorithm

When applying the YOLOv5 algorithm to PFD, its performance optimization has become the focus of research attention. The limitations of YOLOv5 in detection accuracy and speed have been thoroughly analyzed to explore targeted improvement measures. These improvements aim to enhance YOLOv5's detection ability in complex pedestrian flow scenarios through structural adjustments and algorithm optimization. Due to the high similarity of pedestrian targets in TP-PFD, to further improve the accuracy of the model, this study chooses to introduce prior knowledge to make targeted improvements to the network. This study introduces an attention mechanism for the T2T module, which performs another conversion after the image is converted. The specific module architecture is Figure 4.

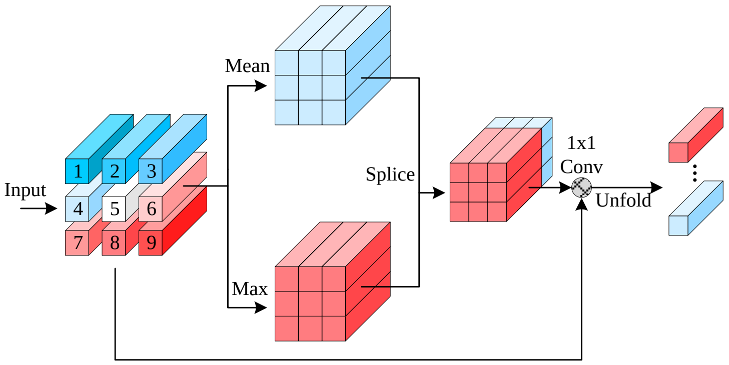


Figure 4: Improved T2T module for TP-PFD

Figure 4 shows the T2T module improvement mechanism that takes into account the similarity of TP-PFD. The process of taking the average value can be expressed as formula (4).

$$A_{mean} = Mean\left(A_i\right) \tag{4}$$

In formula (4), $Mean$ represents averaging. A_{mean} represents the output value processed by $Mean$. A_i represents the input value. The process of taking the maximum value is formula (5).

$$A_{\max} = Max\left(A_i\right) \tag{5}$$

In formula (5), $A_{\max}$ represents the maximum output value, and Max represents taking the maximum value. The attention mechanism is represented by formula (6).

$$Attention = Conv_{1\times1}\left(Cat\left(A_{mean}, A_{\max}\right)\right) \tag{6}$$

In formula (6), $Attention$ represents the output result of the attention mechanism. $Conv_{1\times1}$ represents the convolution operation of 1×1, which is mainly aimed at dimensionality reduction. Cat represents Concate, with the aim of connecting A_{mean} with $A_{\max}$. I_i^a represents the output result after introducing attention mechanism recombination. The purpose of sub-optimization in the study is to further enhance the feature learning performance of the model in TP-PFD. The improved T2T module is Figure 5.

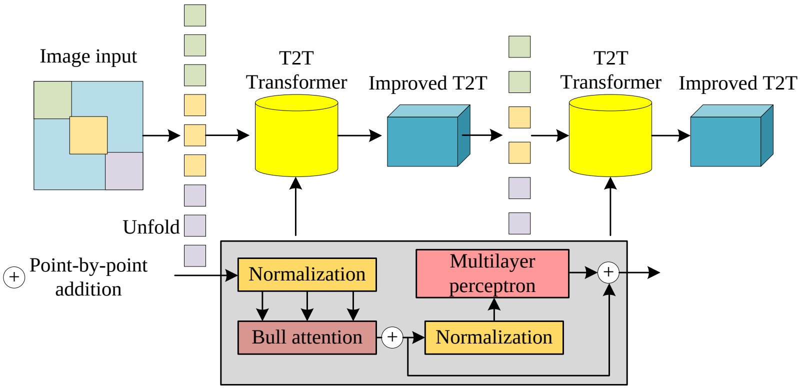


Figure 5: Improved T2T module for PFD at overhead angles

By combining the T2T-ViT module, this study constructs a backbone network with a structure similar to CNN, as shown in Figure 6. It has a similar deep narrow shaped structure as CNN, where S represents the number of stacked modules. The objective of this study is to enhance the learning and recognition performance of the model for pedestrian features under overhead angles. This will be achieved by optimizing the token representation of the images, thus overcoming the problem of similarity in pedestrian targets under this perspective.

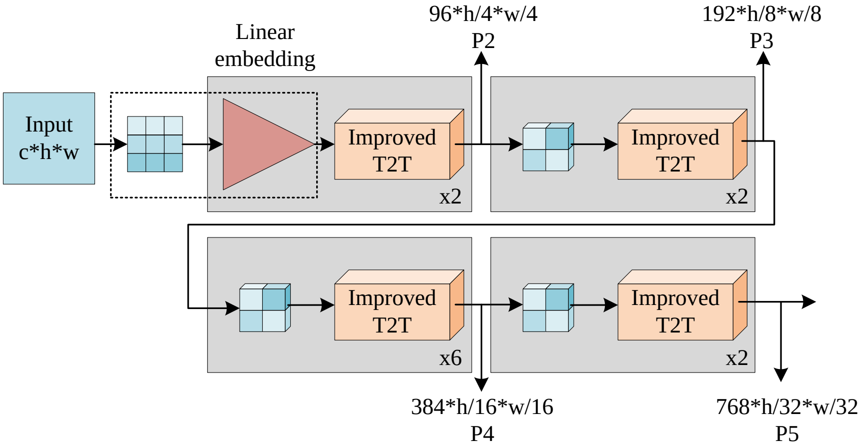


Figure 6: The backbone network structure of overlooking human traffic detection network

In the study, the Deep SORT tracking algorithm is further adopted to achieve accurate tracking of human flow. This algorithm combines Kalman Filtering (KF) [20]. KF essentially achieves optimal prediction of future states through the current historical state. The system state equation of KF is formula (7).

$$X_k = a_{X_{k-1}} + bU_k + W_k \tag{7}$$

In formula (7), X_k and X_{k-1} represent the system state matrices at times k and $k-1$. a and b represent the corresponding system transition matrix. U_k represents the system control matrix at time k. W_k represents the noise impact during the process. The system observation equation is formula (8).

$$Z_k = hX_k + V_k \tag{8}$$

In formula (8), Z_k represents the observation matrix at time k. h represents the system observation matrix. V_k represents the observed noise. For the convenience of calculation, it is assumed that the noise that occurs in the usual process is white noise, which does not change with changes in the system state. Its condition can be expressed as formula (9).

$$\begin{cases} E\left[W_k\right]=0, Cov\left[W_k,W_i\right]=\begin{cases}0, k\neq i\\ Q, k=i\end{cases} \\ E\left[V_k\right]=0, Cov\left[V_k,V_j\right]=\begin{cases}0, k\neq j\\ R, k=j\end{cases}\end{cases} \tag{9}$$

In formula (9), the additional conditions that need to be met are shown in formula (10).

$$Cov\left[W_k,V_k\right]=0 \tag{10}$$

In formulas (10) and (9), Q and R represent the covariance matrices of the corresponding noise, respectively. The state prediction equation of KF is formula (11).

$$X_{k|k-1}=aX_{k-1|k-1}+bU_k \tag{11}$$

In formula (11), $X_{k|k-1}$ represents the system state at time k predicted by time $k-1$. $X_{k-1|k-1}$ represents the optimal prediction system at time $k-1$. The covariance analysis is performed on it as shown in formula (12).

$$P_{k|k-1}=aP_{k-1|k-1}a^T+Q \tag{12}$$

In formula (12), $P_{k|k-1}$ and $P_{k-1|k-1}$ are the covariance matrices corresponding to $X_{k|k-1}$ and $X_{k-1|k-1}$. a^T represents the transposition of the system related state transition matrix. The optimal estimate is expressed as formula (13).

$$X_{k|k}=X_{k|k-1}+K_k\left(Z_k-hX_{k|k-1}\right) \tag{13}$$

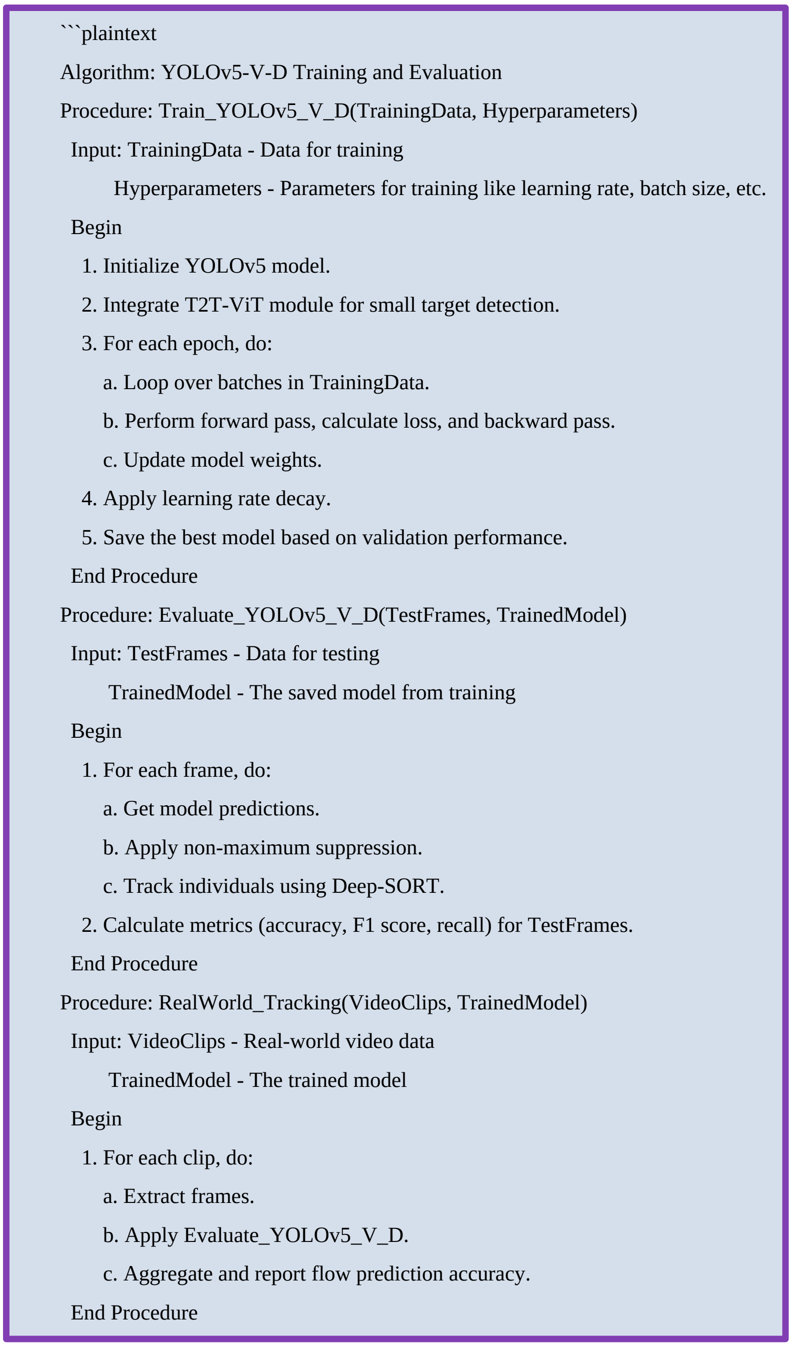
In formula (13), $X_{k|k}$ represents the optimal estimated system state at time k. Z_k represents the system observation state at time k. K_k represents the Kalman gain matrix at time k. h represents the state observation matrix. The calculation of K_k is formula (14).

$$K_k=P_{k|k-1}h^T/\left[hP_{k|k-1}h^T+R\right] \tag{14}$$

In formula (14), R represents the covariance matrix of the noise. The covariance update result of the system state is formula (15).

$$P_{k|k}=\left(1-K_kh\right)P_{k|k-1} \tag{15}$$

In formula (15), $P_{k|k}$ represents the covariance update result of the system state at time k. This study uses YOLOv5 for object detection and combines it with the Deep SORT tracking algorithm to achieve object detection in TP-PFD. The pseudo-code for the research method is shown in Figure 7.



~~~
```plaintext
Algorithm: YOLOv5-V-D Training and Evaluation
Procedure: Train_YOLOv5_V_D(TrainingData, Hyperparameters)
  Input: TrainingData - Data for training
          Hyperparameters - Parameters for training like learning rate, batch size, etc.
  Begin
    1. Initialize YOLOv5 model.
    2. Integrate T2T-ViT module for small target detection.
    3. For each epoch, do:
        a. Loop over batches in TrainingData.
        b. Perform forward pass, calculate loss, and backward pass.
        c. Update model weights.
    4. Apply learning rate decay.
    5. Save the best model based on validation performance.
  End Procedure
Procedure: Evaluate_YOLOv5_V_D(TestFrames, TrainedModel)
  Input: TestFrames - Data for testing
         TrainedModel - The saved model from training
  Begin
    1. For each frame, do:
        a. Get model predictions.
        b. Apply non-maximum suppression.
        c. Track individuals using Deep-SORT.
    2. Calculate metrics (accuracy, F1 score, recall) for TestFrames.
  End Procedure
Procedure: RealWorld_Tracking(VideoClips, TrainedModel)
  Input: VideoClips - Real-world video data
         TrainedModel - The trained model
  Begin
    1. For each clip, do:
        a. Extract frames.
        b. Apply Evaluate_YOLOv5_V_D.
        c. Aggregate and report flow prediction accuracy.
  End Procedure
~~~

Figure 7: Method pseudo-code

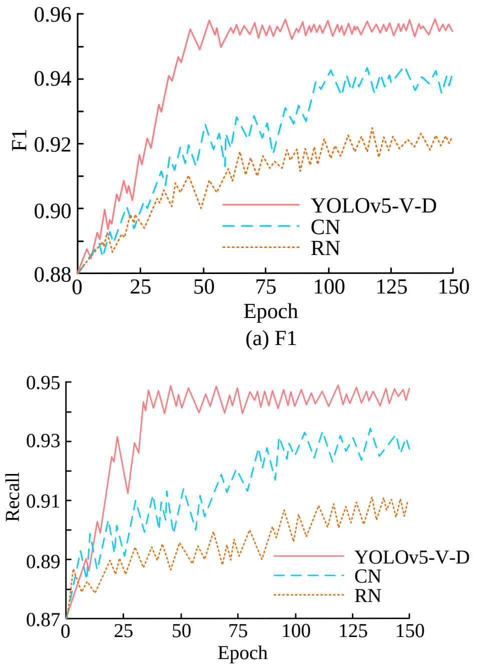
4 Performance testing of TP-PFD model

To test the practicality and usability of the proposed TP-PFD model, this study selects UCF_ CC_ 50 and NWPU Crowd datasets are used to validate this method. Among them, the UCF_ CC_50 dataset contains 50 high-resolution images obtained from different scenes. Each image contains numbers ranging from 94 to 4,543, suitable for population counting studies. NWPU Crowd includes 5,109 images and 2,133,238 person annotations, including some images taken from a top view angle, suitable for crowd counting and positioning. This study randomly selects 80% of the two datasets for training, while the remaining 20% is used for testing. CenterNet (CN) model and RetinaNet (RN) model are selected for comparison with research methods (YOLOv5-ViT-T2T-Deep-SORT, YOLOv5-V-D). To avoid limitations on research due to the device performance, the study chooses to rent a cloud server platform for experimentation. Table 2 provides specific software and hardware details as well as training parameters.

Table 2: Hardware and software details and training parameter settings

Hardware		Software		Training parameter	
Name	Detail	Name	Detail	Hyper-parameters	Detail
Supplier	Microsoft Azure	Linux	Ubuntu Server 20.04 LTS	Learning Rate	0.001
Type	Standard NC6	TensorFlow	2.9	Batch Size	64
CPU	Intel Xeon E5-2690 v3	PyTorch	1.7	Optimizer	Default
		CUDA Toolkit	11.0.194	Weight Initialization	Xavier
RAM	56Gb	cuDNN	7.2.1	Activation Function	Leaky ReLU
GPU	NVIDIA Tesla K80	Python	3.8	Loss Function	Cross-Entropy Loss
MEM	Azure Blob Storage	OpenCV	4.2	Learning Rate Decay	0.1
	Azure Managed Disk	Jupyter Notebook/Lab	-	Epoch	150

Firstly, the F1 and Recall values of the three models are tested, and the results are shown in Figure 8. The improved YOLOv5-V-D has a faster convergence speed, reaching the optimal F1 value and the optimal Recall value in about 61 iterations. In Figure 8 (a), its optimal F1 value is 0.952, which is 0.015 and 0.037 higher than CN and RN, respectively. In Figure 8 (b), its optimal Recall value is 0.947, which is 0.021 and 0.041 higher than CN and RN, respectively.



(a) F1

(b) Recall

Figure 8: F1 and recall tests of three models

Figure 9 shows the results of testing and comparing the Precision recall curves of three models. In Figure 9, the curve area of the research method is superior to the other two models, which proves the superiority of the research method in performance compared to CN and RN. The improved YOLOv5-V-D has better basic performance indicators and stronger practicality.

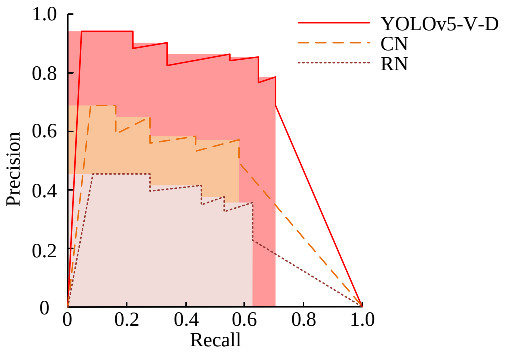
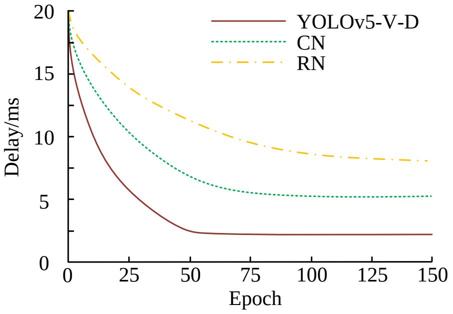
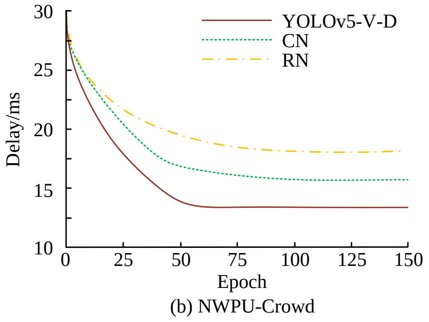


Figure 9: Precision-recall curve tests for three models

The detection delay of three models are tested to examine their usability in practice, testing on two datasets to obtain Figure 10. In Figure 10 (a), the optimal delay performance of YOLOv5-V-D is 2.4ms, leading by 3.5ms and 7.2ms compared to CN and RN, respectively. The performance of the demonstration in Figure 10 (b) has increased, which is speculated to be due to the high difficulty of the dataset. The optimal delay performance of YOLOv5-V-D is 12.0ms, which is 4.3ms and 5.6ms ahead of CN and RN, respectively. Overall, the improved YOLOv5-V-D has better latency performance on both datasets. Its average delay performance is 7.2ms, leading by 3.9ms and 6.4ms compared to CN and RN, respectively.



(a) UCF_CC_50

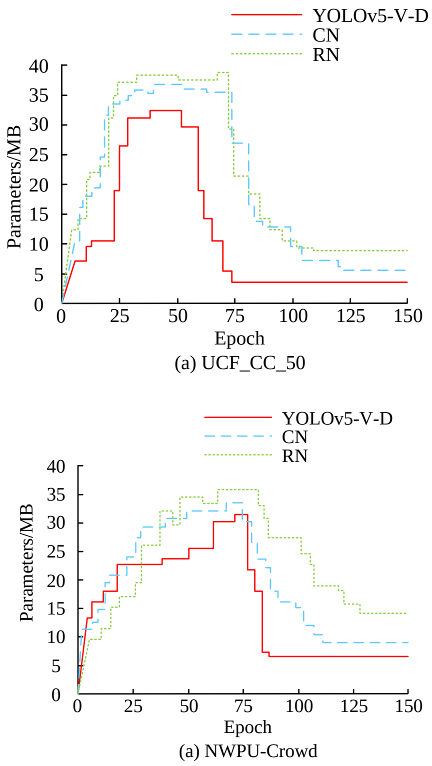


(b) NWPU-Crowd

Figure 10: The delay test results of three algorithms

Figure 11 tests the parameter values of the three models. Figure 11 (a) is the UCF_ CC_50 dataset, and the YOLOv5-V-D performs better than the comparative model. The number of model parameters in Figure 11 (b) has increased due to the increased difficulty of the dataset, but YOLOv5-V-D still performs better than the comparative model. Overall, the improved YOLOv5-V-D has the best parameter performance. It can achieve lower system load and lower computing costs, which is more conducive to embedding and deployment on low performance platforms.

The PFD results of the three algorithms are tested using actual top-down angle shots. To avoid the impact of errors on the test results, this study selects five videos for testing, as shown in Table 3. In Table 3, the improved YOLOv5-V-D has the best accuracy in predicting pedestrian flow. The accuracy of its traffic prediction reaches 98.72%, leading by 20.59% and 28.61%, respectively, compared to CN and RN.



(a) UCF_CC_50

(a) NWPU-Crowd

Figure 11: Parameter test results of three models

Table 3: The actual flow detection test of three models

Video clip	Model	Actual result		Forecast result		Accuracy(%)
		Upflow	Downflow	Upflow	Downflow	
Clip 1	YOLOv5-V-D	9	12	9	12	100.00
	CN			7	11	84.72
	RN			5	9	65.28
Clip 2	YOLOv5-V-D	11	15	11	15	100.00
	CN			14	9	69.29
	RN			11	5	66.67
Clip 3	YOLOv5-V-D	19	14	20	14	97.50
	CN			24	15	86.25
	RN			14	19	73.68
Clip 4	YOLOv5-V-D	21	17	20	17	97.62
	CN			27	21	79.37
	RN			19	29	75.55
Clip 5	YOLOv5-V-D	29	33	29	32	98.48
	CN			21	23	71.06
	RN			46	25	69.40

In summary, the improved YOLOv5-V-D has excellent training efficiency and superior performance compared to both CN and RN models. It has lower latency and parameter quantity, and it has higher accuracy in actual PFD. Using only the UCF_CC_50 and NWPU-Crowd datasets for testing may result in a lack of clear understanding of operational performance in real-world operating environments, leading to unclear limitations in the study. To more fully evaluate the generalization ability of the model, ShanghaiTech and Qnrf datasets containing extreme crowd density and complex dynamic behavior are introduced for testing. On the ShanghaiTech dataset, the model's detection accuracy exhibits a minimal decline of 0.39%, particularly in scenarios involving high passenger density, where its performance remains robust. Data enhancement techniques are then used to simulate a variety of environmental conditions, including random brightness adjustment (brightness variation range ±20%), contrast change (contrast variation range ±15%), noise addition (Gaussian noise and salt and pepper noise), and image transformation to simulate different viewing angles. Under more environmental conditions, the fluctuation range of detection accuracy of the research method is kept within 1.50%. In severe weather conditions including rain, fog, snow and dust, the detection accuracy of the research method has a decrease of nearly 3.00%. To assess the stability of the model over long tracking periods, the study is tested in a continuous video stream. The results show that the YOLOv5-V-D model has a tracking accuracy of more than 95% in the video tracking of up to 2 hours, which proves its long-term stability. It shows that the method has good model generalization performance.

5 Discussion

This paper proposes a PFD method based on top-down view of STT. By integrating ViT architecture and Deep-SORT tracking algorithm, the YOLOv5 model is improved, aiming to improve the accuracy and efficiency of small-target tracking under top-down view. The improved YOLOv5-V-D model outperforms the existing CenterNet and RetinaNet models on several performance metrics. Specifically, the values of F1 and Recall of YOLOv5-V-D reach 0.952 and 0.947, respectively, which are significantly improved compared to CenterNet and RetinaNet models. In addition, the average delay of YOLOv5-V-D is 7.2ms, which also shows better performance in real-time than other existing studies. The introduction of the T2T-ViT architecture has enabled the effective capture of global information in images through SAM. This has facilitated the distinction between the flow of people and the background in complex scenes, thereby enhancing the recognition accuracy of small targets. The integration of T2T-ViT architecture is one of the core innovations of the study. The T2T-ViT model employs a token-to-token mechanism to recursively merge adjacent image blocks, thereby constructing a hierarchical representation. This approach not only enhances the model's capacity to identify subtle targets and intricate textures but also reduces its complexity and computational cost. The integration of the T2T-ViT architecture significantly improves the model's performance when dealing with complex pedestrian dynamics. In the actual overhead angle shot, the YOLOv5-V-D model shows a high pedestrian flow prediction accuracy of 98.72%, which indicates that the model can effectively handle the interaction and occlusion between pedestrians, as well as the challenges brought by dynamic lighting changes. Through experiments on various test sets, it is found that the performance of the model deteriorates in high density human flow scenarios, poor lighting conditions or severe lighting changes. Analysis shows that the main reasons for performance degradation include but are not limited to insufficient representation of small targets in images resulting in insufficient feature extraction, as well as interference factors in complex backgrounds. The stability and accuracy of tracking algorithms are challenged in situations of rapid motion or occlusion. The integration of infrared or depth sensor data can be considered in the solution to supplement the visual information and enhance the robustness of the model to occlusion and illumination changes. More sophisticated image enhancement techniques such as adaptive contrast adjustment and noise reduction algorithms are employed to improve image quality.

6 Conclusion

The detection and tracking of pedestrian flow is of great significance for public safety and space management. The PFT under the overhead angle requires extremely high STD capability of the model. This study aimed to improve the accuracy and efficiency of detection methods to achieve real-time monitoring in various application scenarios. By adopting the improved YOLOv5-V-D model, this method has made significant progress in target tracking and PFD performance. In response to the shortcomings of classical models in small target recognition, the ViT-T2T framework and Deep SORT tracking algorithm were introduced to enhance the sensitivity and detection speed of the model to small targets. After improvement, the model converged after approximately 61 iterations, demonstrating excellent performance. The F1 value and Recall value reached 0.952 and 0.947, respectively, surpassing the performance of the CN and RN models. In addition, the model also performed excellently in terms of latency, with an average time of only 7.2ms, significantly superior to the comparison model. The practical application value of this research result was reflected in the accuracy of pedestrian flow prediction, reaching 98.72%, significantly higher than the CN and RN models. This achievement confirmed the effectiveness of the improved model in handling complex dynamic scenes, especially its potential application in high-density pedestrian environments. Despite the above achievements, the limitations of this method in facing extreme environments and behavior patterns should also be recognized. This study lacks further improvements in the robustness of the model, and future improvements should be made in this area to enhance the applicability and anti-interference ability of the model.

References

[1] X. Xu, Q. Wu, L. Qi, W. Dou, S. B. Tsai, and M. Z. A. Bhuiyan, “Trust-aware service offloading for video surveillance in edge computing enabled internet of vehicles,” IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 3, pp. 1787-1796, 2021. https://doi.org/10.1109/TITS.2020.2995622

[2] Y. Zhang, J. Zhang, and R. Tao, “Key frame extraction of surveillance video based on fractional fourier transform,” Journal of Beijing Institute of Technology, vol. 30, no. 3, pp. 311-321, 2021. https://doi.org/10.15918/j.jbit1004-0579.2021.058

[3] J. Qiu, L. Wang, Y. Hu, and Y. Wang, “Effective object proposals: size prediction for pedestrian detection in surveillance videos,” Electronics Letters, vol. 56, no. 14, pp. 706-709, 2020. https://doi.org/10.1049/el.2020.0850

[4] D. Xi, Y. Qin, and S. Wang, “YDRSNet: an integrated Yolov5-Deeplabv3+real-time segmentation network for gear pitting measurement,” Journal of Intelligent Manufacturing, vol. 34, no. 4, pp. 1585-1599, 2023. https://doi.org/10.1007/s10845-021-01876-y

[5] Z. Zhao, F. N. Khan, Z. A. H. Qasem, B. Deng, Q. Li, Z. Liu, and H. Y. Fu, “Convolutional-neural-network-based versus vision-transformer-based SNR estimation for visible light communication networks,” Optics Letters, vol. 48, no. 6, pp. 1419-1422, 2023. https://doi.org/10.1364/OL.485321

[6] S. Shi, L. Jiang, D. Cao, and Y. Zhang, “Sea-surface small target detection using entropy features with dual-domain clutter suppression,” Remote Sensing Letters, vol. 13, no. 10/12, pp. 1142-1152, 2022. https://doi.org/10.1080/2150704X.2022.2127129

[7] B. Yang, and J. Wang, “An improved helmet detection algorithm based on YOLO V4,” International Journal of Foundations of Computer Science, vol. 33, no. 6/7, pp. 887-902, 2022. https://doi.org/10.1142/s0129054122420205

[8] G. L. Zhi, D. U. Juan, T. Feng, and Z .W. Jia, “Traffic sign recognition using an attentive context region-based detection framework,” Chinese Journal of Electronics, vol. 30, no. 6, pp. 1080-1086, 2021. https://doi.org/10.1049/cje.2021.08.005

[9] L. Bommes, M. Hoffmann, C. Buerhop-Lutz, T. Pickel, J. Hauch, and C. Brabec, “Anomaly detection in IR images of PV modules using supervised contrastive learning,” Progress in Photovoltaics, vol. 30, no. 6, pp. 597-614, 2022. https://doi.org/10.1002/pip.3518

[10] H. Qin, Y. Wu, F. Dong, and S. Sun, “Dense sampling and detail enhancement network: Improved small object detection based on dense sampling and detail enhancement,” IET Computer Vision, vol. 16, no. 4, pp. 307-31, 2022. https://doi.org/10.1049/cvi2.12089

[11] Y. Minegishi, Y. Ohmiya, T. Sano, and M. Tange, “Analysis and modeling of pedestrian flow in a confined corridor focusing on the headway distance and velocity of pedestrians,” Fire Technology, vol. 58, no. 2, pp. 709-735, 2022. https://doi.org/10.1007/s10694-021-01173-3

[12] J. Yang, W. Y. He, T. Zhang, C. Zhang, and B. F. Nan, “Research on subway pedestrian detection algorithms based on SSD model,” IET Intelligent Transport Systems, vol. 14, no. 11, pp. 1491-1496, 2020. https://doi.org/10.1049/iet-its.2019.0806

[13] X. Song, B. Chen, P. Li, B. Wang, and H. Zhang, “PRNet++: learning towards generalized occluded pedestrian detection via progressive refinement network,” Neurocomputing, vol. 482, no. 14, pp. 98-115, 2022. https://doi.org/10.1016/j.neucom.2022.01.056

[14] J. Yang, W. Y. He, T. Zhang, C. L. Zhang, L. Zeng, and B. F. Nan, “Research on subway pedestrian detection algorithms based on SSD model,” IET Intelligent Transport Systems, vol. 14, no. 11, pp. 1491-1496, 2020. https://doi.org/10.1049/iet-its.2019.0806

[15] S. Zhang, X. Yang, Y. Liu, and C. Xu, “Asymmetric multi-stage CNNs for small-scale pedestrian detection,” Neurocomputing, vol. 409, no. 7, pp. 12-26, 2020. https://doi.org/10.1016/j.neucom.2020.05.019

[16] A. Ali, “A framework for air pollution monitoring in smart cities by using IoT and smart sensors,” Informatica, vol. 46, no. 5, pp. 129-138, 2022. https://doi.org/10.31449/inf.v46i5.4003

[17] J. Guo, X. Zhang, Y. Dong, Z. Xue, and B. Huang, “Terrain classification using mars raw images based on deep learning algorithms with application to wheeled planetary rovers,” Journal of terramechanics, vol. 108, no. 8, pp. 33-38, 2023. https://doi.org/10.1016/j.jterra.2023.04.002

[18] Q. Zhang, Y. Wang, L. Song, M. Han, and H. Song, “Using an improved YOLOv5s network for the automatic detection of silicon on wheat straw epidermis of micrographs,” Journal of Field Robotics, vol. 40, no. 1, pp. 130-143, 2023. https://doi.org/10.1002/rob.22120

[19] A. Ali, “Remote monitoring of lab experiments to enhance collaboration between universities,” Informatica, vol. 46, no. 2, pp. 169-177, 2022. https://doi.org/10.31449/inf.v46ix.xxxx

[20] S. Yang, Z. Chen, X. Ma, X. Zong, and Z. Feng, “Real-time high-precision pedestrian tracking: a detection-tracking-correction strategy based on improved SSD and Cascade R-CNN,” Journal of Real-Time Image Processing, vol. 19, no. 2, pp. 287-302, 2022. https://doi.org/10.1007/s11554-021-01183-y