IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 December 2023, accepted 2 February 2024, date of publication 13 February 2024, date of current version 21 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365501

RESEARCH ARTICLE

Pedestrian Tracking Algorithm for Video Surveillance Based on Lightweight Convolutional Neural Network

HONGLEI WEI, XIANYI ZHAI, AND HONGDA WU
School of Mechanical Engineering and Automation, Dalian Polytechnic University, Dalian, Liaoning 116034, China

Corresponding author: Honglei Wei (weihl2005@163.com)

This work was supported in part by the Liaoning Provincial Department of Education 2021 Annual Scientific Research Funding Program under Grant LJKZ0535 and Grant LJKZ0526; in part by the 2021 Annual Comprehensive Reform of Undergraduate Education Teaching, Dalian Polytechnic University, under Grant JGLX2021020 and Grant JCLX2021008; and in part by the Graduate Innovation Fund, Dalian Polytechnic University, under Grant 2023CXYJ13.

ABSTRACT The Efficient Convolution Operators for Tracking (ECO) algorithm has garnered considerable attention in both academic research and practical applications due to its remarkable tracking efficacy, yielding exceptional accuracy and success rates in various challenging contexts. However, the ECO algorithm heavily relies on the deep learning Visual Geometry Group (VGG) network model, which entails complexity and substantial computational resources. Moreover, its performance tends to deteriorate in scenarios involving target occlusion, background clutter, and similar challenges. To tackle these issues, this study introduces a novel enhancement to the pedestrian tracking algorithm. Specifically, the VGG network is substituted with a lightweight MobileNet v2 model, thereby reducing computational demands. Additionally, a Double Attention Networks (A2-Net) module is incorporated to augment the extraction of crucial information, while pre-training techniques are integrated to expedite model convergence. Experimental results demonstrate that the C-ECO algorithm achieves comparable accuracy and success rates to the conventional ECO algorithm, despite reducing the model size by 27.96% and increasing the tracking frame rate by 46.11%. Notably, when compared to other prevalent tracking algorithms, the C-ECO algorithm exhibits an accuracy of 82.20% and a success rate of 64.72%. These findings underscore the enhanced adaptability of the C-ECO algorithm in complex environments, offering a more lightweight model while delivering superior tracking capabilities.

INDEX TERMS Machine vision, target tracking, deep learning, efficient convolution operator, pedestrian tracking.

I. INTRODUCTION

Target tracking is an important branch of machine vision and is the technical basis for intelligent visual surveillance, visual behavior analysis, and human-computer interaction [1]. Target tracking refers to the technique of predicting the location of a target in each frame of a subsequent video sequence based on the given target location in the input video sequence [2]. Pedestrian detection and tracking is a research hotspot in the field of computer vision, which can be applied to traffic monitoring, video surveillance, security, and other fields, and has certain application value and challenges. There has been a remarkable upsurge of interest in automated crowd monitoring within the computer vision community. Modern deep learning techniques have enabled the development of fully automated crowd monitoring applications based on visual analysis. Even with the magnitude of the issue, the substantial technological progress, and the unwavering interest from the research community, there are still several challenges that demand attention [3], [4].

2010 CVPR, Bolme et al. [5] first applied correlation filtering to the field of tracking, and based on his idea,

The associate editor coordinating the review of this manuscript and approving it for publication was Xuebo Zhang.

algorithms using correlation filtering for target tracking have appeared one after another, and the tracking effect has the tracking results are getting better and better. Moridvaisi et al. [6] surmount KCF's limitations through the lens of the Tracking-Learning-Detection (TLD) framework and devised an algorithm that concurrently trains two classifiers, employing a semi-supervised co-training learning algorithm. Subsequently, they subject the proposed method to rigorous scrutiny against TB-100 datasets, juxtaposed with its counterparts. Yang et al. [7] used KCF-based SOT to learn discriminative target appearance that relied on hand-crafted deep features and used the prediction results to refine detection errors in new ways and eliminated tracking errors caused by uncorrelated algorithms. Sanagavarapu and Pullakandam [8] proposed the method using the Kernelized Correlation Filter (KCF) object tracking technique. The segmented region is encoded by the complexity-efficient Scalable HEVC (SHVC) to meet the resolution of an end-user device. The complexity of SHVC is decreased by using the Convolutional Neural Network (CNN) and Long- and Short-Term Memory (LSTM) to predict the Coding Tree Unit (CTU) structure. The results show that the proposed method decreases the bitrate significantly for video sequences without degradation in Peak Signal-to-Noise Ratio (PSNR). A tracking method that integrates the objectness-bounding box regression (O-BBR) model and a scheme based on kernelized correlation filter (KCF) is proposed by Mbelwa et al. [9]. The scheme based on KCF is used to improve the tracking performance of FM and MB. For handling drift problems caused by OCC and IV, we propose objectness proposals trained in bounding box regression as prior knowledge to provide candidates and background suppression. Finally, scheme KCF as a base tracker and O-BBR are fused to obtain the state of a target object. Khan et al. [10] proposed a new criterion based on the hybridization of multiple cues i.e., average peak correlation energy (APCE) and confidence of squared response map (CSRM), which is presented to enhance the tracking efficiency. They updated the occlusion detection module adaptive learning rate adjustment module, and drift handling using an adaptive learning rate model based on hybridized criterion, and integrated all these modules to propose a new tracking scheme. Degli-Esposti et al. [11] proposed a new algorithm for object tracking in SWIR imaging, using a kernelized correlation filter (KCF) as a basic tracker. To overcome occlusions, they proposed the use of the Kalman filter as a predictor and a method to expand the object search area. To cope with outliers, Huber's M-robust approach is applied, so this paper proposes robustification of the Kalman filter by introducing a nonlinear Huber's influence function in the Kalman filter estimation step. To make a balance between desired estimator efficiency and resistance to outliers, a new adaptive M-robustified Kalman filter is proposed. This is achieved by adjusting the saturation threshold of the influence function using the detection confidence information from the basic KCF tracker. Liang et al. [12] proposed Spatio-Temporal adaptive and Channel selective Correlation Filters (STCCF) for robust tracking, selecting a set of target-specific features from high dimensional features, STCCF can not only alleviate the over-fitting problem and reduce the computational cost, but also enhance the discriminability and interpretability of the learned filters.

With the rapid development of deep learning in recent years, many scholars use deep learning networks to extract image features and fuse them with relevant filters to perform target tracking.

Zdarsky et al. [13] introduced a deep learning-based approach that uses the video frames of low-cost web cameras. Using DeepLabCut (DLC), an open-source toolbox for extracting points of interest from videos, they obtained facial landmarks critical to gaze location and estimated the point of gaze on a computer screen via a shallow neural network. Tested for three extreme poses, this architecture reached a median error of about one degree of visual angle. Abdelali et al. [14] introduced a wholly automated methodology for Multiple Hypothesis Detection and Tracking (MHDT) in the domain of video traffic surveillance. The presented framework integrates the Kalman filter with data association-based tracking techniques, employing the YOLO detection method, to adeptly monitor vehicles in intricate traffic surveillance scenarios. Empirical findings substantiate that the proposed approach exhibits resilience in discerning and tracing the trajectories of vehicles under diverse circumstances, including scale variations, stationary vehicles, rotations, fluctuating lighting conditions, and instances of occlusion. Zhang et al. [15] introduced a pioneering approach known as Harris Hawks Optimization with deep learning-enhanced automated face detection and tracking (HHODL-AFDT). The HHODL-AFDT model, as proposed, incorporates a Faster Region-Based Convolutional Neural Network (RCNN) for face detection and leverages the Harris Hawks Optimization (HHO) for hyperparameter optimization. The optimized Faster RCNN model presented in this context impeccably discerns facial features and feeds this information into the face-tracking model through a regression network (REGN). The face tracking, facilitated by the REGN model, makes use of features extracted from adjacent frames to anticipate the facial target's location in subsequent frames. Almuqren et al. [16] presented an effective method to track an object based on a combination of feature hierarchies of CNNs, they combined several feature hierarchies and compute the more discriminative map to track the object, a novel method of feature hierarchies integration based on Kullback–Leibler (KL) divergence is adopted. Ahmed et al. [17] unveiled an intricate multi-person tracking framework, thoughtfully intertwined with 5G infrastructure. Employing a top-view perspective, this framework yields an expansive scope of the observed scene or field of vision. The essence of person tracking is encapsulated within a deep learning-driven tracking-by-detection framework, wherein detection duties are seamlessly executed by the YOLOv3

model, and the subsequent tracking operations are orchestrated by the Deep SORT algorithm. To further elevate the precision of the detection model, a transfer learning approach is artfully employed. In this methodology, a detection model capitalizes on a pre-trained foundation, enriched with an additional layer meticulously fine-tuned using a top-view dataset. Zhang et al. [18] proposed a robust adaptive learning visual tracking algorithm, HOG features, CN features, and deep convolution features are extracted from the template frame and search region frame, respectively, and analyzed the merits of each feature and perform feature adaptive fusion to improve the validity of feature representation.

Although target tracking algorithms have been developed over many years, the current algorithms still face challenges in accurately tracking targets that experience occlusion, background clutter, or leave the field of view. Additionally, the deep learning network models, while highly effective, are complex and have a large number of parameters. Consequently, they require substantial computational resources and place higher demands on computer hardware.

To address the above problems, this study proposes a lightweight convolutional neural network-based C-ECO tracking algorithm [19] based on the ECO tracking algorithm based on deep learning and correlation filtering. The ECO algorithm has two implementation forms, ECO based on convolutional features and ECO_HC based on artificial features [20]. Combining the accuracy and speed considerations, this experiment chooses the ECO algorithm based on convolutional features for optimization and improvement. The main contributions of this study are as follows:

(1) In response to the complex deep learning VGG network model in the ECO algorithm, which occupies large computational resources, a lightweight MobileNet v2 is used instead to perform feature information extraction, which effectively reduces resource consumption and improves tracking speed.

(2) In order to improve the target feature extraction ability of the convolutional network, the A2-Net module is added to MobileNet v2, which effectively improves the extraction effect of the network on important information with a small increase of computing parameters, and significantly improves the training efficiency and tracking accuracy.

(3) Introducing the pre-training model in the training stage effectively accelerates the model convergence speed, significantly shortens the training time, and effectively improves the accuracy and success rate of the tracking algorithm.

The remaining chapters of this paper are organized as follows: in Section I, the basics are introduced, in Section II, the construction of the C-ECO algorithm is introduced, in Section III, the model is subjected to ablation experiments, comparison tests, and in Section IV, the whole paper is summarized and in Section V, an outlook on future work is given.

II. FOUNDATIONAL THEORIES AND PROPOSED METHOD

A. THE OBJECT TRACKING ECO ALGORITHM

The ECO target tracking algorithm is improved from the continuous convolutional tracking algorithm C-COT [21]. The algorithm finally achieves target localization and filter update by applying the convolutional features of the target in the input image video, the directional gradient histogram feature HOG (histogram of gradients) and the color channel feature CN (color-names) [22], [23].

The ECO algorithm mainly includes the processes of feature extraction, continuous convolution operation, convolution operation of factorization, generation of sample space model and correlation filtering operation [24]. First, the interpolation operation is performed for the features x in the search region of the target to be detected as shown in Equation (1) [25].

$$J_d\left\{x^d\right\}(t)=\sum_{n=0}^{N_d-1}x^d[n]\,b_d\left(t-\frac{T}{N_d}n\right) \tag{1}$$

where: x^d denotes the d-channel characteristic of x, $J_d\left\{x^d\right\}(t)$ is a function on $t\in[0,T)$ that represents the result of the interpolation operation of x^d, $x^d[n]\in\mathbb{R}^{Nd}$ is a function with respect to $n\in\{0,\cdots,N_d-1\}$, Nd indicates resolution, $b_d\left(t-\frac{T}{N_d}n\right)$ denotes the d-channel interpolation function. The interpolation results for channels 1 to D are denoted by $J\{x\}(t)\in\mathbb{R}^D$, abbreviated as $J\{x\}$. After that, the filter is simplified using principal component analysis and the response score $S_{Pf}\{x\}$ obtained by convolving with $J\{x\}$ is shown in equation (2) [26].

$$S_{Pf}\{x\}=Pf*J\{x\}=f*P^TJ\{x\} \tag{2}$$

where: f denotes the filter with channel number D, denotes the convolution calculation, P is the projection matrix of D rows and C columns, and P^T denotes its transpose matrix. The position of the score maximum, i.e., the new position of the target, is obtained by optimizing $S_{Pf}\{x\}$ using the Gaussian Newton algorithm. Finally, the data set is compressed using a Gaussian mixture model, and the error of the convolutional response score $S_{Pf}\{x\}$ of the training sample and the current filter f with the Gaussian label y_0 of the training sample is taken as L_2 parametric, and the penalty term is added to obtain the loss function as shown in equation (3).

$$E(f)=\sum_{m=1}^{M}\pi_m\left\|S_{Pf}\{\mu_m\}-y_0\right\|_{L^2}^2+\sum_{c=1}^{C}\left\|\omega f^c\right\|_{L^2}^2 \tag{3}$$

where: μ_m and π_m are the mean and weight of the training samples, respectively; M is the total number of training samples; ω is the penalty term of f. P is only calculated in the first frame and is kept constant when f is updated using the conjugate gradient algorithm to re-solve (3) every 6 frames thereafter [27].

In summary, ECO takes three ways to improve by reducing the filter, optimizing the training set and reducing the filter update frequency, which effectively improves the tracking speed.

B. LIGHTWEIGHT NETWORK MobileNet v2

The ECO algorithm uses convolutional networks of VGG19 and ResNet50, which have deeper networks, better feature

TABLE 1. MobileNet v2 network structure.

Input	Operator	t	c	n	s
224×224×3	Conv2d	-	32	1	2
112×112×32	Bottleneck	1	16	1	1
112×112×16	Bottleneck	6	24	2	2
56×56×24	Bottleneck	6	32	3	2
28×28×32	Bottleneck	6	64	4	1
28×28×64	Bottleneck	6	96	3	2
14×14×96	Bottleneck	6	160	3	2
7×7×160	Bottleneck	6	320	1	1
7×7×320	Conv2d 1×1	-	1280	1	1
7×7×1280	Avg pool 7×7	-	-	1	-
1×1×k	Conv2d 1×1	-	k	-	

extraction, and higher tracking accuracy, but the overly complex networks and a huge number of parameters take up a lot of computational resources and require higher hardware, which leads to an increase in computational cost [28]. The target tracking task requires high speed, so it is necessary to build a lightweight convolutional network model to reduce the model size and improve the detection speed while guaranteeing accuracy. MobileNet has a simple streamlined structure with the advantages of a small number of parameters and low latency. MobileNet network structure is shown in Table 1, where $t$is the expansion factor, c is the number of channels, n is the block number, and s is the step size [29], [30].

1) DEPTHWISE SEPARABLE CONVOLUTION

MobileNet v2 is mainly composed of depth separable convolution (DSC), the standard convolution operation is split into a Depthwise convolution (DW) and a pointwise convolution (PW) [31]. The comparison of the convolution is shown in Figure 1. For the feature map obtained by Depthwise convolution, a 1×1 convolution kernel is used in the point-by-point convolution to perform the convolution operation, and the final output feature layer after point-by-point convolution has the same dimension as the standard convolution [32].

2) CONTRAST BETWEEN DEPTHWISE SEPARABLE CONVOLUTION AND TRADITIONAL CONVOLUTIONAL NETWORK

The number of parameters and the amount of computation of Depthwise Separable convolution is compared with the standard convolution to get the ratio of the number of parameters (4) and the ratio of the amount of computation (5). Generally speaking, N is larger, $1/N$ is negligible, and D_K indicates the size of the convolution kernel. The number of parameters and computation of Depthwise Separable convolution is reduced to about $1/D_K^2$ of the original one, and if the common 3×3 convolution kernel is used, it can be reduced to about 1/9 of the original one [33], [34]. It can be seen that Depthwise Separable convolution significantly reduces the number of operations and parameters, which can effectively reduce the complexity of the network and improve the speed of target tracking.

$$\frac{D_K \times D_K \times M + M \times N}{D_K \times D_K \times M \times N} = \frac{1}{N} + \frac{1}{D_K^2} \tag{4}$$

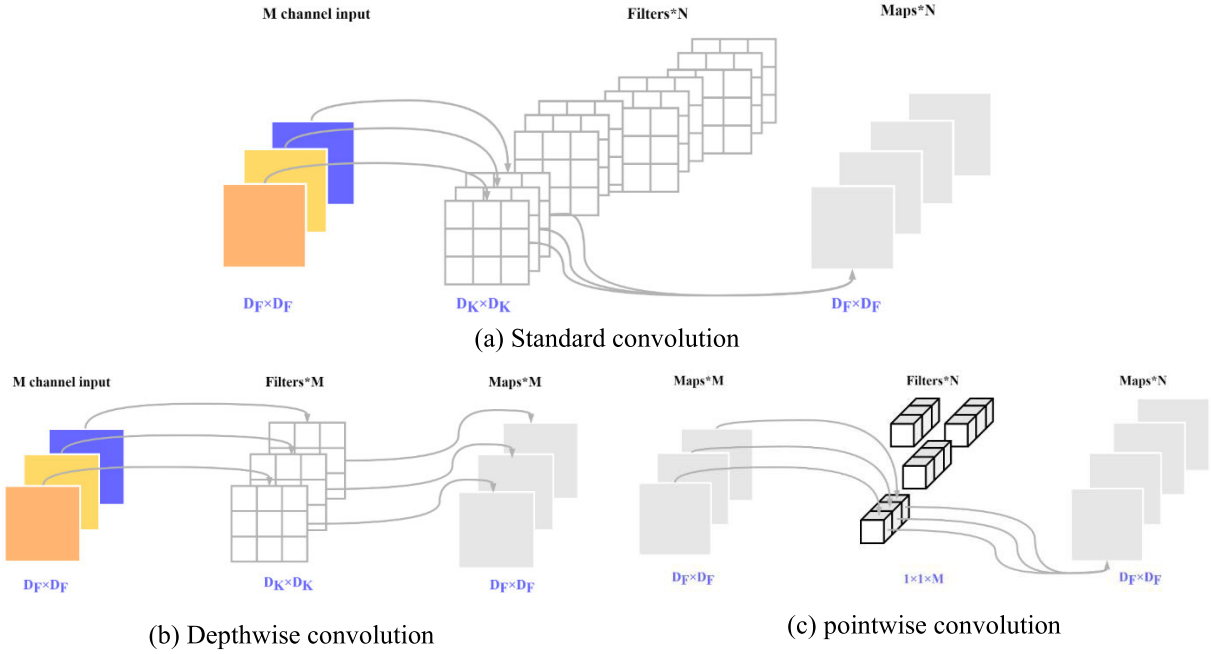
$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{5}$$

3) A2-NET ATTENTION MODULE

During the process of network training, as the volume of information to be acquired grows, the complexity of the model also tends to rise. Consequently, this heightened complexity necessitates increased computational capacity from the hardware on which the model is deployed. The attention mechanism plays a pivotal role in this context by sieving and selecting a small fraction of significant information from a substantial volume of data. By concentrating predominantly on this essential information, the attention mechanism effectively disregards the majority of relatively unimportant data [35].

The fundamental concept of A2-Net revolves around gathering the pivotal features of the entire space into a concise set, followed by an adaptive distribution to each location. This enables subsequent convolutional layers to sense the features of the entire space even without an extensive receptive field. The A2-Net module introduces a dual attention block specifically designed to efficiently capture and distribute long-distance features. This architectural design showcases its potential for enhancing image and video recognition performance, as it effectively models quadratic feature statistics and adapts feature assignments.

The central premise of the A2-Net module involves two primary steps. Initially, it gathers crucial features from the complete space and condenses them into a concise set. Subsequently, these pivotal features are adaptively distributed



(a) Standard convolution

(b) Depthwise convolution

(c) pointwise convolution

FIGURE 1. Comparison of standard convolution and Depthwise Separable Convolution. Figure 1(a) shows the standard convolution, Figure 1(b) shows the Depthwise convolution and Figure 1(c) shows the pointwise convolution.

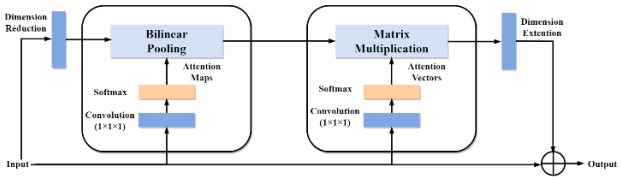


FIGURE 2. A2-Net module structure.

TABLE 2. Performance comparison of A2-net on imagenet-1K dataset.

Module	Backbone	Top1-acc	Top5-acc
ResNet	ResNet-50	75.3%	92.2%
	ResNet-152	77.0%	93.3%
SENet	ResNet-50	76.7%	93.4%
A2-Net	ResNet-50	77.0%	93.5%

to each location, allowing subsequent convolutional layers to perceive the features of the overall space without necessitating an extensive receptive field. The first-level attention operation within the A2-Net selectively gathers crucial features from the complete space, ensuring the integration of vital information. Meanwhile, the second-level attention operation employs an additional attention mechanism to dynamically allocate subsets of pivotal features to complement each specific spatio-temporal location within the higher-level task [36]. For a visual representation of the A2-Net module's structure, refer to Figure 2, which depicts the module's components and architecture.

A2-Net shares some similarities with SENet, covariance pooling, Non-local, and Transformer. However, it sets itself apart through its first attention operation, which implicitly calculates second-order statistics of pooled features. This unique approach allows A2-Net to capture intricate appearance and motion correlations that elude global average pooling, a technique employed in SENet. Additionally, the second attention operation in A2-Net dynamically allocates features from a concise collection, providing a more efficient alternative to the exhaustive feature correlation utilized by Non-local and Transformer, which examines correlations between features from all locations and specific positions.

Table 2 shows the performance comparison between A2-Net and the famous attention network SENet using ImageNet-1K as the dataset. The commonly used metrics Top1-acc and Top5-acc are selected for performance comparison. Top-1 Accuracy refers to the Accuracy that the top-ranked category matches the actual results, and Top-5 accuracy refers to the accuracy that the top-five categories contain the actual results.

It can be seen from Table 1 that in terms of two indicators Top1-acc and Top5-acc, the prediction effect of using A2-Net is improved compared with that on SENet and ResNet.

4) CONSTRUCTION OF C-ECO TRACKING ALGORITHM

The C-ECO pedestrian tracking algorithm proposed in this study is based on the ECO tracking algorithm. In order to increase the ability of the network to extract feature information and improve the accuracy of target detection, the A2-Net module is introduced into MobileNet v2. By adding A2-Net module after the second, fourth and sixth layer of bottleneck, the final C-MobileNet is obtained by replacing the bottleneck in the original network, and the improved

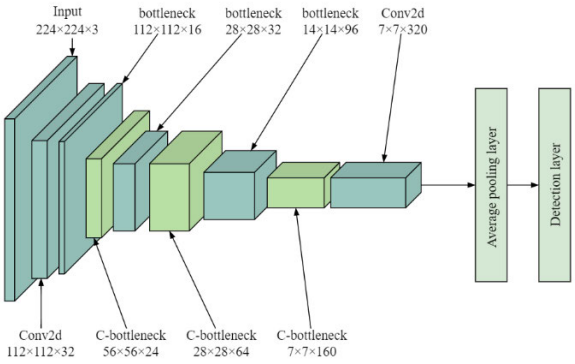


FIGURE 3. C-MobileNet network structure.

C-MobileNet network structure is shown in Figure 3. The addition of three A2-Net modules only adds fewer training parameters and operations, but brings a great improvement in the ability to extract important information in the feature map.

The ECO algorithm is enhanced by incorporating C-MobileNet, which replaces the VGG network utilized in the original algorithm. This modification leads to the creation of the C-ECO algorithm, which is based on a lightweight convolutional neural network. In this paper, the C-ECO algorithm is designed to improve upon the limitations of the VGG19 network, characterized by its complex structure and redundant parameters. Instead, the lightweight MobileNet v2 is employed for efficient feature extraction. Additionally, the A2-Net attention module is introduced to enhance recognition performance. By balancing feature extraction capability and network complexity, the C-ECO algorithm achieves a lightweight architecture. Ultimately, the improved C-ECO algorithm is utilized for precise target tracking. Figure 4 shows the flow of the improved C-ECO algorithm.

III. EXPERIMENT AND ANALYSIS

1) EXPERIMENTAL ENVIRONMENT AND DATASET

The GPU used in this experiment is NVIDIA GeForceRTX3060 with 6G of memory; the CPU is Intel Core i7-12700H with 2.70GHz and 32GB of RAM; the OS is Windows 11, the programming environment is python3.9, the programming software is PyCharm2021.3.1, and the CUDA version is 11.6. In terms of experimental parameter setting, the initial learning rate was set as 0.001, and the epoch was set as 200.

In order to evaluate the tracking algorithm performance and show the tracking effect, two data sets, OTB-50 and OTB-100, containing 50 and 100 videos respectively, are selected for testing.

2) ABLATION EXPERIMENTS

In order to verify the feasibility of applying the A2-Net module in MobileNet, ablation experiments are conducted. The ECO algorithm was introduced for comparison with ECO algorithm combined with MobileNet and C-ECO algorithm, respectively, using the dataset OTB-100, and four metrics were used to measure model strengths and weaknesses, namely model size, accuracy, success rate, and FPS.

(1) Accuracy rate: the percentage of video frames in which the distance between the center point of the target location (bounding box) estimated by the tracking algorithm and the center point of the manually labeled (ground-truth) target is less than a given threshold.

(2) Success rate: define the overlap score (OS), the bounding box obtained by the tracking algorithm (denoted as a), and the box obtained by ground-truth (denoted as b), the overlap rate is defined as $OS = |a \cap b| / |a \cup b|$, $|\cdot|$ indicates the number of pixels in the region. When the OS of a frame is greater than the set threshold, the frame is considered as Success and the percentage of successful frames to all frames is the Success rate.

(3) FPS: The number of frames per second that the tracking algorithm processes the image.

The experimental results are shown in Table 3.

According to the data presented in Table 3, when MobileNet v2 is used for feature extraction, the algorithm model size is reduced by 30.29% compared to the VGG network. Although there is a slight decrease in accuracy and success rate, the FPS (frames per second) shows some improvement. Furthermore, with the inclusion of the A2-Net module, the model size increases slightly but remains 27.96% smaller than the VGG network model. The accuracy and success rates are also comparable to the VGG network, differing by less than 1%. Notably, the FPS improves significantly, increasing from 16.57 FPS to 24.21 FPS, representing a remarkable 46.11%improvement. These results demonstrate the superior performance of the C-ECO algorithm, which incorporates the lightweight MobileNet network and the A2-Net module.

Algorithms No. 2, No. 3 and No. 4 are used as the basic feature extraction network to conduct target detection tests on pedestrians in the video to verify their detection performance. In order to more clearly display the differences between algorithms, the detection results are displayed, as shown in Figure 5.

Figure 5 shows the comparison diagram of target detection for algorithms 2,3, and 4. The detection results of ResNet-50 algorithm are shown in Figure 5(a1-a4). There is target error detection in pedestrian detection, and the umbrella in the upper part of the figure is wrongly detected as a passenger. And the pedestrian detection effect on the road is not good. The detection results of MobileNet v2 algorithm are shown in Figure 5(b1-b4). The pedestrian detection is relatively accurate, and there is almost no problem of error detection. However, due to its simple network structure, the pedestrian detection effect of small targets on the right edge of the image in Figure 5(b4) is not good. In general, the pedestrian detection effect of algorithm 3 is significantly better than that of algorithm 2. The detection effect after adding A2-Net attention module to MobileNet v2 is shown in Figure 5(c1-c4). Due to the addition of the attention mechanism, the algorithm

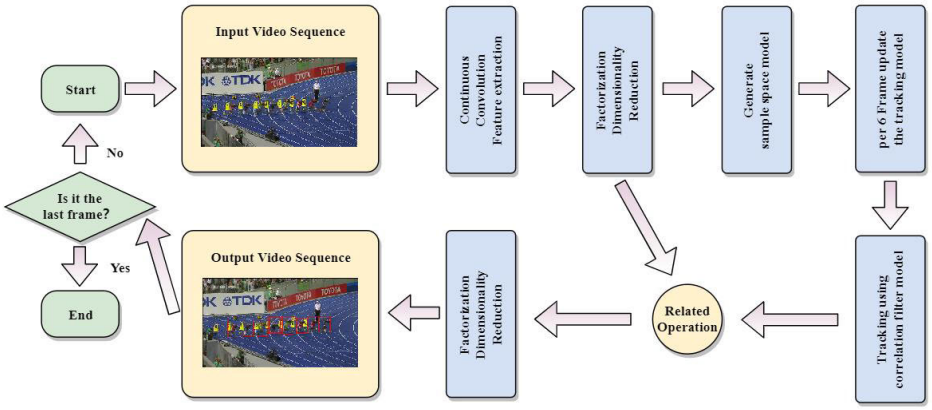


FIGURE 4. C-ECO algorithm flow.

TABLE 3. Ablation experiments.

Algorithm number	Algorithm	Model size (MB)	Accuracy (%)	Success (%)	FPS
1	ECO + VGG19	36.45	82.59	70.10	16.57
2	ECO + ResNet-50	33.90	83.74	71.16	17.33
3	ECO + MobileNet v2	25.41	79.17	67.85	26.85
4	ECO + MobileNet v2+ A2-Net	26.26	81.66	69.29	24.21

TABLE 4. Comparison experiment with and without pre-training.

Model	Pre-training	Precision (%)	Success (%)
C-ECO	√	83.68	67.35
C-ECO-N		81.06	64.88

has a stronger ability to extract features and a better ability to detect small targets than Algorithm 3 and Algorithm 2. It can be concluded from Figure 5 and Table 3 that the introduction of MobileNet v2 and A2-Net module can significantly improve the target detection ability, reduce the size of the model, and improve the speed of the pedestrian tracking algorithm.

3) MODEL PRE-TRAINING

Pre-training refers to the process of initially training a model on a large-scale dataset and subsequently fine-tuning it on specific downstream task data. This approach can accelerate model convergence and significantly reduce training time. To investigate the impact of the C-MobileNet feature extraction network in the proposed C-ECO algorithm on target tracking performance with and without pre-training, experiments were conducted on the OTB-100 dataset. Specifically, C-MobileNet was tested with and without pre-training. For pre-training, the MobileNet v2 model was selected, pre-trained on the ImageNet dataset, and its parameters were fine-tuned. The experimental results are summarized in Table 4.

From Table 4, it can be seen that the accuracy rate of the model with pre-training is 2.62% higher and the success rate is 2.47% higher than that of the model without pre-training (C-ECO without pre-training, C-ECO-N), from which it can be concluded that the model with pre-training is more effective in tracking the target, and therefore the model with pre-training is used in all subsequent sections.

4) TRACKING ALGORITHM PERFORMANCE COMPARISON EXPERIMENTS

In order to validate the effectiveness of the C-ECO algorithm, it is compared against several mainstream correlation filter tracking algorithms. The comparison algorithms include Kernel Correlation Filter (KCF), Discriminative Correlation Filter (DCF), Discriminative Scale Space Tracker (DSST), Spatially Regularized Correlation Filters (SRDCF), and Background-Aware Correlation Filters (BACF). By conducting this comparison, the performance and advantages of the C-ECO algorithm can be assessed in relation to these established tracking algorithms.

The comparison experiment uses the dataset OTB-50, which consists of 50 video sequences manually labeled with the true position of the target, classified according to different interference factors as occlusion (OCC), motion blur (MB), background clutters (BC), illumination variation (IV), low resolution (LR), scale variation (SV), deformation (DEF), out of view (OV), in plane rotation (IPR), fast motion (FM), out of The tracking (OPR), accuracy and tracking success rate of C-ECO and other tracking algorithms for different types of videos are shown in Tables 5 and 6, respectively, and the bolded data are the optimal data.

Based on the data presented in Tables 5 and 6, it is evident that the C-ECO algorithm proposed in this paper achieves a high accuracy rate when evaluated on the OTB-50 dataset.

FIGURE 5. Comparison of target detection effects before and after algorithm improvement. Figure 5(a1-a4) is the detection result of algorithm 2, Figure 5(b1-b4) is the detection result of algorithm 3, and Figure 5(c1-c4) is the detection result of algorithm 4.

Furthermore, it outperforms mainstream correlation filtering algorithms, exhibiting an average accuracy rate of 82.04%and a success rate of 64.72%. In complex scenarios such as FM, OCC, and MB, the C-ECO algorithm demonstrates a superior performance, surpassing the ECO algorithm by 0.51% in terms of accuracy and 0.23% in terms of success rate. These improvements highlight the enhanced accuracy and success rate achieved by the C-ECO algorithm over the ECO algorithm.

The accuracy and success rate curves of the C-ECO algorithm and mainstream tracking algorithms are shown in Figure 6.

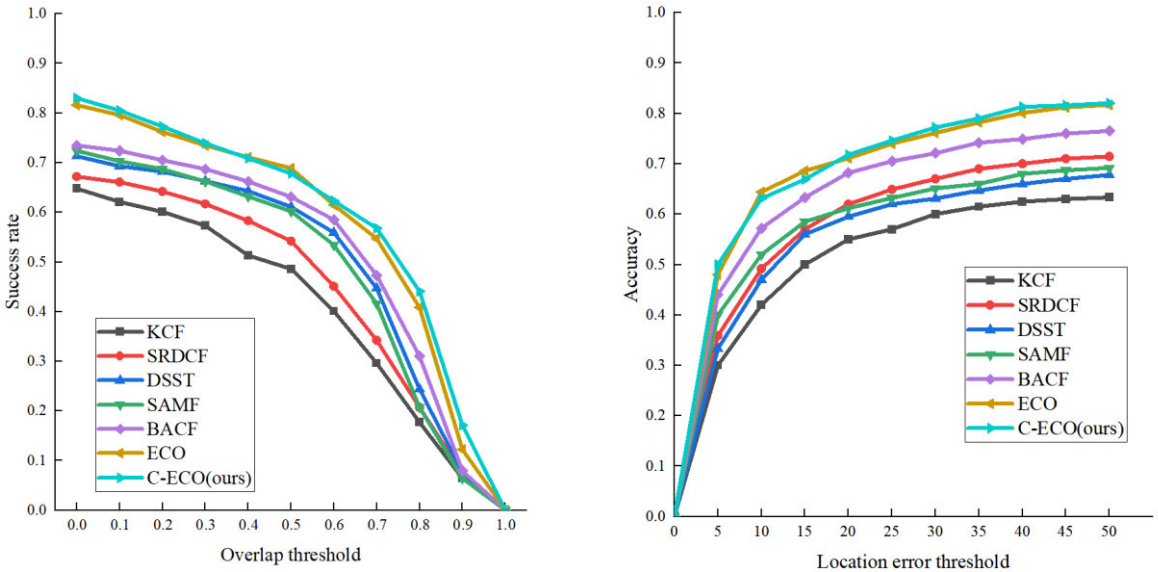
From Figure 6(a), it can be seen that when the position error threshold is less than 20, the accuracy value of the C-ECO algorithm in this paper is slightly lower than the ECO algorithm, and at the same time, it has a large lead compared with other mainstream tracking algorithms, and after the position error threshold is greater than 20, C-ECO

TABLE 5. Accuracy of each tracking algorithm on OTB-50 (%).

Algorithm	Type of video challenge											Average
	FM	BC	OCC	OPR	OV	LR	IPR	MB	DEF	IV	SV	
KCF	63.54	73.01	62.18	66.77	51.36	55.72	69.21	58.64	60.50	71.94	63.85	63.34
SRDCF	69.17	77.35	70.40	72.69	64.05	64.19	75.51	69.38	73.13	76.98	73.24	71.46
DSST	65.75	74.27	66.91	70.52	58.77	60.32	73.18	65.96	69.30	73.08	67.76	67.80
SAMF	66.58	75.46	67.34	70.31	62.98	67.10	72.29	68.16	68.72	72.63	69.44	69.18
BACF	76.49	78.06	71.88	78.12	**77.60**	**78.83**	79.27	72.11	73.56	77.61	78.32	76.53
ECO	80.33	81.56	79.64	**83.47**	76.81	77.70	**86.79**	78.41	84.35	**84.76**	**85.77**	81.69
C-ECO (ours)	**82.64**	**81.75**	**82.33**	82.58	77.35	78.67	84.04	**81.28**	**84.65**	83.46	85.48	**82.20**

TABLE 6. Success rate of each tracking algorithm on OTB-50 (%).

Algorithm	Type of video challenge											Average
	FM	BC	OCC	OPR	OV	LR	IPR	MB	DEF	IV	SV	
KCF	42.85	49.84	44.68	43.96	42.80	31.58	46.25	44.61	42.70	46.92	39.34	43.23
SRDCF	49.36	57.19	55.11	49.29	47.64	40.82	55.73	52.47	53.38	58.72	50.94	51.88
DSST	51.34	62.42	53.29	57.93	45.47	47.37	58.24	54.26	57.04	62.16	56.21	55.07
SAMF	53.43	61.89	56.32	62.12	49.38	54.03	62.10	57.49	58.03	60.63	55.79	57.38
BACF	58.92	59.34	56.61	61.00	**56.75**	55.69	62.40	57.33	59.48	62.22	57.99	58.88
ECO	67.58	**68.35**	62.08	**62.35**	54.71	54.82	**70.31**	67.50	**70.06**	**68.92**	62.72	64.49
C-ECO (ours)	**68.93**	66.40	**64.37**	61.81	55.23	**56.00**	68.49	**70.19**	69.49	67.86	**63.17**	**64.72**



(a) success plots (b) accuracy plots

FIGURE 6. Accuracy and success curves of 7 tracking algorithms on the OTB-100 dataset.

algorithm overtakes ECO algorithm and continues to lead. From Figure 6(b), it can be seen that when the IoU setting is less than 0.5, the gap between ECO and C-ECO success rate is not obvious, and when IoU is greater than 0.6 C-ECO is in the leading position and significantly better than other tracking algorithms.

The main challenges of the video in Figure 7(a1-a4) are SV, OCC, and LR. from Figure 7(a1), it can be seen that KCF, ECO, and C-ECO are able to track the target more accurately when the target is moving without disturbance, where C-ECO successfully tracked the small target. When a new target appeared as shown in Figure 7(a2), KCF tracked the shadow of the target incorrectly, and ECO and C-ECO were able to track it correctly. When the initial transformation as shown in Figure 7(a3-a4) occurred and the moving target was shaded and separated, both KCF and

FIGURE 7. Comparison of tracking effects. For clarity of display, only KCF, ECO and the algorithm C-ECO are used in this paper. black boxes are KCF, yellow boxes are ECO and blue boxes are C-ECO.

ECO lost the original target, and C-ECO was able to track accurately.

The main challenges of the video in Figure 7(b1-b4) are IV, DEF, MB. from Figure 7(b1-b2), it can be seen that KCF shows tracking drift when the target is occluded, as shown in Figure 7(b3-b4) when the video has motion blur and has a large angle change and a complex background, KCF prediction frame shows a large range of drift, ECO has a similarly colored background for the C-ECO algorithm is still able to track more accurately.

The principal challenges in the video sequence depicted in Figure 7(c1)-(c4) encompass issues such as Scale Variation (SV), Object Occlusion (OCC), Deformation (DEF), and Object Perspective Changes (OPR). From the analysis of frame c2, it becomes apparent that as the complexity of occlusion scenarios intensifies, featuring rapid movements of pedestrians, interplay of street lamps, and mutual obstruction among pedestrians engaged in typical walking, the tracking efficacy of the KCF and ECO algorithms is notably compromised. In contrast, the proposed C-ECO framework maintains a commendable level of tracking performance under these demanding conditions.

Frame c3 highlights that, in the case of swiftly moving pedestrians, the KCF algorithm has regrettably lost track of the target entirely, and ECO, though valiant, still struggles to maintain effective tracking. Remarkably, C-ECO manages to uphold a higher degree of tracking accuracy even in the face of such dynamic scenarios.

In frame c4, we observe that when confronted with the challenge of tracking small targets within complex scenes rife with occlusions, both KCF and ECO have forfeited the ability to track the target. In contrast, C-ECO exhibits resilience and

maintains the ability to accurately track the target, thus showcasing its remarkable robustness in these intricate situations.

The primary challenges encountered in the video sequence depicted in Figure 7(d1)-(d4) encompass Illumination Variation (IV), Scale Variation (SV), Object Occlusion (OCC), and Deformation (DEF). Notably, from the scrutiny of frame d2, it becomes evident that when the tracked target shares a color proximity with the environmental background, both the KCF and ECO algorithms are susceptible to tracking drift.

Further insights from frames d3 and d4 reveal that as the tracked target progressively recedes from the camera, diminishing in size and encountering occlusions, both KCF and ECO manifest varying degrees of tracking drift. In contrast, owing to its augmented feature extraction capabilities, C-ECO demonstrates a heightened resistance to tracking failures in these challenging scenarios.

The experimental results presented above emphasize the superior performance of the C-ECO algorithm proposed in this paper compared to other classical correlation filter tracking algorithms. The C-ECO algorithm not only achieves higher accuracy and success rates in tracking, but it also exhibits a significantly smaller feature extraction model size compared to the ECO algorithm prior to optimization. Additionally, the algorithm effectively improves the frame rate for video tracking. The comparison with other classical correlation filter tracking algorithms highlights the strength and competitiveness of the C-ECO algorithm. Its enhanced accuracy and success rates solidify its position as a reliable and efficient tracking solution. Furthermore, the reduced model size contributes to its practicality and resource efficiency, while the improved frame rate enriches the user experience during video tracking tasks. These findings demonstrate the significance of the C-ECO algorithm in advancing correlation filter tracking methods and establishing it as a promising choice for various tracking applications.

IV. CONCLUSION AND EXPECTATIONS

In this study, we propose a novel C-ECO tracking algorithm that leverages a lightweight convolutional neural network. The algorithm employs MobileNet v2 for efficient feature extraction, integrates the A2-Net module to enhance feature representation, and incorporates a pre-trained model to expedite training. The primary objective is to improve tracking accuracy and success rates. Experimental results demonstrate that the C-ECO algorithm outperforms the previous ECO algorithm employing VGG Net, exhibiting a 27.96% reduction in model size and a 46.11% improvement in frame rate. Importantly, these improvements are achieved without compromising the accuracy and success rate achieved before the enhancements. When compared with six other mainstream tracking algorithms, including ECO, the C-ECO algorithm consistently ranks at the top, boasting an average accuracy rate of 82.04% and a success rate of 64.72%. The lightweight pedestrian tracking algorithm proposed in this paper showcases its ability to effectively detect and track pedestrians in various complex scenarios. This research provides a new perspective for intelligent monitoring applications, contributing to advancements in the field.

V. FUTURE WORK AND PROSPECTS

While the pedestrian tracking algorithm proposed in this paper, based on a lightweight convolutional neural network, has demonstrated favorable results in terms of tracking speed and accuracy, it is important to acknowledge the existing limitations and room for improvement.

Firstly, the experiments conducted in this study primarily encompassed scenarios with good lighting conditions, favorable weather conditions, and indoor monitoring scenes. To provide a more comprehensive evaluation of the algorithm's performance, future research should include pedestrian monitoring videos captured under poor lighting conditions at night, as well as videos recorded in adverse weather conditions such as rain, snow, and fog. By expanding the dataset to encompass these challenging scenarios, the algorithm's robustness and generalizability can be further scrutinized.

Despite efforts to simplify the algorithm's architecture in this study, the inference operation still requires a substantial amount of computing resources due to the inherent complexity of the model and its network. As part of our future work, we aim to optimize and refine the model to minimize computational demands, enabling it to be executed efficiently on mobile devices while maintaining or even improving its performance.

It is also essential to note that this algorithm has its limitations and areas for further refinement. Additional experiments can be conducted to address these deficiencies and improve the overall effectiveness of the proposed pedestrian tracking algorithm. By embracing these future endeavors, we aspire to achieve superior results and make significant advancements in the field of pedestrian monitoring and tracking.

REFERENCES

[1] M. Kumar and S. Mondal, "Recent developments on target tracking problems: A review," *Ocean Eng.*, vol. 236, Sep. 2021, Art. no. 109558.

[2] N. Mahmoudi, S. M. Ahadi, and M. Rahmati, "Multi-target tracking using CNN-based features: CNNMTT," *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 7077–7096, Mar. 2019.

[3] M. A. Khan, H. Menouar, and R. Hamila, "Visual crowd analysis: Open research problems," 2023, *arXiv:2308.10677*.

[4] M. A. Khan, H. Menouar, and R. Hamila, "Revisiting crowd counting: State-of-the-art, trends, and future perspectives," 2022, *arXiv:2209.07271*.

[5] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[6] H. Moridvaisi, F. Razzazi, M. A. Pourmina, and M. Dousti, "An extended KCF tracking algorithm based on TLD structure in low frame rate videos," *Multimedia Tools Appl.*, vol. 79, nos. 29–30, pp. 20995–21012, Apr. 2020.

[7] H. Yang, S. Gao, X. Wu, and Y. Zhang, "Online multi-object tracking using KCF-based single-object tracker with occlusion analysis," *Multimedia Syst.*, vol. 26, no. 6, pp. 655–669, Dec. 2020.

[8] K. S. Sanagavarapu and M. Pullakandam, "Object tracking based surgical incision region encoding using scalable high efficiency video coding for surgical telementoring applications," *Radioengineering*, vol. 31, no. 2, pp. 231–242, May 2022.

[9] J. T. Mbelwa, Q. Zhao, Y. Lu, F. Wang, and M. E. Mbise, "Visual tracking using objectness-bounding box regression and correlation filters," *J. Electron. Imag.*, vol. 27, no. 2, p. 1, Mar. 2018.
[10] B. Khan, A. Jalil, A. Ali, K. Alkhaledi, K. Mehmood, K. M. Cheema, M. Murad, H. Tariq, and A. M. El-Sherbeeny, "Multiple cues-based robust visual object tracking method," *Electronics*, vol. 11, no. 3, p. 345, Jan. 2022.
[11] V. Degli-Esposti, F. Fuschini, H. L. Bertoni, R. S. Thomä, T. Kürner, X. Yin, and K. Guan, "IEEE access special section editorial: Millimeter-wave and terahertz propagation, channel modeling, and applications," *IEEE Access*, vol. 9, pp. 67660–67666, 2021.
[12] Y. Liang, Y. Liu, Y. Yan, L. Zhang, and H. Wang, "Robust visual tracking via spatio-temporal adaptive and channel selective correlation filters," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107738.
[13] N. Zdarsky, S. Treue, and M. Esghaei, "A deep learning-based approach to video-based eye tracking for human psychophysics," *Frontiers Hum. Neurosci.*, vol. 15, pp. 1–10, Jul. 2021.
[14] H. A. I. T. Abdelali, H. Derrouz, Y. Zennayi, R. O. H. Thami, and F. Bourzeix, "Multiple hypothesis detection and tracking using deep learning for video traffic surveillance," *IEEE Access*, vol. 9, pp. 164282–164291, 2021, doi: 10.1109/ACCESS.2021.3133529.
[15] J. Zhang, Y. Liu, H. Liu, J. Wang, and Y. Zhang, "Distractor-aware visual tracking using hierarchical correlation filters adaptive selection," *Appl. Intell.*, vol. 52, no. 6, pp. 6129–6147, Sep. 2021.
[16] L. Almuqren, M. A. Hamza, A. Mohamed, and A. A. Abdelmageed, "Automated video-based face detection using Harris hawks optimization with deep learning," *Comput., Mater. Continua*, vol. 75, no. 3, pp. 4917–4933, 2023.
[17] I. Ahmed, M. Ahmad, A. Ahmad, and G. Jeon, "Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: Within 5G infrastructure," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 11, pp. 3053–3067, Oct. 2020.
[18] W. Zhang, Y. Du, Z. Chen, J. Deng, and P. Liu, "Robust adaptive learning with Siamese network architecture for visual tracking," *Vis. Comput.*, vol. 37, no. 5, pp. 881–894, Apr. 2020.
[19] Y. Wang, H. Huang, X. Huang, and Y. Tian, "ECO-HC based tracking for ground moving target using single UAV," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 1–12, Jul. 2020.
[20] P. Wang, M. Sun, H. Wang, X. Li, and Y. Yang, "Convolution operators for visual tracking based on spatial–temporal regularization," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 5339–5351, Jan. 2020.
[21] S. Shen, S. Tian, L. Wang, A. Shen, and X. Liu, "Improved C-COT based on feature channels confidence for visual tracking," *J. Adv. Mech. Design, Syst., Manuf.*, vol. 13, no. 5, 2019, Art. no. JAMDSM0096.
[22] R. Zhang, Y. Zheng, C. C. Y. Poon, D. Shen, and J. Y. W. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern Recognit.*, vol. 83, pp. 209–219, Nov. 2018.
[23] P. M. Raju, D. Mishra, and R. K. S. S. Gorthi, "Detection based long term tracking in correlation filter trackers," *Pattern Recognit. Lett.*, vol. 122, pp. 79–85, May 2019.
[24] D. Yuan, X. Chang, P. Huang, Q. Liu, and Z. He, "Self-supervised deep correlation tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 976–985, 2021.
[25] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106977.
[26] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5596–5609, Nov. 2019.
[27] Z. Liang and J. Shen, "Local semantic Siamese networks for fast tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 3351–3364, 2020.
[28] J. Zhang, J. Sun, J. Wang, and X.-G. Yue, "Visual object tracking based on residual network and cascaded correlation filters," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 8, pp. 8427–8440, Sep. 2020.
[29] J. Chen, D. Zhang, M. Suzauddola, and A. Zeb, "Identifying crop diseases using attention embedded MobileNet-V2 model," *Appl. Soft Comput.*, vol. 113, Dec. 2021, Art. no. 107901.
[30] B. Singh, D. Toshniwal, and S. K. Allur, "Shunt connection: An intelligent skipping of contiguous blocks for optimizing MobileNet-V2," *Neural Netw.*, vol. 118, pp. 192–203, Oct. 2019.
[31] A. Michele, V. Colin, and D. D. Santika, "MobileNet convolutional neural networks and support vector machines for palmprint recognition," *Proc. Comput. Sci.*, vol. 157, pp. 110–117, Jan. 2019.
[32] U. Kulkarni, M. S. Meena, S. V. Gurlahosur, and G. Bhogar, "Quantization friendly MobileNet (QF-MobileNet) architecture for vision based applications on embedded platforms," *Neural Netw.*, vol. 136, pp. 28–39, Apr. 2021, doi: 10.1016/j.neunet.2020.12.022.
[33] X. Zhai, H. Wei, Y. He, Y. Shang, and C. Liu, "Underwater sea cucumber identification based on improved YOLOv5," *Appl. Sci.*, vol. 12, no. 18, p. 9105, Sep. 2022.
[34] H. Fu, G. Song, and Y. Wang, "Improved YOLOv4 marine target detection combined with CBAM," *Symmetry*, vol. 13, no. 4, p. 623, Apr. 2021.
[35] K. Xu, Z. Wang, J. Shi, H. Li, and Q. C. Zhang, "A2-Net: Molecular structure estimation from cryo-EM density, volumes," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 1230–1237.
[36] Y. Chen, X. Zhang, W. Chen, Y. Li, and J. Wang, "Research on recognition of fly species based on improved RetinaNet and CBAM," *IEEE Access*, vol. 8, pp. 102907–102919, 2020.

HONGLEI WEI was born in 1973. He received the Ph.D. degree from the Dalian University of Technology. He is currently an Associate Professor with Dalian Polytechnic University. His research interests include machine vision, deep learning, and object tracking.

XIANYI ZHAI was born in 1998. He received the bachelor's degree from Dalian Maritime University. He is currently pursuing the master's degree with Dalian Polytechnic University. His research interests include machine vision, object detection, and object tracking.

HONGDA WU was born in 1998. He received the bachelor's degree from Dalian Polytechnic University, where he is currently pursuing the master's degree. His research interests include automated control and intelligent algorithms.