IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 November 2023, accepted 4 December 2023, date of publication 7 December 2023,
date of current version 14 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340914

RESEARCH ARTICLE

Dark-SORT: Multi-Person Tracking in Underground Coal Mines Using Adaptive Discrete Weighting

RUI WANG[1], **JINGZHAO LI**[1], **AND ZHI XU**[2]
[1]School of Artificial Intelligence, Anhui University of Science and Technology, Huainan 232001, China
[2]School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232001, China

Corresponding author: Jingzhao Li (Rwang1281@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 51874010, and in part by the Huaibei Major Science and Technology Project under Grant Z2020004.

ABSTRACT Tracking-by-detection is a popular paradigm for Multi-Object Tracking (MOT), but the problems of unstable tracking and frequent ID transitions still occur due to the low illumination, point light sources, and high dust in the underground coal mine space. In this respect, this paper proposes a Dark-SORT personnel tracking algorithm for downhole environment characteristics. First, a video image enhancement method is designed to enhance the video image quality and improve the localization accuracy of the detector for the dim and unevenly distributed light environment in the well. Second, an Adaptive Discrete-weighted Attention Module (ADAM) is designed, which consists of an Enhanced Discrete Channel Attention (EDCA) module and an Adaptive Discrete Spatial Attention (ADSA) module. EDCA enables the network to capture richer information at different scales by adaptively processing different channels according to their importance and feature scales. The ADSA approach enhances the linkage between different locations within the same region, combines different pooling strategies to highlight important regions, and reduces the focus on overexposed regions. Finally, the OC-SORT tracking algorithm is introduced to solve the error accumulation problem based on the motion model and incorporate the appearance feature information to improve the stability of target tracking. We conducted a comparison test on the self-built dataset MINE-MOT, and the HOTA, MOTA, DetA, AssA, IDF1, AssRe, and FPS metrics of the Dark-SORT tracking algorithm based on the YOLOv7 target detection model were 67.4, 92.6, 80.3, 46.8, 61.7, 65.7, and 23, respectively, which was the best in terms of accuracy and stability of all the models involved in the test.

INDEX TERMS Attention mechanism, computer vision, target detection, target tracking, image enhancement.

I. INTRODUCTION

The coal mine underground space is narrow and dim, light distribution is mixed, and the staff needs to work with large machinery and equipment, making the undercover operation personnel face many safety risks, such as collision, entrapment, fall, etc. For this reason, it is essential to monitor and track the activities of underground personnel to issue warnings about potential safety risks promptly.

Early target-tracking algorithms primarily relied on traditional machine vision techniques, including methods like template matching and feature-based approaches. For instance, features like Histogram of Oriented Gradients (HOG) [1] were employed to describe the target and were combined with machine learning techniques to facilitate target tracking. These methods perform well in dealing with changes in target appearance but are susceptible to background interference. Accompanied by the progress of deep learning technology, neural network-based target detection algorithms realize automatic learning and representation of

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

target features by end-to-end training on large-scale datasets, obtaining better detection results.

An exemplary series of methods, including the R-CNN [2], [3], [4] family, ushered in a new era in target detection by harnessing convolutional neural networks. These pioneering techniques involve feature extraction across candidate regions, employing classifiers and regressors to ascertain the target's class and precise location. The YOLO (Look Only Once) [5], [6], [7], [8] family divides the image into grids and predicts the category and location information of objects within each grid. These end-to-end methods have good accuracy and real-time performance, promoting further development of target-tracking algorithms. MOT algorithms usually contain Tracking-By-Detection (TBD), Joint Detection and Tracking (JDT), and Joint Detection and Embedding (JDE) [9] paradigms. TBD is a multi-object tracking method based on the detector output with high tracking accuracy. JDT methods combine the detector and the tracker as a whole, emphasizing the sharing of information between detection and tracking, which helps to improve the tracking accuracy but usually requires a more complex neural network architecture. The JDE approach combines detection and tracking and feature embedding of targets, emphasizing target identification and identity association, and is therefore suitable for scenarios where specific targets need to be identified and tracked.

Video images from underground coal mines face many problems, such as dim environments, geographic complexity, uneven exposure, blurred details, and noise, and it is a challenge for detectors to find targets in these images. Image enhancement processing and attention mechanisms are commonly used to solve this challenge. However, the suddenness of coal safety accidents still needs to consider the real-time requirements of the algorithms. Image enhancement algorithms tend to become more complex with the development of recent years, both traditional methods [10], [11], [12] and deep learning methods [13], [14], [15], while images under normal light cannot be provided for end-to-end training in coal mines. Moreover, coal mine underground images are very different from the imaging characteristics of most scenes, which are characterized by less information in the high exposure region and lower contrast in the target region. This may lead to unsatisfactory accuracy enhancement of ordinary attention mechanisms. For this reason, it is necessary to design a video image enhancement processing method that takes into account the image enhancement effect and real-time performance and to construct an attention mechanism applicable to the image features of underground coal mines in order to improve the localization accuracy of the detector and thus enhance the stability of the target tracking algorithm. The following research is carried out in this paper.

(1) To address the problem of poor imaging quality due to the dark and uneven distribution of light in the underground environment of coal mines. First, the image's brightness and contrast are enhanced based on the CLAHE algorithm. Secondly, the HSV spatial transformation is applied to the improved image, and the improved two-dimensional gamma function adjusts the regions where the light components are too strong and weak to achieve illumination balance.

(2) The Adaptive Discrete-weighting Attention Module (ADAM) design comprises EDCA and ADSA components. EDCA determines channel importance by employing global discrete pooling. Convolutional operations and activation functions derive a one-dimensional weight distribution for each channel. Subsequently, it designs a separable convolutional module with adaptable scale and depth based on these weight distributions. This design enables the network to capture varying information at different scales across different channels.

On the other hand, ADSA enhances inter-channel connections within the same spatial region. It combines diverse pooling strategies to fine-tune the network's attention, mainly focusing on target regions and reducing overexposure. This enhancement contributes to improved detection accuracy, especially in scenarios characterized by dim backgrounds and low contrast.

(3) A high-performance tracking algorithm based on a pure motion model is introduced. Combining the Observation Centered Momentum (OCM) module and Observation Centered Recovery (OCR) module proposed by OC-SORT, as well as the introduction of appearance information features to help reduce the accumulation of errors based on the motion model and improve the re-identification ability of the tracking algorithm, which in turn improves the stability of the target tracking process.

II. RELATED WORK

Bewley et al. [16] proposed to predict the future target position using the Kalman filter by weighting the prediction and detection frames to get the filter value and then tracking using the Hungarian matching algorithm. The experimental results show the method has a good tracking effect in an interference-free environment. Bochinski et al. introduced a fast Multiple Object Tracking (MOT) system [17] that correlates neighboring frames based on spatial overlap. This system is known for its simplicity and computational speed. However, densely populated scenes with complex object overlaps may encounter frequent identity (ID) conversion challenges. Wang [18] et al. proposed a pedestrian tracking algorithm based on multi-granularity. The pedestrian features extracted by the neural network are combined with basic color features to determine the results of the tracking algorithm, and the tracking results are modified according to the target detection results. Bae and Yoon [19] utilized Linear Discriminant Analysis (LDA) to extract re-recognition features for the target object. This approach enhances re-recognition performance and demonstrates greater robustness compared to alternative algorithms. A bi-directional network GRU was proposed in [20] to build and retain candidate trajectories with high confidence in sparse environments based on object features extracted from CNN and RCNN neural networks.

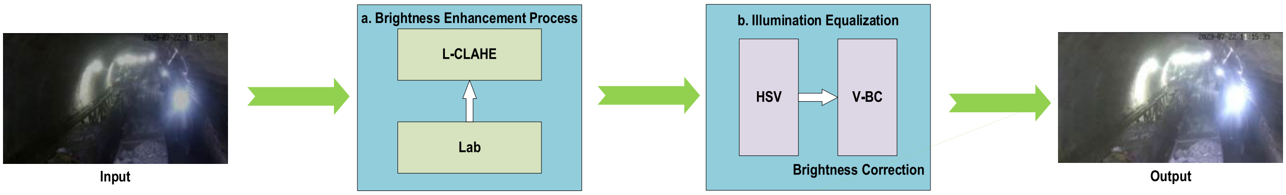
Input
a. Brightness Enhancement Process
L-CLAHE
Lab
b. Illumination Equalization
HSV
V-BC
Brightness Correction
Output

FIGURE 1. Image enhancement processing flowchart. (a) is the brightness and contrast adjustment module. (b) is the illumination adjustment module.

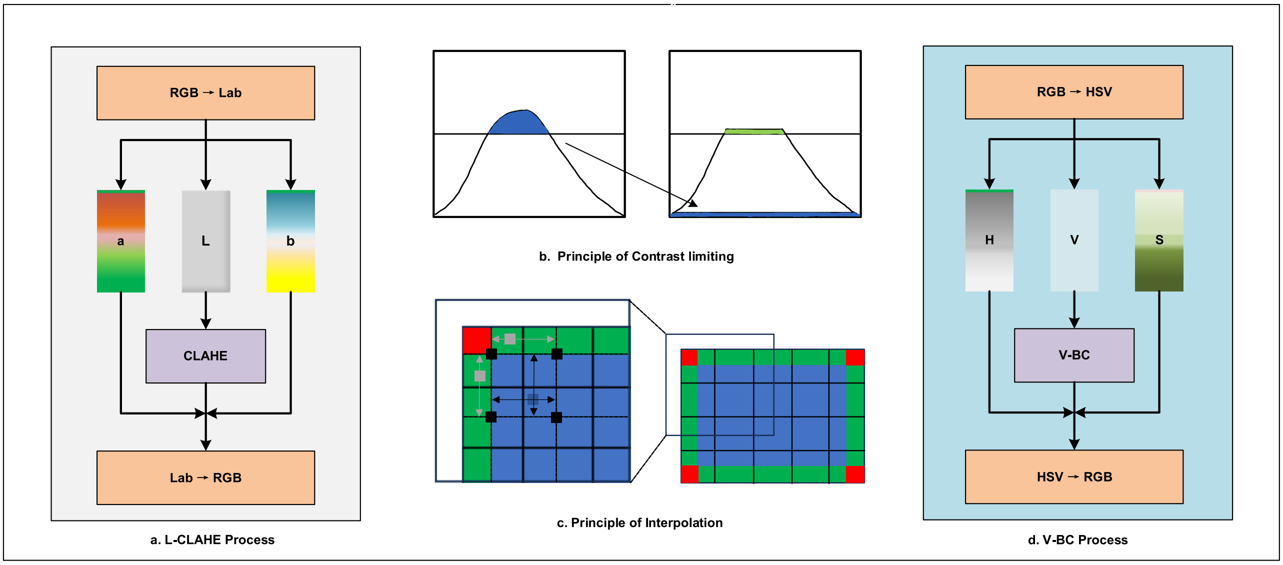

FIGURE 2. (a) Flowchart of the L-CLAHE algorithm. (b) The noise suppression schematic. (c) Schematic diagram of interpolation. (d) Flowchart of the V-BC process algorithm.

In their study [21], Wojke et al. recombined appearance information to enhance feature extraction performance.

Based on this improvement, the algorithm reduces the number of identity transitions when faced with long-term occlusion situations. Yu et al. [22] conducted an in-depth study on high-performance detection based on appearance features and deep learning. They demonstrated that the proposed methodology consistently yields improved results in real-time and offline tracking scenarios for Multiple Object Tracking (MOT). A single-shot multi-object tracking system was developed in the literature [23] by improving the dilation convolution and adding a spatial channel attention mechanism. The system was evaluated on three commonly used multi-object tracking datasets and achieved good tracking results. In their study, referenced as [24], the authors enhanced the YOLO network, devised the architecture for the DeepSORT pedestrian tracking method, and incorporated the Kalman filter algorithm for precise motion state estimation of pedestrians. The validation results show that the method reduces pedestrian targets' leakage and false detection rates. Literature [25] proposed a novel parallel converter network architecture and designed transformer-1 module, transformer-2 module, and feature fusion head (FFH) based on the attention mechanism to achieve robust visual tracking. Gu et al. [26] have elaborated a local feature information association module (LFIA) and a global feature information fusion module (GFIF) based on the attention mechanism, which can effectively utilize contextual information and feature dependencies to enhance feature information. Literature [27] proposes a multi-feature response map adaptive fusion strategy that adaptively weights and fuses different features to solve the tracking failure problem due to occlusion. Literature [28] has designed a shared encoder dual-pipeline converter architecture based on a hybrid attention mechanism, which incorporates a concise tracking prediction network to obtain an efficient tracking representation. Literature [29] extracted the motion feature of the tracked target on the time axis and added this feature to the computation of the cost matrix to alleviate the tracking instability problem due to occlusion. In the literature [30], in order to reduce the annotation cost and ensure the diversity of the selected samples, a multi-frame collaborative active learning-based method, and a nearest-neighbor discrimination method are proposed for the screening of the training samples. The results show that the method is competitive in terms of tracking accuracy and speed compared to state-of-the-art trackers.

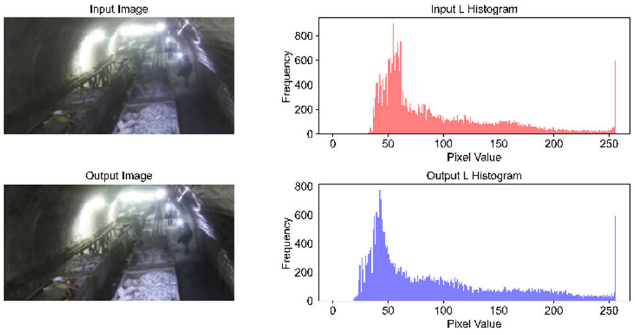

FIGURE 3. **Comparison of L-CLAHE treatment results and L-component histograms.**

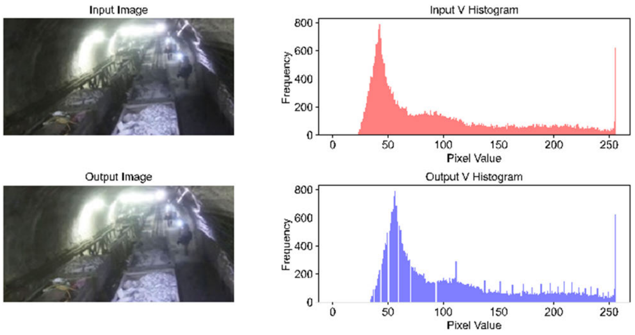

FIGURE 4. **Comparison of V-BC processing results and V-component histograms.**

III. MATERIALS AND METHODS

A. IMAGE ENHANCEMENT METHOD

Video surveillance images are characterized by low overall brightness and contrast as well as a wide range of brightness due to the distributed light in the well. In this section, a video image enhancement method with image brightness, contrast adjustment, and illumination equalization is designed for the characteristics of underground coal mine images. Its processing flow is shown in Fig. 1.

Step(a) in Fig. 1 is a brightness and contrast enhancement processing module that uses a limiting contrast adaptive histogram equalization method to enhance the overall brightness and contrast of the image. Step (b) is a light equalization module that performs light unevenness correction using a modified two-dimensional gamma function to balance overly dark and overexposed regions.

Constrained Contrast Adaptive Histogram Equalization (CLAHE) [31] involves dividing the image into small chunks and performing histogram equalization on each chunk separately to perform local contrast enhancement for each chunk. The L-CLAHE process is shown in Fig. 2(a), where the image is first converted to Lab space, the "L" component corresponds to luminance, the "a" and "b" components correspond to chromaticity, and then the "L" component is processed by the CLAHE method, which avoids over-enhancement that leads to the loss of noise or local details. Limiting the contrast is to limit the grayscale distribution of the image, which can remove the noise generated when expanding the contrast, and the pixel points of the portion exceeding the threshold are uniformly distributed to lower grayscale values, the principle of which is shown in Fig. 2(b). Adaptive histogram equalization is the interpolation operation, and the principle of interpolation operation is shown in Fig. 2(c). In the figure, the histogram, histogram accumulation function, and the corresponding transformation function are derived for each square, the blue area pixels are obtained by bilinear interpolation with the transformation function of its four neighbors, the green area pixels are obtained by linear interpolation with the transformation function of its two neighbors, and the red area pixels are obtained by adopting its transformation function. Fig. 3 presents the outcomes of L-CLAHE processing, showcasing a notable enhancement in both brightness and contrast in the output image when compared to the input image, all while maintaining minimal noise interference.

The V-BC processing flow is shown in Fig. 2(d). First, the measures of HSV spatial conversion can separate the hue component H, saturation component S, and luminance component V and extract the luminance component V with the help of the multi-scale Gaussian function $G(x, y)$. The expression is:

$$G(x, y) = \lambda \exp\left(-\frac{x^2 + y^2}{c^2}\right) \tag{1}$$

where c is the scale parameter, and λ is the normalization factor. $G(x, y)$ can be realized by adjusting c and the weighting coefficient to extract the optical component V. The final expression of the optical component estimate is obtained as:

$$I(x, y) = \sum_{k=1}^{M} \omega_k \left[L(x, y)G_k(x, y)\right] \tag{2}$$

where $I(x, y)$ is the estimated value of the illumination component; $L(x, y)$ is the input image, ω_k is the weighting coefficient of the illumination component; $k = 1,2,\ldots M$ is the number of scales, and the weighting coefficient of the scales satisfies $\sum_{k=1}^{3} \omega_k = 1$. Secondly, the improved 2D gamma function is constructed to correct the illumination component, and finally, the reconstructed image is returned to the RGB space. The expression of the improved 2D gamma function is:

$$V(x, y) = 255\left(\frac{L(x, y)}{m}\right)^{\gamma^2} \tag{3}$$

$$\gamma = \left(\frac{1}{2}\right)^{\frac{m-I(x,y)}{m}} \tag{4}$$

where $V(x, y)$ is the corrected luminance value, the illumination component is $I(x, y)$, γ is the correction parameter, and m is the luminance mean value of the illumination component.

The results of V-BC processing are shown in Fig. 4. Compared with the input image, the light and dark distribution of the output image is more uniform, and the dark details

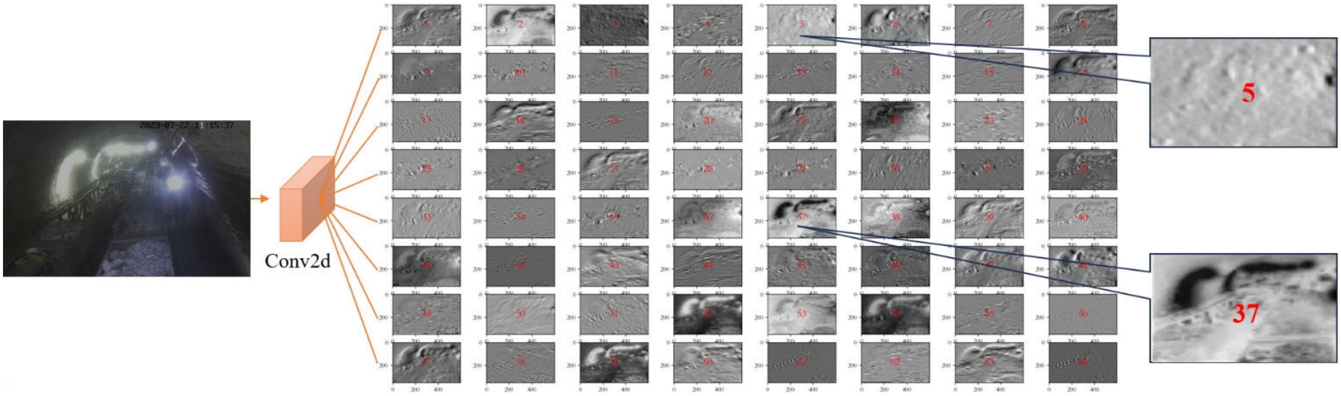

FIGURE 5. Shallow convolutional network 64-channel output feature map visualization results.

are emphasized. For the sake of the following discussion, we name this image enhancement method IE.

B. ADAPTIVE DISCRETE-WEIGHTED ATTENTION MODULE

1) LIMITATIONS PROPOSED

The precision of the target detection method serves as a fundamental pillar for maintaining the continuous and accurate tracking of targets along with reliable ID assignments. The YOLO series represents a single-stage target detection algorithm embracing the concept of one-step regression classification. Notable for its remarkable detection precision, swift computational performance, and streamlined network architecture, it has found extensive utility across various domains, spanning industries, transportation, and medical applications. In recent years, the YOLO series has been continuously developed. For example, WANG et al. further improved the YOLOv7 target detection algorithm based on the YOLOv5 target detection algorithm, which has been improved in both accuracy and speed. However, the environmental characteristics of dim and point light sources in underground coal mines are challenging for detectors and require targeted improvements in the feature extraction capabilities of the network. Attentional mechanisms have been shown to be an effective measure. Jie et al. [32] et al. proposed SENet, which obtains a scalar by global average pooling of the feature layer as its weighting factor, allowing channels with a larger number of parameters to be given higher importance by the network. CBAM [33] incorporates the SENet channel attention part while adding a new serial spatial attention module. However, for coal mine underground images, the above algorithm fails to solve the following two problems.

(1) It is known from practice and experience that shallow CNNs can usually capture lower-level features such as edges, textures, and simple shapes. Traditional attention algorithms tend to use global average pooling or global maximum pooling as the basis for judging the importance of feature layers; this determination is effective in deep CNNs, but it can be inaccurate in shallow networks, especially for images with low contrast and high exposure characteristics. Through experimental analysis, we visualize the feature layer output from the 64 channels of the shallow convolutional layer of the network and perform a global average pooling operation for each channel. From the experimental results, it can be seen that Fig. 5(5) is the maximum value of average pooling. However, there are better solutions among the 64 channels, both in terms of the portrayal of texture and edge information and subjective perception.

(2) Dark ambient point light sources form an overexposed region characterized by high contrast at the edges and low information content at the center. By visualizing the feature maps of the 64-channel output of the convolutional network, we find that the network pays particular attention to this overexposed region, which is expressed by the fact that most of the channels have more texture and edge information portrayed in this region, but in fact, this region is not the target region. Although the target region is more informative the overall contrast is low, through Fig. 5, we find that the convolutional network pays less attention to the target region, which is expressed by the fact that most of the channels do not even have texture and edge information for this region.

Typically, features represented by channels with higher information content may contain significant edge and texture information, while channels with lower information content may present smoother and more homogeneous features. According to Shannon's theorem, information entropy is a concept that measures the amount of information in a system, and the more chaotic the distribution of pixel values, the more information a feature layer contains. Therefore, we can use the information entropy as a criterion for determining the importance of the feature map, the formula of which is shown below:

$$H(X) = -\sum\nolimits_{i} p(x_i)log_2(p(x_i)) \qquad (5)$$

where $p(x_i)$ is the probability of x_i appearing. While the operation of information entropy is more complex, which has a greater impact on the real-time nature of the network, the standard deviation is also an indicator for evaluating the degree of system fluctuations, and the calculation is relatively

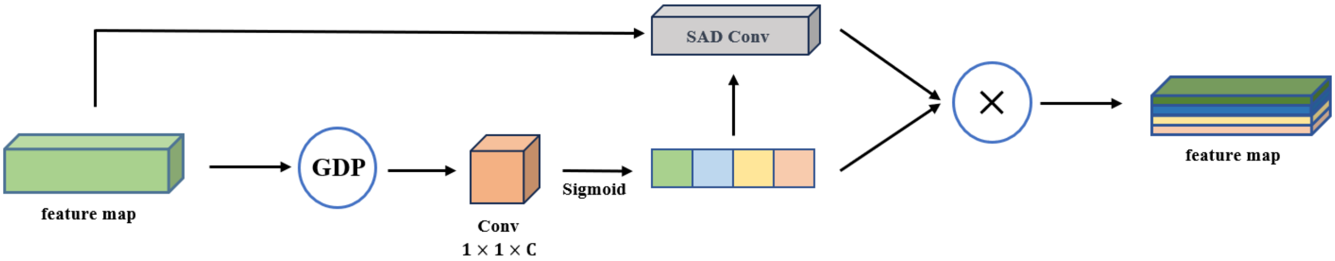

FIGURE 6. EDCA module structure diagram.

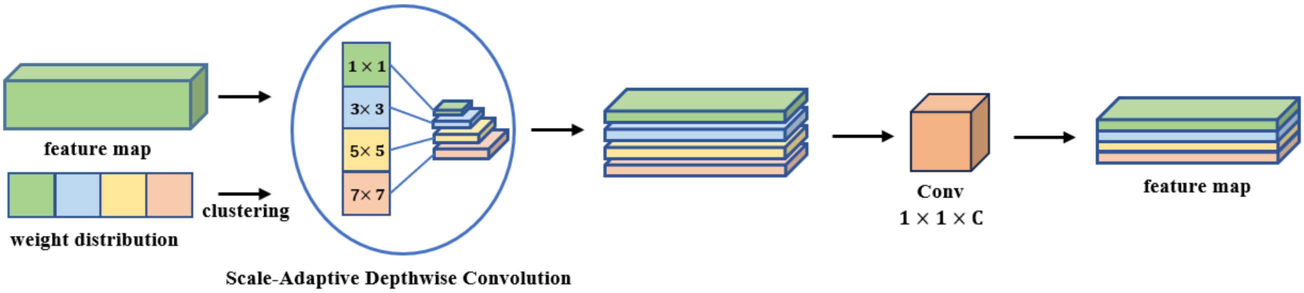

FIGURE 7. SAD Conv module structure diagram.

simple, for this reason, we introduced Global Discrete Pooling (GDP). Its expression is:

$$S(x)_{i,j} = \sqrt{\frac{1}{k^2}\sum_{m=0}^{k-1}\sum_{n=0}^{k-1}\left(x_{is+m,js+n} - A(x)_{i,j}\right)^2} \qquad (6)$$

where $S(x)_{i,j}$ denotes the local discrete pooling at feature maps (i, j), $A(x)_{i,j}$ denotes the local mean pooling at feature maps (i, j), $x_{is+m,js+n}$ is the corresponding position of the input feature maps, k denotes the window size of pooling, and s denotes the step size of pooling. Using this method for global discrete pooling of the 64-channel convolutional output layer, we obtain the results shown in Fig. 5(37), which better highlights some texture and edge details of the scene and contains richer information.

2) DESIGN OF ADAM ATTENTION MECHANISM

ADAM is designed in this section for such image features and the above experimental results. This attention module consists of an Enhanced Discrete Channel Attention (EDCA) module and an Adaptive Discrete Spatial Attention (ADSA) module in series. The main idea of EDCA is to guide the convolutional neural network to apply the channels' information more efficiently to improve the feature extraction capability of the convolutional neural network in the target detection task in low-light scenes. The module measures the importance of each channel in the task by analyzing the degree of discrete elements within the channel and, in combination with depth-separable convolution, automatically adjusts the size of the convolutional kernel according to the degree of importance of the channel, thus optimizing the feature extraction process. The structure of the EDCA channel-attention mechanism is shown in Fig. 6.

The channel attention mechanism evaluates the importance of each channel of the feature map using a global discrete pooling approach, learns the relative importance of different channels through a convolutional layer, and obtains a weight distribution between them via an activation function. Combining weight distribution and depth-separable convolution [34], we designed the Scale-Adaptive Depthwise Convolution (SAD Conv) module with the structure shown in Fig. 7. This module combines the weight distribution to configure different sizes of convolution kernels for different channels so that the network can adapt to features at different scales and capture beneficial feature information at various scales more finely. Finally, the weight distribution is multiplied with the output of the SAD Conv module in a multiplication operation to obtain the empowered feature map. The formula for the EDCA operation is shown below:

$$\begin{aligned} EDCA(x) = {} & sigmoid(Conv^{1\times 1}(GDP(x)) \\ & \times SADConv(x) \end{aligned} \qquad (7)$$

Based on the clustering of the weight distribution, the SAD Conv module adaptively divides the channels into four parts and assigns four different scales of convolution kernels 1×1, 3×3, 5×5, and 7×7 respectively. Accordingly, the channels with lower information content use more delicate convolution methods to extract more detailed features, while the channels with higher information content use more coarse convolution methods to extract the overall features. The size of the feature map is kept constant by adjusting the step size

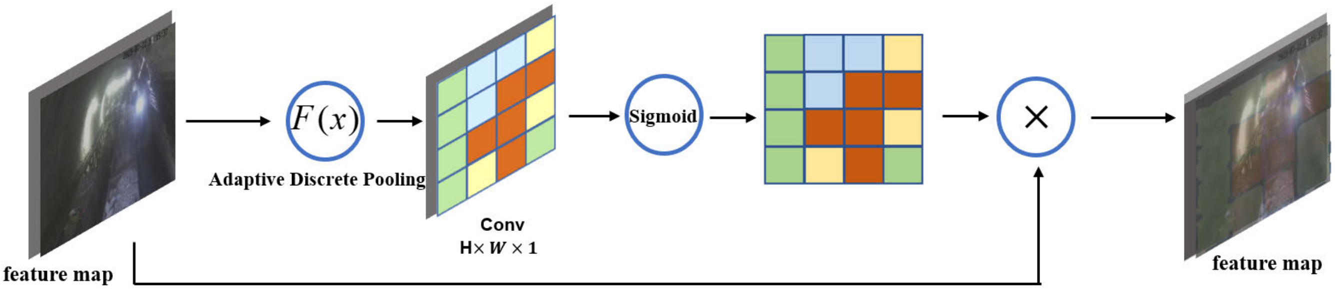

FIGURE 8. Adaptive discrete pooling weighted spatial attention mechanism architecture diagram.

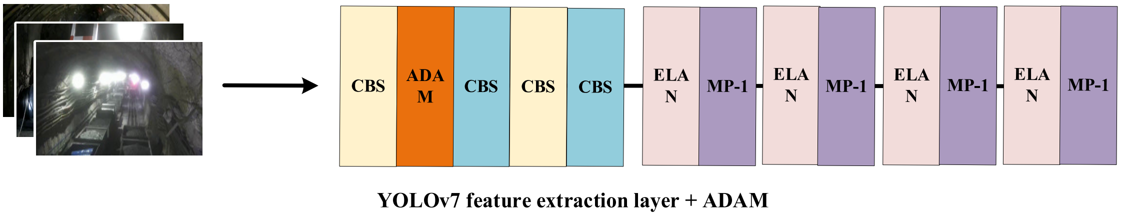

FIGURE 9. Network structure of YOLOv7 feature extraction layer with added ADAM module.

and padding, and the channels are finally combined by point-by-point convolution. EDCA allows different channels to be personalized according to their importance and feature scales, thus enabling the network to capture richer information at different scales.

ADSA is a spatial attention mechanism based on adaptive discrete pooling weighting, which is designed to enhance the deep learning network's ability to attend to and recognize the target region under conditions such as dim background, exposure, and low contrast. The structure of ADSA is shown in Fig. 8.

This attention module performs local discrete pooling on the input feature map and adaptively adjusts the subsequent pooling strategy according to the local discrete pooling values, using different local pooling methods for different regions. Compared with the traditional algorithm based on a single pixel at the same location in different channels, ADSA enhances the connection between different locations in the same region, combines discrete pooling and uniform pooling to differentiate between overexposed bright and dark regions, and gives different pooling strategies to highlight or reduce the weight of a specific region, to enhance the network's attention to the region with high information content and low contrast, and to reduce the network's attention to the overexposed region. The expression for $F(x)$ is:

$$F(x)=\begin{cases} GMP_{i,j}^{k}+GAP_{i,j}^{k},\, S_{i,j}>GDP; & A_{i,j}<GAP\\ GAP_{i,j}^{k},\, S_{i,j}>GDP; & A_{i,j}>GAP\\ GMP_{i,j}^{k},\, S_{i,j}<GDP; & A_{i,j}<GAP\\ GDP_{i,j}^{k},\, S_{i,j}<GDP; & A_{i,j}>GAP \end{cases} \quad (8)$$

where $GMP_{i,j}^{k}$, $GAP_{i,j}^{k}$, $GDP_{i,j}^{k}$ denote global maximum pooling, global average pooling, and global discrete pooling for the region of size $k \times k$ at (i, j) respectively. $S_{i,j}$, $A_{i,j}$ are local discrete pooling and local uniform pooling, respectively. GDP and GAP are global discrete pooling and global average pooling for feature maps, respectively. By doing so, the advantages of different pooling methods can be fully utilized. From Eq. 8, it can be seen that for the region of $S_{i,j} > GDP; A_{i,j} < GAP$, which has more texture and edge information and lower overall brightness, using global maximum pooling and global average pooling added together helps to increase the network's attention to this region; For regions of $S_{i,j} > GDP; A_{i,j} > GAP$, which have more texture and edge information and higher overall brightness, the global average pooling approach helps to preserve the overall correlation between features. $S_{i,j} < GDP; A_{i,j} < GAP$ is an over-dark region that contains less information, and the use of global maximum pooling helps to enhance the extraction of detailed features in this region of the network; $S_{i,j} < GDP; A_{i,j} > GAP$ region has a low degree of discretization relative to the full map but has a high luminance, which is an overexposed region that contains less information, and the use of global discrete pooling helps to diminish the attention of this region of the network.

ADAM can provide more accurate and robust processing of images with dim backgrounds and inconspicuous target features in tasks such as target tracking and target detection, improving the performance of deep learning networks in complex scenes. In this research, we introduce the ADAM module into the early layers of the multi-scale feature extraction network within the YOLOv7 framework. This inclusion serves the purpose of aiding the model in discerning and enhancing critical patterns and details within the initial low-level features. Consequently, it significantly improves the detector's accuracy in object localization. For a visual

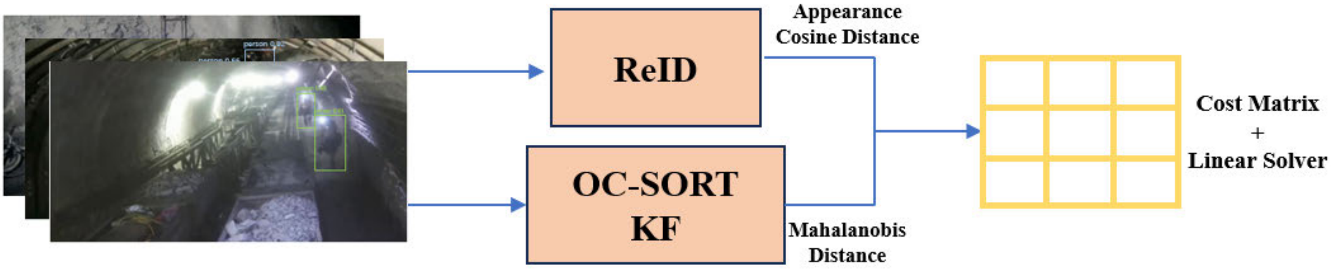

FIGURE 10. SOC-SORT structure diagram.

representation of this architectural modification, please refer to Fig. 9.

C. ADDITIONS AND EXTENSIONS TO OC-SORT

SORT consists of Kalman filtering and Hungarian matching algorithms and is a widely used tracking algorithm. The principle is to linearly distribute the detection frame IoU obtained by the target detection algorithm and the tracking frame IoU predicted by the Kalman filtering algorithm into the Hungarian algorithm to correlate the inter-frame ID. Kalman filtering consists of two main processes: prediction and correction. In the prediction phase, the filter utilizes the estimate of the previous state to predict the current state. The prediction model is as follows:

$$\hat{x}'_k = A\hat{x}_{k-1} + Bu_k \tag{9}$$

$$P'_k = AP_{K-1}A^T + Q \tag{10}$$

where $\hat{x}'_k$ denotes the predicted value at the moment k, A is the state transfer matrix, $\hat{x}_{k-1}$ denotes the estimated value at the moment $k-1$, B and u_k denote the model system parameters. P'_k denotes the covariance matrix between the predicted value and the true value, P_{K-1} denotes the error covariance matrix between the filtered value and the true value at moment $k-1$, and Q is the state noise covariance matrix. As can be seen from eq. 9. the predicted value is linearly related to the estimated value at the previous moment, and such an approach is effective when the frame rate is high. However, when the detector loses the target, the measurements lost during this period update the combined error of the Kalman filter parameters in a quadratic manner over time. Moreover, while detecting the movement of the target can be approximated as a linear model, using high frame rate video increases the sensitivity of the system to state noise.

The strategy of OC-SORT [35] is to focus the design philosophy of the tracker on an observation-centered approach rather than relying solely on state estimation to take full advantage of the object's kinematic momentum and incorporate it into the correlation phase. When a trajectory is interrupted due to loss, an observation-centered re-update (ORU) method eliminates errors accumulated during the untracked period. In addition, an observation-centered momentum (OCM) term is introduced into the correlation cost to enhance the tracker's performance further.

In the ORU method, a re-update operation is performed when the object is not tracked for some time and then detected again by the detector. This means that we need to know the object's trajectory during the untracked period and thus readjust the Kalman Filter (KF) parameters to reflect the actual motion more accurately. So, we need to observe the object's state at the beginning and the end of the untracked period to generate the object's virtual trajectory during the untracked period. Then, the observed data of the virtual trajectory is compared with the latest actual observation to update the parameters of the Kalman filter. This approach allows the updating process to no longer be affected by errors introduced through virtual updating, thus improving the tracking accuracy.

The assumption of linear motion of a target requires that the direction of motion of the target remains consistent. However, in practice, noise and uncertainty make it impossible to utilize motion direction consistency accurately. To address these issues, OC-SORT proposes a method to reduce the noise in the motion computation using observed data rather than estimated data (OCM). Since the observation data is unaffected by the time error amplification problem, it can be used to calculate the motion direction more accurately. At the same time, the concept of velocity consistency is introduced to help with target correlation by adding a velocity consistency (momentum) term to the cost matrix. Velocity consistency means that the change in velocity is consistent between different states of an object, even if the time interval is large. The associated cost formula for OC-SORT is shown in equation 11:

$$C(\hat{G}, Z) = C_{IOU}(\hat{G}, Z) + \lambda C_v(\hat{G}, Z, V) \tag{11}$$

$$\begin{aligned} C(\hat{G}, Z) = \eta D_M(\hat{G}, Z) + (1-\eta)D_A(\hat{G}, Z) \\ + \lambda C_v(\hat{G}, Z, V) \end{aligned} \tag{12}$$

where $\hat{G}$ is the estimated state matrix of the target, Z is the state matrix of the detection, and V is the direction containing the existing trajectory computed from the two previous time-difference observations. C_{IOU} calculates the IoU value between the negative pair (Intersection over Union) detection frame and the prediction frame. C_v computes the orientation of the trajectory and the difference in orientation formed by the historical and new detections of the trajectory,

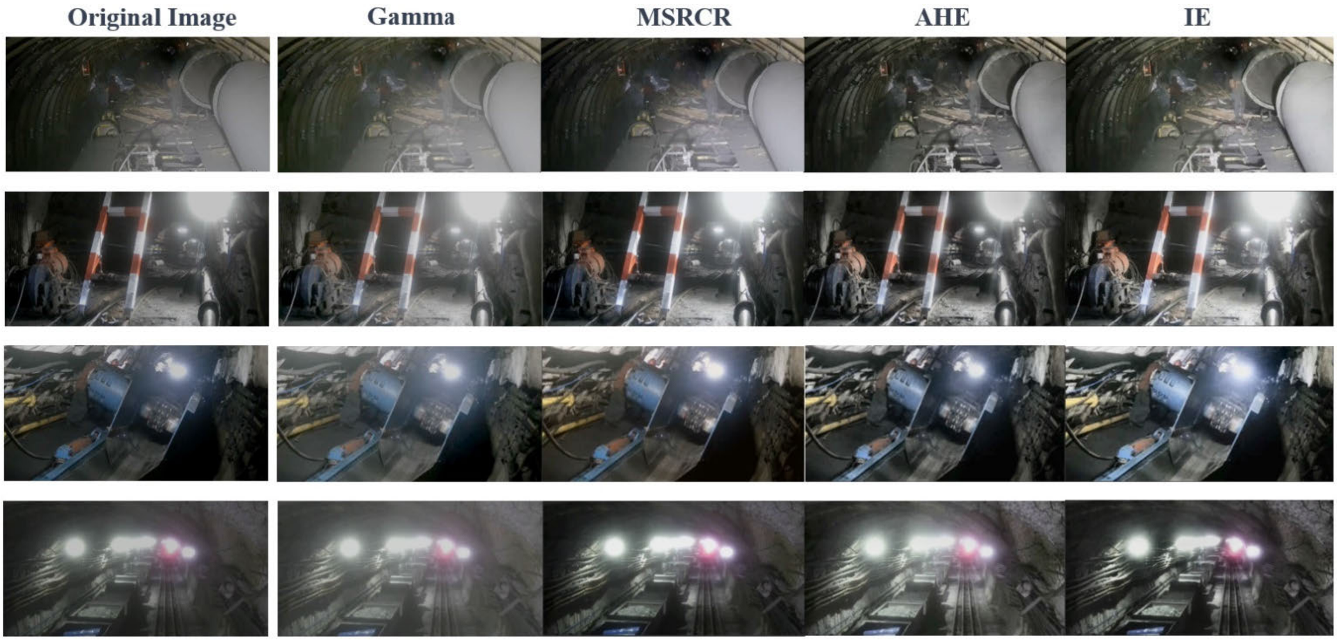

FIGURE 11. Comparative tests of image enhancement algorithms.

and λ is the weighting factor. This method uses the detected values associated with the trajectory for direction calculation, which avoids error accumulation in the estimation state. However, when the target is moving fast or the frame rate is low, the bounding box overlap between successive frames may be low or even zero, which creates a challenge for re-recognition, for this reason, we use the Mahalanobis distance instead of the IoU and add the appearance information into the cost matrix; then, the new association cost formula is shown in Eq. 12. Mahalanobis distance is a distance metric based on a covariance matrix that takes into account the shape and size variations of targets and can more accurately characterize the similarity between targets when dealing with situations with uncertainty, overlapping targets, or nonlinear motion. The visual appearance information of a target is usually represented by feature vectors, which include color histograms, texture features, shape features, etc. Cosine distance is a method to measure the angle between two vectors in vector space that is used to determine whether their directions are similar or not. By calculating the cosine distance between target appearance information, the similarity between different targets can be measured, thus helping to determine the association relationship between targets. Where D_M and D_A calculate the Mahalanobis distance and appearance cosine distances between the detection and prediction frames, respectively, and is the weight factor; by adjusting the weighting coefficients, it is possible to change the weighting of the Mahalanobis distance and appearance cosine distances in the cost matrix. In Eq. 12, we set η to 0.6 and λ to 0.2. For the convenience of the following discussion, we name the improved algorithm Supplementary-OC SORT (SOC-SORT). Its structure is shown in Fig. 10.

IV. RESULTS

A. EXPERIMENTS ENVIRONMENT

The test operating environment is Ubuntu18.04.5 system, the CPU is Intel Core i7-12700K @5.0GHz; the running memory is 64GB, and the GPU is NVIDIA GeForce RTX3090 24GB. The software environment is Python 3.6.8, based on the Pytorch 1.10.0 architecture.

B. DATA DESCRIPTION

The video data used in the experimental part of this paper were provided by a mining group in Huainan, Anhui Province. Firstly, the video is processed by frame extraction to get 10,000 pictures containing operating personnel, and then the blurred and repeated pictures are screened to get 8,000 pictures. Finally, after frame classification of personnel by annotation software, the data set is divided according to the ratio of 7:2:1, and 5600 training sets, 1600 validation sets, and 800 test sets are obtained to get the data set MINE-TD used in the personnel detection part of this paper. Based on the above video data, we produced MINE-MOT, a personnel tracking dataset suitable for underground coal mines. MINE-MOT labeled the consecutive frame images with bounding boxes and categories by automatic annotation, and labeled the target ID by manual annotation. MINE-MOT contains multiple underground coal mine scenes, four training sequences and four test sequences, totaling about 4500 frame Images.

C. ANALYSIS OF TEST RESULTS

In order to evaluate the effectiveness of the image enhancement algorithm introduced in this research for enhancing features within coal mine underground images, a rigorous comparative analysis is necessary. We selected four images

TABLE 1. Comparative results of testing image enhancement algorithms on a single image.

Method	MV	AG	MSE	LOE	PSNR	SSIM	TC(CPU/GPU)
Gamma	**74.17**	18.11	4278.3	283.8	11.7	0.71	**0.02/0.01**
MSRCR	67.33	22.13	2577.5	227.5	13.4	0.83	0.42/∼
AHE	72.53	21.45	5391.2	**657.1**	17.2	0.66	0.07/0.02
IE	71.24	**24.89**	**2413.6**	317.2	**21.5**	**0.96**	0.09/0.02

TABLE 2. Model training hyperparameter settings.

hyperparameters	value	hyperparameters	value
epochs	10000	LR	0.001
batch-size	16	LR decay epoch	2000，10000
momentum	0.9	decay	0.0001

TABLE 3. Combined test results based on the MINE-TD-test dataset.

MINE-TD				
Model	Param/M	GFLOPs	FPS f/s	AP(0.5)
YOLOv7	36.5	103.2	52	80.17
YOLOv7+IE	36.5	103.2	41	85.65
YOLOv7+EDCA	38.2	112.7	49	86.61
YOLOv7+ADSA	37.4	108.9	48	83.07
YOLOv7+ADAM	39.1	118.4	44	88.77
YOLOv7+ADAM+IE	39.1	118.4	41	92.41

of different scenes from the dataset MINE-TD and the current mainstream algorithms for comparison tests respectively, and the test results are shown in Fig. 11.

In our experimental setup, we subjected the images to four distinct image processing methods: Gamma correction [36], MSRCR [37], AHE [38], and the algorithm proposed in this paper. Fig. 11 visually showcases the results of these processes. It is worth noting that while Gamma correction significantly enhances overall image brightness, it falls short in providing a well-defined contrast improvement strategy; MSRCR is an image enhancement algorithm based on the principle of Multi-Scale Retinex (MSR) [39], where the introduction of the color recovery mechanism for improving image performance under different lighting conditions. However, its high computational complexity and parameter dependency make it difficult to be applied in real-time processing tasks. The AHE algorithm is prone to the problem of local contrast imbalance when dealing with regions with low brightness and contrast, which may bring more noise.

In this paper, the algorithm combines the limiting contrast adaptive histogram equalization and adaptive illumination adjustment strategies to improve the overall brightness and contrast of the coal mine underground image while providing adaptive brightness adjustment for over-darkened and over-exposed regions. The comparison results of the four images show that the images are well-balanced in brightness and contrast and do not introduce too much noise, which boosts the subsequent target detection and tracking tasks. Concurrently, we introduced evaluation metrics such as mean value (MV), average gradient (AG), mean square error (MSE), luminance order error (LOE) [40], peak signal-to-noise ratio (PSNR), and structural similarity index metric (SSIM) to compare the above algorithms for the test. The experimental results are shown in Table 1.

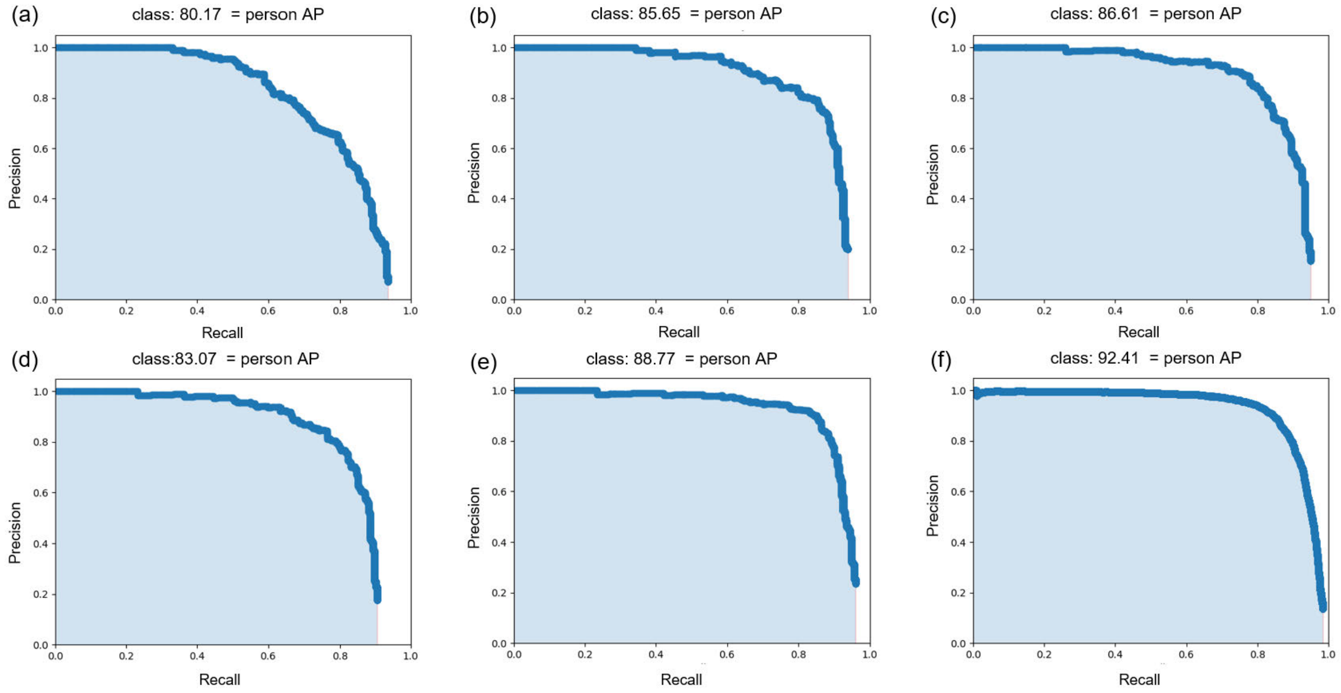

FIGURE 12. AP (0.5) plot obtained based on the combined strategy of Table 3. (a) YOLOv7; (b) YOLOv7+IE; (c) YOLOv7+EDCA; (d) YOLOv7+ADSA; (e) YOLOv7+ADAM; (f) YOLOv7+ADAM+IE.

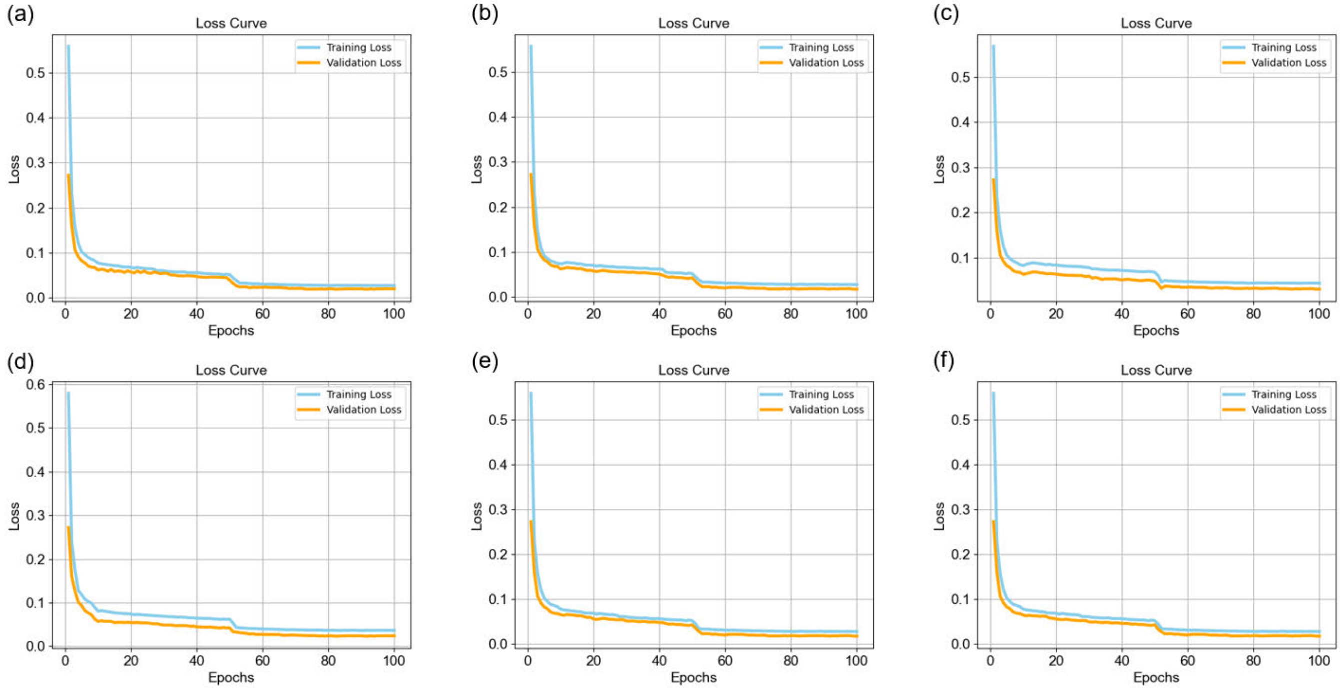

FIGURE 13. Based on the training loss map under the combined strategy in Table 3. (a) YOLOv7; (b) YOLOv7+IE; (c) YOLOv7+EDCA; (d) YOLOv7+ADSA; (e) YOLOv7+ADAM; (f) YOLOv7+ADAM+IE.

Time complexity (TC) is used to reflect the processing speed of the image enhancement algorithm. From Table 1, it can be seen that Gamma has the best effect for image brightness enhancement, but the clarity is not good; the MSRCR algorithm has more balanced indexes in general but has higher computational complexity due to its multi-scale Gaussian filtering operation. IE algorithm is an optimization algorithm based on AHE, which achieves the best of three

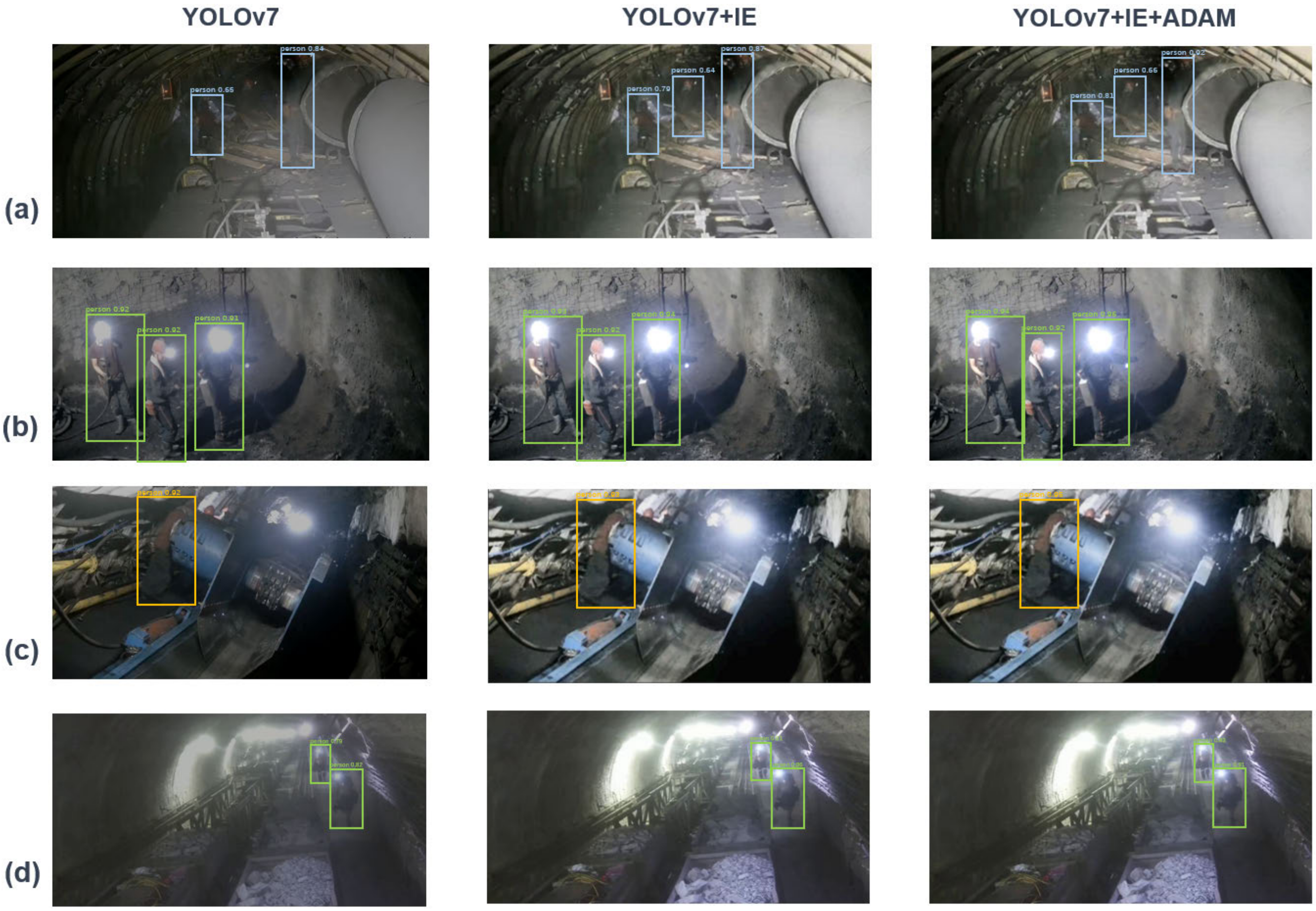

FIGURE 14. Detection effect of different combinations of algorithms in coal mine underground scene.

indexes with a small amount of increase in time complexity. Combining subjective visual evaluation and objective index evaluation, the IE algorithm has certain advantages in the coal mine underground image enhancement processing task.

We conducted a comprehensive comparative evaluation using the YOLOv7 target detection network on the MINE-TD dataset to assess the efficacy of the ADAM module introduced in this article. Detailed information about the test hyperparameters and the outcomes of the ablation experiments can be found in Table 2 and 3. The precision-recall curves and loss curves under different combination strategies based on Table 3 are shown in Fig. 12 and Fig. 13, respectively.

To objectively evaluate the performance of the model, we introduce evaluation metrics such as the number of parameters (Param), Giga Floating Point Operations per Second (GFLOPs), the Frame rate Per Second (FPS), and the Average Precision (AP) to reflect the equilibrium relationship between the time complexity and the space complexity of different combination strategies. The parametric quantity is the sum of all parameters in the model, which is closely related to the size of the disk space occupied by the model and can be used to measure the spatial complexity of the model. GFLOPs refer to the billions of floating-point calculations that the model can accomplish per second, which is used to measure the time complexity of the model; FPS refers to the number of image frames that the model can process per second.

We introduced IE in the test set validation phase of the model training stage and accelerated the image preprocessing based on the IE method using the CUDA module and parallel computing, which improved the AP (0.5) by 5.48% compared to the original YOLOv7 without increasing the model parameters, but at the same time, the IE method limited the maximum FPS of the model. We split the ADAM module and performed a combination of experiments, but of course, the location where they were added to the network remained unchanged. The EDCA module enhances the feature extraction capability of the network while directing the network's attention to channels that contain more information content, dramatically enhancing the model's AP value with a small reduction in FPS. The ADSA module adaptively adjusts the pooling strategy for both the over-dark and overexposed regions to highlight the target region, again achieving a 3% improvement in AP value. The ADSA module adaptively adjusts the pooling strategy for over-dark and overexposed regions to highlight the target regions and also

TABLE 4. Results of comparative tests of different algorithms on the MINE-TD-test dataset.

MINE-TD				
Model	Param/M	GFLOPs	AP(0.5)	FPS f/s
YOLOX-X	99.1	281,9	81.74	26
YOLOv5x	86.7	205.9	81.33	33
YOLOv7-x	71.3	189.9	82.17	39
Faster-RCNN	84.6	200.1	86.58	13
YOLOv7+CBAM	36.6	103.6	83.49	48
YOLOv7+ADAM	39.1	118.4	88.77	44
YOLOv7+ADAM+IE	39.1	118.4	92.41	41

TABLE 5. Results of an ablation study of the MINE-MOT-val set.

MINE-MOT							
OC-SORT	IE	ADAM	SOC-SORT	HOTA↑	IDF1↑	AssA↑	MOTA↑
√				62.2	58.2	45.7	85.6
√	√			63.6	60.1	46.1	86.1
√	√	√		65.2	61.2	46.5	88.3
√	√	√	√	67.4	63.3	47.1	91.9

TABLE 6. Results of an ablation study of the MOT17-val set.

MOT17							
OC-SORT	IE	ADAM	SOC-SORT	HOTA↑	IDF1↑	AssA↑	MOTA↑
√				63.7	77.5	63.6	78.0
√	√			64.6	79.1	64.1	78.7
√	√	√		64.9	79.2	65.5	79.0
√	√	√	√	65.1	81.2	66.1	79.6

achieves a 3.1% increase in AP value. The ADAM module consists of the EDCA module and ADSA module connected serially to enhance the early expression of target features in the shallow network. On the coal mine underground dataset MINE-TD, the combined YOLOv7+ADAM strategy obtains an 8.6% improvement in the AP value. The combined YOLOv7+ADAM strategy combined with the IE method has a significant advantage over the original YOLOv7, even though the FPS is limited to 41 f/s. but the AP value is 92.41%, which is a significant advantage over the original YOLOv7.

To visualize the advanced detection effect of different combination algorithms, we conducted detection tests on four different scenarios of coal mines underground, and the detection effect is shown in Fig. 14.

As can be seen from the row of Fig. 14(a), for this dark scene, the IE algorithm effectively improves the overall brightness and contrast of the image, so that the target is separated from the background and thus recognized by the detector. Benefitting from the enhanced image quality, the YOLOv7 network, augmented with the ADAM module, demonstrates an increased capacity to prioritize significant channels and regions. Consequently, this augmentation contributes to an overall elevation in target confidence levels. Observing rows (b), (c), and (d) in Fig. 14, it becomes evident that the detector's performance experiences significant improvements across a range of lighting conditions and light/dark distribution scenarios. This enhancement underscores its versatility and suitability for a variety of environmental settings.

To demonstrate the superiority of the proposed algorithms in targeting the task of target detection in underground coal mine images, we conduct a comparison test of different algorithms on the self-constructed dataset MINE-TD. We introduced single-stage and two-stage target detection algorithms, including YOLOX [41], YOLOv5x, YOLOv7-x, Faster-RCNN [4], etc., as well as the YOLOv7 algorithm inserted with a CBAM module, respectively. The experimental results are shown in Table 4.

We selected the extra-large parameter versions of the YOLO-X, YOLOv5x, and YOLOv7-x series of models to obtain higher accuracy. As can be seen from the comparison tests in Table 4, the attention mechanism can substantially improve the accuracy of the model with lower model parameters, while achieving the advantage of detection speed. Among all the test results, the YOLOv7+ADAM+IE combination strategy achieves optimal performance in terms of accuracy.

Although OC-SORT strives to address the limitations of the Kalman filter-based tracker SORT, i.e., the loss of tracking accuracy due to low-accuracy detectors, occlusions, or fast nonlinear motions, the positional accuracy of the detectors is still the basis for stable tracking by the tracker. SOC-SORT inherits the overall process of OC-SORT with the addition of appearance-matching information. In order

TABLE 7. Comparative experimental results based on the MINE-MOT-test dataset and the MINE-MOT-test dataset enhanced by the IE algorithm (blue textured section).

MINE-MOT							
Algorithm	HOTA↑	MOTA↑	DetA↑	AssA↑	IDF1↑	AssRe↑	FPS↑
FairMOT [43]	47.3	82.2	56.4	21.2	35.8	63.6	26
TransTrk [44]	45.2	83.7	62.3	23.6	36.8	57.2	10
MOTR [45]	50.2	78.6	61.1	40.2	48.7	59.2	-
TransMOT [46]	52.2	83.5	67.7	41.3	51.5	66.4	9
SORT	44.5	85.8	65.9	33.9	50.2	49.1	41
DeepSORT	55.2	87.7	67.2	37.2	48.8	50.2	14
ByteTrack [47]	57.1	87.4	67.7	36.8	49.7	69.0	27
StrongSORT [48]	59.7	89.8	72.1	38.4	51.8	-	-
OC-SORT	57.2	87.6	74.9	41.7	56.2	63.2	27
Deep OC-SORT [49]	61.3	91.7	78.4	45.9	58.0	65.1	17
Dark-SORT	67.4	92.6	80.3	46.8	61.7	65.7	23
FairMOT [43]	49.7	84.1	59.4	21.7	37.4	64.9	26
MOTR [45]	51.6	80.7	64.1	43.1	51.6	59.9	-
TransMOT [46]	53.8	85.1	67.8	42.4	53.2	66.8	9
ByteTrack [47]	58.7	89.2	69.4	37.1	51.5	69.3	27
Deep OC-SORT [49]	63.6	91.9	79.7	46.2	59.4	65.3	17
Dark-SORT	67.4	92.6	80.3	46.8	61.7	65.7	23

TABLE 8. Comparative experimental results based on the MOT17-test dataset and the MOT17-test dataset enhanced by the IE algorithm (blue textured section).

MOT17					
Algorithm	HOTA↑	MOTA↑	AssRe↑	AssA↑	IDF1↑
FairMOT [43]	59.3	73.7	63.6	58.0	72.3
MOTR [45]	57.2	71.9	59.2	55.8	68.4
TransMOT [46]	61.7	76.7	66.5	59.9	75.1
ByteTrack [47]	63.1	80.3	68.2	62.0	77.3
Deep OC-SORT [49]	64.9	79.4	70.1	65.9	80.6
Dark-SORT	65.4	77.9	71.7	63.8	78.2
FairMOT [43]	60.1	73.9	63.4	58.8	73.9
MOTR [45]	56.9	72.3	60.3	55.3	68.7
transMOT [46]	62.1	77.1	66.7	60.4	75.5
ByteTrack [47]	63.4	80.8	68.4	62.1	76.8
Deep OC-SORT [49]	65.7	79.3	70.4	65.6	80.9
Dark-SORT	65.4	77.9	71.7	63.8	78.2

to verify the enhancement effect of different combination strategies on the OC-SORT algorithm, we conducted ablation experiments (based on the same detector, YOLOv7) on the datasets MINE-MOT and MOT17 [42], respectively, and the results are shown in Table 5 and 6.

Our findings indicate that integrating both Image Enhancement (IE) and the ADAM module leads to enhanced tracking algorithm accuracy, while the inclusion of appearance information notably improves identity-matching accuracy. For the convenience of the following discussion, we name the combination of IE+ADAM+SOC-SORT as Dark-SORT. We performed comparative tests between the state-of-the-art tracking algorithms and Dark-SORT on the MINE-MOT and MOT17 datasets based on the evaluation metrics of HOTA, MOTA, IDF1, DetA, AssA, AssRe, and FPS (based on the same detector YOLOv7), respectively.

In addition, considering the generality of the algorithms and in order to validate the algorithms on a multi-target tracking dataset, we performed comparative experiments on MINE-MOT and MOT17 after image enhancement of all baseline algorithms. The test results are shown in Table 7 and 8.

The Higher Order Tracking Accuracy (HOTA) metric serves as a crucial evaluation measure for assessing the performance of multi-object tracking algorithms. While traditional tracking evaluation metrics focus only on the tracking of a single object, HOTA is more comprehensive as it

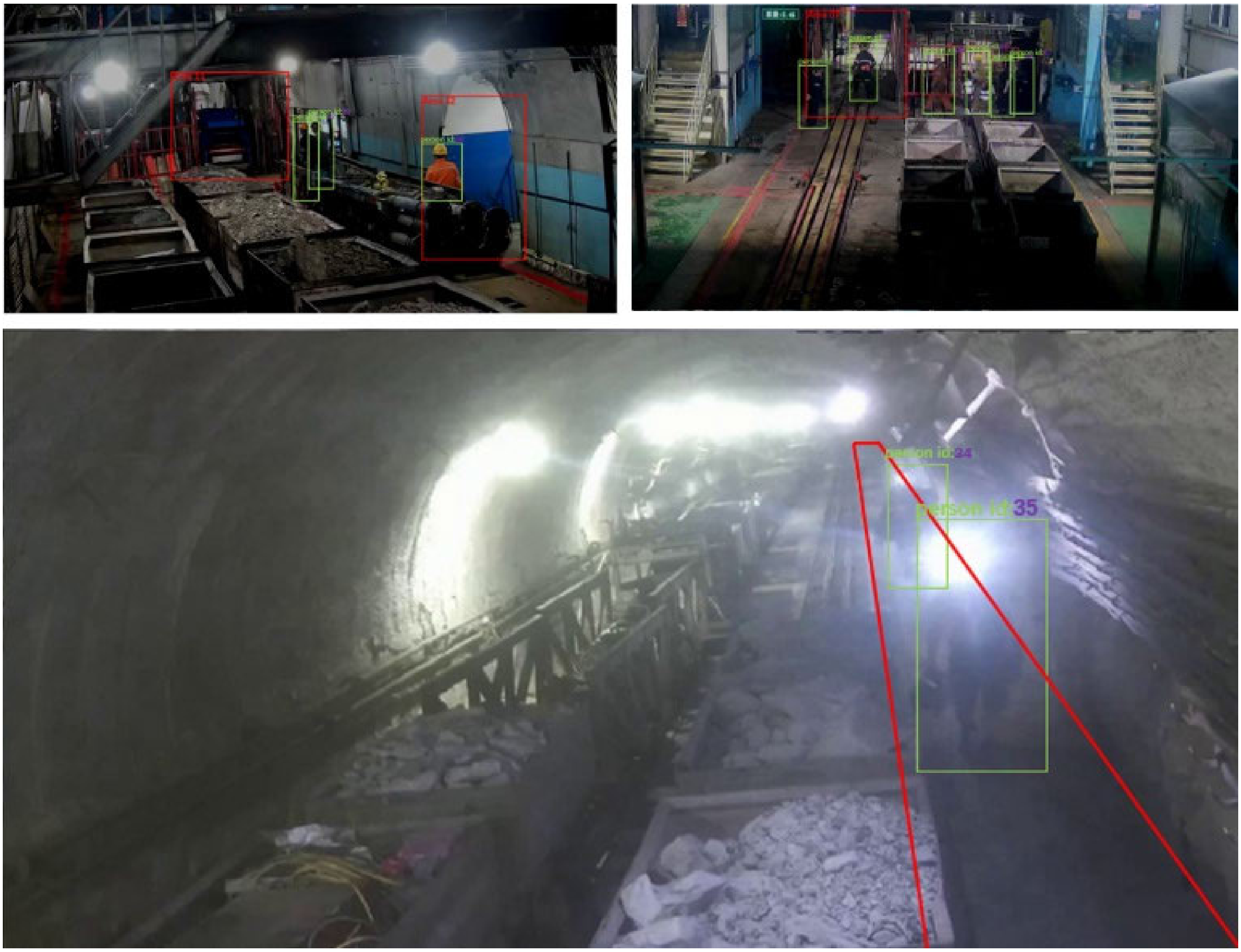

FIGURE 15. **Tracking effect graph of Dark-SORT in a real application.**

considers both localization and identity assignment accuracy. The Multi-Object Tracking Accuracy (MOTA) metric is a widely adopted measure for assessing the performance of multi-object tracking algorithms. It measures the accuracy of the tracking system by considering various factors such as false positives, false negatives, and identity conversion. Detection-based tracking Accuracy (DetA) offers a holistic evaluation of tracking accuracy, taking into account both the critical elements of detection and tracking. AssA is a measure of association accuracy. IDF1 combines precision and recall to provide a more comprehensive performance evaluation. Association Recall Error (AssRe) measures whether the tracker is able to correctly associate the tracked trajectory with the real target. For re-identification purposes, we employed the SBS50 model available in the fast-reid [50] library. Dark-SORT combines OC-SORT, appearance information, and the method designed in this paper to solve the target tracking problem in a dark-light environment.

By analyzing Table 7 and 8, it can be seen that Dark-SORT accounts for the optimal scores for all metrics in the MINE-MOT dataset. On MOT17, several indicators are evaluated optimally, which shows that Dark-SORT can be generalized to similar problems. As can be seen from Table 8, the Dark-SORT algorithm has good accuracy and tracking stability on multi-class object tracking datasets with wide generalization. The practical application effect of Dark-SORT is shown in Fig. 15.

V. DISCUSSION

Computer vision, an advanced technology, finds extensive industrial applications for automating target detection, recognition, and analysis by examining image and video data. Computer vision plays a pivotal role in the coal industry, particularly in enhancing safety monitoring, optimizing personnel management, and facilitating efficient accident response. However, limited by the low light, high dust, and point light sources in underground coal mines, detection, and tracking for underground targets in coal mines is still a severe challenge.

In this paper, we perform an in-depth analysis of the distinctive features within underground coal mine images while also reviewing relevant research in the field. The research on tracking-by-detection is carried out in terms of image enhancement, enhanced network feature extraction capability, and an improved motion model, respectively, which is dedicated to improving the accuracy of target detection and tracking algorithms in underground coal mines. Firstly, we improve the picture imaging quality by enhancing contrast, brightness, and illumination equalization under the demand of real-time video processing. Secondly, we analyze the attributes present in underground coal mine images through the visualization of the network's output layer. Additionally, we scrutinize the constraints and shortcomings of existing algorithms; based on this, we have devised novel enhancement techniques based on discrete pooling and introduced weighted attention mechanisms to enhance the network's feature extraction capabilities. Finally, we introduce OC-SORT and add appearance feature information on this basis to improve tracking stability and recognition ability. We performed individual comparison tests for each method, using the MINE-TD and MINE-MOT datasets to validate the viability and efficacy of our proposed methods.

The code and models are available at https://github.com/RWAUST123/DarkSORT-main123.

VI. CONCLUSION

The conclusions were as follows:

- The proposed video image enhancement method applies to coal mine underground scenes. While considering the real-time nature of the algorithm, the imaging quality is improved by adjusting the image brightness, contrast, and light equalization, which improves the detector's accuracy.
- In the coal mine underground scenario, the detection accuracy of the detector (YOLOv7+ADAM) network with the addition of the ADAM module is 88.77%, while the detection accuracy of the YOLOv7 network is 80.17%. The ADAM module, designed for coal mine underground scenarios and shallow network feature expression, has good performance.
- The introduction of the OC-SORT algorithm and the addition of appearance information as a supplement to re-identification improves the stability of person tracking. Through experiments on the MINE-MOT dataset, we verify the feasibility and superiority of the whole algorithm. Through a series of experiments on the publicly available dataset MOT17, we verify the applicability of the algorithm.

Looking ahead to the next phase of our research, we will concentrate on tackling the unique challenges the coal mine operating environment poses. These challenges include low visibility and difficulty re-identifying personnel due to their similar attire.

REFERENCES

[1] Y. Pang, Y. Yuan, X. Li, and J. Pan, "Efficient HOG human detection," *Signal Process.*, vol. 91, no. 4, pp. 773–781, Apr. 2011.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[3] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[5] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[7] C.-Y. Wang, A. Bochkovskiy, and H. M. Liao, "Scaled-YOLOv4: Scaling cross stage partial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13024–13033.

[8] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[9] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, vol. 16, no. 16, 2020, pp. 107–122.

[10] W. Wang, X. Yuan, X. Wu, and Y. Liu, "Fast image dehazing method based on linear transformation," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1142–1155, Jun. 2017.

[11] S. Park, K. Kim, S. Yu, and J. Paik, "Contrast enhancement for low-light image enhancement: A survey," *IEIE Trans. Smart Process. Comput.*, vol. 7, no. 1, pp. 36–48, Feb. 2018.

[12] Y.-F. Wang, H.-M. Liu, and Z.-W. Fu, "Low-light image enhancement via the absorption light scattering model," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5679–5690, Nov. 2019.

[13] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic photo adjustment using deep neural networks," *ACM Trans. Graph.*, vol. 35, no. 2, pp. 1–15, May 2016.

[14] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.

[15] S. Park, S. Yu, M. Kim, K. Park, and J. Paik, "Dual autoencoder network for retinex-based low-light image enhancement," *IEEE Access*, vol. 6, pp. 22084–22093, 2018.

[16] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2016, pp. 3464–3468.

[17] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2017, pp. 1–6.

[18] Z. Y. Wang, D. Q. Miao, C. R. Zhao, S. Luo, and Z. H. Wei, "A pedestrian tracking algorithm based on multi-granularity feature," *IEEE Trans. Comput.*, vol. 57, no. 6, pp. 996–1002, 2020.

[19] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1218–1225.

[20] C. Ma, C. Yang, F. Yang, Y. Zhuang, Z. Zhang, H. Jia, and X. Xie, "Trajectory factory: Tracklet cleaving and re-connection by deep Siamese bi-GRU for multiple object tracking," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6.

[21] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[22] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "POI: Multiple object tracking with high performance detection and appearance feature," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 36–42.

[23] G. Li, X. Chen, M. Li, W. Li, S. Li, G. Guo, H. Wang, and H. Deng, "One-shot multi-object tracking using CNN-based networks with spatial-channel attention mechanism," *Opt. Laser Technol.*, vol. 153, Sep. 2022, Art. no. 108267.

[24] X. Chen, Y. Jia, X. Tong, and Z. Li, "Research on pedestrian detection and DeepSort tracking in front of intelligent vehicle based on deep learning," *Sustainability*, vol. 14, no. 15, p. 9281, Jul. 2022.

[25] F. Gu, J. Lu, and C. Cai, "RPformer: A robust parallel transformer for visual tracking in complex scenes," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[26] F. Gu, J. Lu, and C. Cai, "A robust attention-enhanced network with transformer for visual tracking," *Multimedia Tools Appl.*, vol. 82, no. 26, pp. 40761–40782, Nov. 2023.

[27] J. Zhang, H. Liu, Y. He, L.-D. Kuang, and X. Chen, "Adaptive response maps fusion of correlation filters with anti-occlusion mechanism for visual object tracking," *EURASIP J. Image Video Process.*, vol. 2022, no. 1, p. 4, Dec. 2022.

[28] F. Gu, J. Lu, C. Cai, Q. Zhu, and Z. Ju, "Repformer: A robust shared-encoder dual-pipeline transformer for visual tracking," *Neural Comput. Appl.*, vol. 35, no. 28, pp. 20581–20603, Oct. 2023.

[29] J. Wang, S. Xuan, H. Zhang, and X. Qin, "The moving target tracking and segmentation method based on space-time fusion," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 12245–12262, Mar. 2023.

[30] D. Yuan, X. Chang, Q. Liu, Y. Yang, D. Wang, M. Shu, Z. He, and G. Shi, "Active learning for deep visual tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[31] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. VLSI Signal Process.-Syst. Signal, Image, Video Technol.*, vol. 38, no. 1, pp. 35–44, Aug. 2004.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[33] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[34] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[35] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9686–9696.

[36] S.-C. Huang, F.-C. Cheng, and Y.-S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1032–1041, Mar. 2013.

[37] D. J. Jobson, "Retinex processing for automatic image enhancement," *J. Electron. Imag.*, vol. 13, no. 1, p. 100, Jan. 2004.

[38] T. K. Kim, J. Ki Paik, and B. S. Kang, "Contrast enhancement system using spatially adaptive histogram equalization with temporal filtering," *IEEE Trans. Consum. Electron.*, vol. 44, no. 1, pp. 82–87, Feb. 1998.

[39] Z. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-scale retinex for color image enhancement," in *Proc. 3rd IEEE Int. Conf. Image Process.*, vol. 3, Sep. 1996, pp. 1003–1006.

[40] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, Sep. 2013.

[41] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[42] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*.

[43] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3069–3087, Nov. 2021.

[44] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: Multiple object tracking with transformer," 2020, *arXiv:2012.15460*.

[45] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-end multiple-object tracking with transformer," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 659–675.

[46] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "TransMOT: Spatial-temporal graph transformer for multiple object tracking," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4859–4869.

[47] Y. Zhang, "ByteTrack: Multi-object tracking by associating every detection box," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 1–21.

[48] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "Strong-SORT: Make DeepSORT great again," *IEEE Trans. Multimedia*, to be published.

[49] G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani, "Deep OC-SORT: Multi-pedestrian tracking by adaptive re-identification," 2023, *arXiv:2302.11813*.

[50] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A PyTorch toolbox for general instance re-identification," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 9664–9667.

RUI WANG received the bachelor's degree in communication engineering from the College of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, China, in 2021. He is currently pursuing the master's degree in intelligent manufacturing engineering with the School of Artificial Intelligence, Anhui University of Science and Technology. His current research interests include computer vision and artificial intelligence.

JINGZHAO LI received the M.A. degree from the China University of Mining and Technology, in 1992, and the Ph.D. degree from the Key Laboratory of Power Electronics and Power Drives, Hefei University of Science and Technology, in 2003. He is currently a Professor with the School of Electrical Information and Engineering, Anhui University of Science and Technology, China. He has published more than 100 papers in domestic and international academic journals and conference proceedings. His research interests include computer control, the Internet of Things technology, and embedded systems. These papers are embodied more than 60 times by SCI and EI and are cited more than 100 times by others.

ZHI XU received the bachelor's degree in electrical engineering from the College of Electrical and Information Engineering, Anhui University of Science and Technology (AUST), Huainan, China, in 2018, where he is currently pursuing the Ph.D. degree in mechanical engineering. His current research interests include edge computing and artificial intelligence.