J. Meas. Sci. Instrum., 2025, 16(1): 128-141. http://jmsi.publish.foundersss.cn
DOI: 10.62756/jmsi.1674-8042.2025013 jmsi@nuc.edu.cn

A novel email-based smart remote image surveillance camera

LIU Xinhao[1], MENG Lingjun[1*], LIU Feng[1], ZHOU Xiaotong[2], WANG Jiacheng[2]

1. School of Instrument and Electronics, North University of China, Taiyuan 030051, China;
2. School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China

***Corresponding author:** MENG Lingjun (menglingjun@nuc.edu.cn)

Received: August 8, 2024 **Revised:** November 3, 2024 **Accepted:** November 11, 2024

Abstract: Aiming at the problems of difficult deployment and access of surveillance system server, as well as high operation and maintenance cost, a remote surveillance camera is designed based on RK3566 chip, which is controlled and transmits data via email platform. Firstly, to address the impact of environmental factors such as weather and light on image quality, a deep neural network (DNN) image exposure correction network is employed to rectify images with abnormal exposure. Additionally, a back propagation (BP) neural network is utilized to fit a curve relating the brightness difference to the gamma value of images before and after exposure correction, thereby adjusting the gamma value of the camera. Secondly, to enhance the precision of YOLOv5 algorithm in differentiating between anomalies in nighttime imagery, infrared image data are employed, and a context-aware light-weight label assignment head and coordinate attention mechanism are incorporated into the model to augment the model's detection accuracy and recall rate for small targets. Furthermore, to meet the demand for reporting of abnormal situations in unattended environments, an automatic target identification and reporting process has been designed which combines YOLOv5 algorithm with the frame-difference motion detection algorithm. The camera has been tested for compatibility with the current mainstream commercial email platforms. The mean time required for transmitting a single image file via the email platform is less than 10 s, while the mean time for transmitting a short video is less than 60 s. The BP network's average training loss is 0.015, and the average testing loss is 0.013, which basically meets the precision requirements for gamma adjustment. The improved YOLOv5 algorithm achieved an mAP@0.5 of 91.5% and a recall rate of 85.5%, effectively enhancing the accuracy of small object detection.

Key words: email transmission; exposure correction; back propagation (BP) neural network; gamma value; YOLOv5

0 Introduction

As the level of automation and intelligence in industrial production environments continues to increase, the role of remote surveillance systems becomes increasingly significant. Through the application of various image processing technologies[1] and target detection algorithms[2,3], the surveillance camera can achieve target detection and automatic reporting function in an unmanned environment. The design of surveillance cameras is trending towards miniaturization and intelligence. Miniaturization leads to reduced power consumption. Magno et al.[4] proposed a multimodal wireless smart camera equipped with a pyroelectric infrared sensor and a solar energy harvester. This self-powered camera can operate perpetually in an outdoor scenario. Also to reduce hardware power consumption during video encoding, Liu et al.[5] designed a video encoder for surveillance systems consuming only 41 mW. Additionally, intelligence corresponds to the integration of various artificial intelligence (AI) algorithms, such as you only look once (YOLO)[6], MobileNet[7], and region convolutional neural network (RCNN)[8] for image recognition and target detection. Among them, the YOLO algorithm offers advantages such as high detection speed and accuracy, making it suitable for deployment on embedded devices[9,10], therefore, it becomes the most widely used target recognition algorithm in the surveillance domain. For instance, Wei et al.[11] applied HD-YOLO algorithm to fisheye camera surveillance devices, and design the radius-aware loss function to adapt the impact of fisheye distortion. Abdusalomov et al.[12] employed the improved YOLOv3 algorithm for fire monitoring systems, and deployed the algorithm on an embedded device. Khazukov et al.[13,14] utilized YOLO algorithm in traffic cameras to analyze traffic flow. Similarly, YOLO algorithm can also be used for vehicle

speed detection in the traffic monitoring[15]. Besides target detection algorithms, surveillance cameras are generally equipped with motion detection algorithms such as frame differencing[16], and RetinaNet[17].

In summary, current designs of remote surveillance cameras generally exhibit the following characteristics:

1) Streaming media and cloud services: Video and image data from surveillance systems are transmitted via streaming servers or cloud servers and can be viewed through specific applications.

2) High definition: Modern surveillance cameras typically support high resolutions and automatic exposure, providing clear image quality. And many cameras offer low-light capabilities, enabling clear video capture in low-illumination environments, with some also featuring infrared night vision.

3) AI algorithms: Built-in motion detection algorithms can identify and record moving objects within the frame and trigger alerts. Utilizing AI, these cameras can recognize specific objects.

4) Automatic reporting: Based on the analysis of surveillance content, the system can automatically report any anomalies such as unauthorized personnel entry or the presence of smoke and fire on the premises.

Although the use of streaming services or cloud servers for video data transmission in surveillance cameras ensures real-time data availability, it leads to elevated operation and maintenance costs. In remote industrial environments with lower real-time requirements (e. g., unmanned power stations), this complicates equipment deployment and functional expansion instead.

For the above issues, this paper presents the design of a smart remote surveillance camera based on email platform, which does not require the construction of a dedicated server or the submission of a public IP application. Instead, it utilises an ordinary email for both control and data transmission. Furthermore, it is compatible with numerous commercial email platforms, which can be accessed at any time or place, making the usage scenarios more flexible. Additionally, the security of commercial email platforms is typically superior to that of self-hosted servers. The utilization of such platforms can guarantee data security while concurrently reducing server operation and maintenance costs.

In the design of remote surveillance cameras, this research focuses on addressing the following practical issues:

1) In the industrial environment, the surveillance camera is typically situated in locations distant from the target area, such as the eaves of a building or a high tower. As a result, the target often occupies a small portion of the overall image, which places high demands on the algorithm's capability for detecting and localizing small targets.

2) The improper setting of exposure parameters is the primary cause of ineffective image results when the outdoor surveillance camera captures images. If the exposure time or gamma is not set correctly, the resulting image will be either too bright or too dark overall. Moreover, in practical applications, the automatic exposure mode of the camera either responds slowly and takes a long time to adjust, or the automatically adjusted exposure effect is suboptimal.

3) It is of particular importance to control power consumption in the context of solar-powered remote surveillance camera. However, in the context of the RK3566 platform, the target detection algorithm requires invoking the neural network processing unit (NPU) for acceleration[18], which necessitates the utilisation of additional hardware resources and results in increased power consumption. Therefore, the start timing and duration of the target detection algorithm need to be strictly controlled.

Above all, the main contributions of this research are as follows:

1) A remote surveillance camera based on email system has been designed, wherein a device-side program monitors the email server to automatically parse control command emails as well as compose and return image data emails.

2) An image exposure correction process has been developed, along with a self-constructed dataset comprising image brightness differences and camera gamma value for back propagation (BP) neural network fitting. Initially, the brightness of abnormally exposed images is corrected using a deep neural network (DNN) exposure correction network, with the corrected images serving as reference to calculate the brightness difference relative to the original images. Subsequently, the camera gamma value is adjusted using a BP neural network that fits the brightness difference and gamma value, thereby obtaining normally exposed image data.

3) The YOLOv5 algorithm has been improved, adding the context-aware light-weight label assignment head (CLLAHead) detection head to the YOLOv5 algorithm to improve the detection of small targets, and enhances the coordinate attention mechanism in the backbone network to improve the detection accuracy and to correct the localisation errors.

4) An automatic target identification and reporting process has been designed which combines the YOLOv5 algorithm with the frame-difference motion detection

algorithm. Upon detecting the motion target, the YOLOv5 algorithm is invoked to identify the target within the image data, and the NPU resources are released immediately after the identification. Subsequently, the device determines whether to establish a network connection and transmit the image data back based on the target recognition results.

The remaining sections are organized as follows. Section 1 introduces the design concept and methodology of the surveillance camera, including the method of email transmission, the image exposure correction process, the improved YOLOv5 algorithm, and the automatic target identification and reporting process. The experimental results are shown in Section 2. Conclusions are presented in Section 3.

1 Design methods

The surveillance camera is controlled by the RK3566 chip. It is mainly composed of a camera, a solar charging module, and a 4G network module. The device receives email control information and returns image data emails through the network module. The overall block diagram is shown in Fig.1.

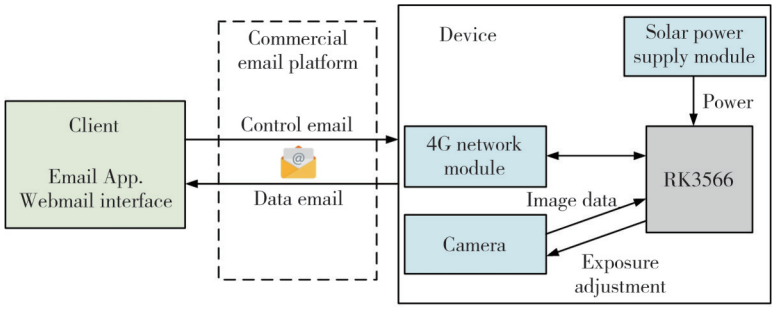


Fig. 1 Overall block diagram

In the software component, the surveillance camera is based on Ubuntu 22.04 system with kernel version 5.10. The primary function is to monitor the email server and automatically parse the commands contained within emails when new emails are sent from the specified email address. Subsequently, it executes the corresponding functions based on these commands, such as capturing snapshots or recording short videos. Upon completion of these tasks, the acquired data is automatically compiled into an email as attachment and sent back to the client's email address. The overall workflow is illustrated in Fig.2.

Process 1: Software design concept

```
while Listening the email server do
    if Receive new email then
        Parsing the command in email
        thread func
            Perform the corresponding function
        thread reply
            End of function execution then
                Composition reply email
                Send to the client email address
```

Fig. 2 Software design concept

In the "corresponding functions" section of the process, there are operations that significantly consume software and hardware resources, mainly including the image exposure correction and the target detection function with improved YOLOv5 algorithm. The process is executed using a multi-threaded approach[19], whereby distinct threads are bound to discrete CPU cores through affinity settings, as illustrated in Fig. 3. This approach is employed to prevent the overloading of shared hardware resources by concurrent tasks.

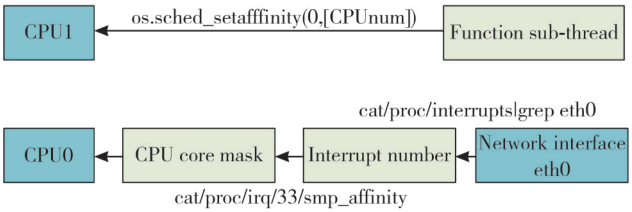


Fig. 3 Affinity setting

1.1 Email transmission

Email transfer function relies on Internet message access protocol (IMAP) and simple mail transfer protocol (SMTP)[20], The IMAP is employed to search for new emails, retrieve and parse control commands within the emails. The SMTP is used to add images or short videos as attachments to the returned emails and to send the emails. The email transmission process is illustrated in Fig.4.

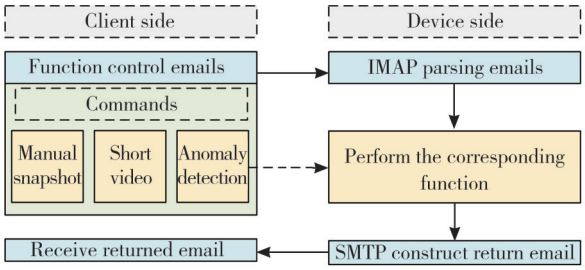


Fig. 4 Flow chart of email transmission

Two mailboxes are employed in this design, the mailbox at the device side is utilized to receive user commands. The mailbox at the client side is used to control the remote camera and receive feedback data. In practical tests, the total number of emails in the inbox will impact the time required for the programme to scan the inbox email list. Therefore, used emails in device side mailbox are deleted at the conclusion of each function execution cycle. The specific flow of email parsing and data email return is illustrated in Fig.5. The program requires the email account and the specific licence to log in and monitor the IMAP and SMTP servers. Furthermore, the device can be managed hierarchically through email accounts with different levels of permissions. This method facilitates the maintenance of remote cameras while ensuring data security.

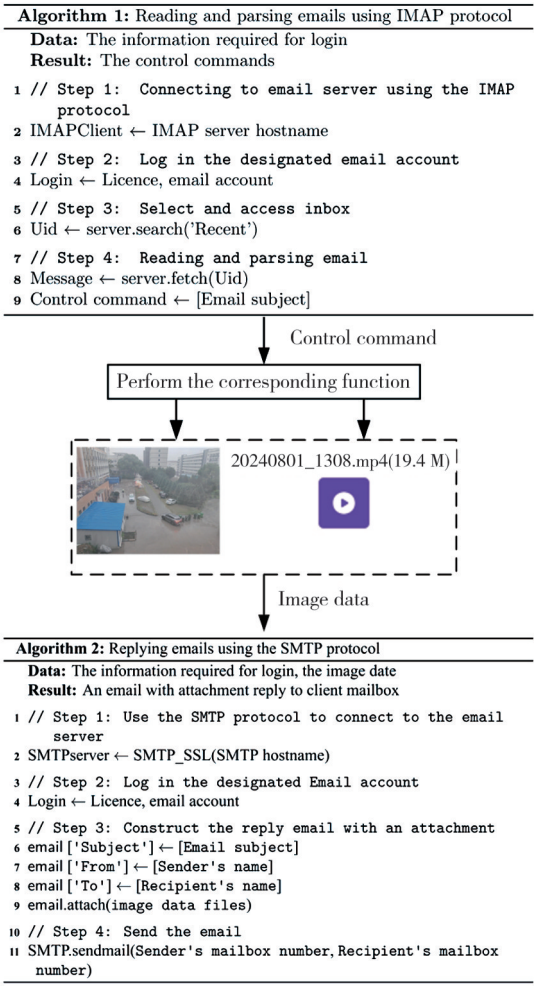


Fig. 5 Email parsing and feedback process

1.2 Image exposure correction

In order to guarantee optimal image quality, it is essential to undergo a preliminary processing stage prior to the recognition and transmission functions being carried out on the image. The primary cause of suboptimal image quality in outdoor surveillance camera is the utilisation of inadequate exposure settings. Digital cameras can adjust the exposure value to change the brightness of the image. This exposure value can be either manually controlled or automatically adjusted in auto exposure (AE) mode. In the latter case, the camera measures the amount of light received from the scene through-the-lens (TTL) and then compensates for any fluctuations in brightness levels by adjusting the exposure value (EV) accordingly[21]. Exposure errors may be attributed to a multitude of factors, including lens metering errors, inadequate lighting conditions, significant fluctuations in scene brightness levels, and operator errors in manual mode. These errors are introduced at the outset of the acquisition process and subsequently by the camera's image signal processor (ISP), which generates the RGB image in a non-linear manner, rendering it difficult to rectify them in the final generated image[22]. While current cameras basically have an automatic exposure mode, however, the camera's auto exposure mode tends to be slow and ineffective.

In order to address this issue, this research employs an image exposure correction network that incorporates a Laplacian pyramid and a DNN[23]. Then, the camera's gamma value is adjusted using a curve fitted by a BP neural network. The overall flow of image preprocessing for exposure correction is shown in Fig.6.

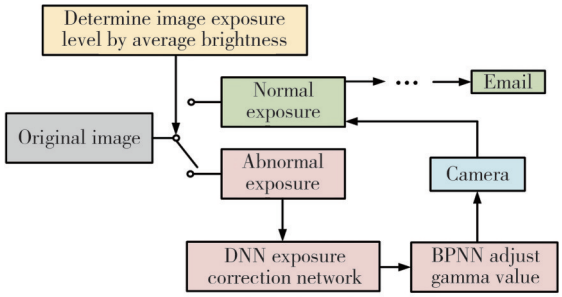


Fig. 6 Image exposure correction process

In this process, the DNN-corrected images serve as intermediate data for luminance difference reference. And based on this reference, BP neural network adjusts the gamma value, bringing image exposure back to normal level.

The DNN initially repairs the global color information and subsequently enhances the image details. The network is divided into multiple scales, each containing a sub-network for processing images of different resolutions to more effectively capture global and local exposure information. Next, the feature information at different scales is fused and the image size is restored by up-sampling operation of the transposed convolutional layer. Finally, the feature map is converted into an exposure-restored image by the convolutional layer.

However, this network introduces additional noise, resulting in a slight degradation of image quality. Consequently, the DNN-corrected images are not directly transmitted to the client as final images but serve only as intermediate data for reference.

After the process of utilising the DNN network for rectifying anomalies in image exposure, the discrepancies between the repaired and unrepaired images are employed to facilitate the adjustment of the camera's exposure parameters. The parameters that exert a significant influence on the camera's exposure effect are typically the exposure time and gamma value. The gamma value is a parameter that controls image brightness and contrast. It is used to adjust the total response curve of an image to align with the nonlinear visual perception of the human eye. Initially, the exposure time is adjusted based on auto exposure to roughly determine the range of exposure levels. Subsequently, the repaired and unrepaired images are transformed from the BGR color space to the YUV color space, and the luminance difference in the Y-channel is calculated. A self-constructed dataset comprising gamma values and average luminance difference is created based on the aforementioned luminance difference. The dataset comprises 4 668 data pairs, each pair containing the mean luminance difference between the image exhibiting an exposure anomaly and its corresponding corrected image, as well as the gamma value associated with that luminance change. Given that on most cameras can the impact of gamma value on image brightness be described as follows: when gamma is greater than 1 and gradually increasing, the image will become brighter; when gamma is less than 1 and gradually decreasing, the image will become darker, therefore the average brightness difference in the dataset is represented by signed data rather than absolute values. In the case of an exposure time that has been roughly determined, the brightness of the image will not change significantly. Based on the actual test results, the brightness difference in the dataset has been set within the range of -80 to 135; the gamma value range is between 0 and 3.0. It should be noted that the adjustable range of the gamma value may have different mapping relationships in different cameras.

The BP neural network is constructed to derive the gamma value that should be corrected for different values of luminance difference. The network comprises two hidden layers: an input node representing the mean luminance difference, and an output node representing the gamma value. The combination of the average luminance difference of the images before and after processing by the exposure correction network and the fitted curve of this BP neural network constitutes the entire image exposure correction process.

1.3 Target detection

The image after exposure correction is fed into the target detection process. A clear image with uniform brightness distribution can significantly enhance the accuracy of target detection.

The target detection process of the surveillance camera is implemented using the YOLOv5 algorithm. To improve the ability of small infrared target localization, the coordinate attention mechanism is added to the algorithm's backbone network, and the CLLAHead detection head is added to improve the model's small target knowledge accuracy by using a wider range of scene information.

The flow of the modified algorithm is shown in Fig.7.

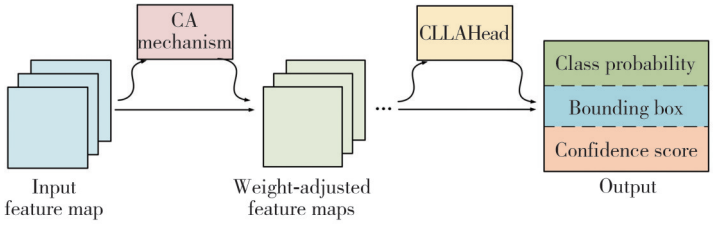


Fig. 7 Schematic diagram of improved algorithm

1.3.1 Coordinate attention mechanism

In YOLOv5, the attention mechanism is employed to augment the feature extraction module, thereby enabling the model to more effectively concentrate on salient regions within the image and consequently enhance the accuracy of the detection process. Infrared images are typically characterized by low contrast and indistinct edges. To enhance the model's capacity to detect minute infrared targets, the coordinate attention mechanism is introduced to improve feature representation[24]. The structure of the coordinate attention mechanism with feature map scale change is illustrated in Fig.8. Here, C, H, and W denote the number of channels, height, and width of the feature map, respectively; and r denotes the

scaling factor, which is employed to reduce the number of channels in the feature map, thereby reducing the computational complexity. Firstly, two one-dimensional vectors are obtained respectively by the average pooling of the input feature map in the horizontal (W axis) and vertical (H axis) directions. Subsequently, a concat operation is performed, and the channels are compressed using 1×1 Conv. Following that, the spatial information in the vertical and horizontal directions is normalized and activated through batch normalization (BN) and non-linear operations, and then encoded. The horizontal and vertical feature maps are subjected to an ascending dimension operation using 1×1 Conv, and a sigmoid activation function is used to respectively obtain the weights in both directions. After applying the sigmoid function, all values are compressed into the (0, 1) range. Consequently, the feature vectors after the sigmoid transformation are represented using more similar colors in the figure. Finally, the weights are multiplied with the input feature maps to obtain the output feature maps after adding the coordinate attention mechanism[25].

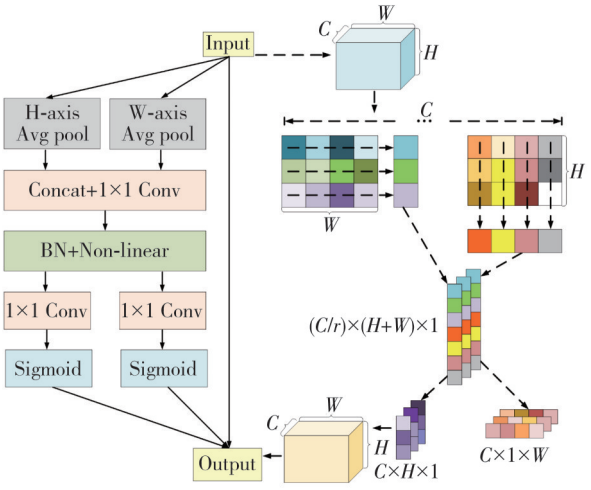


Fig. 8 Coordinate attention mechanism

The distinctive aspect of the coordinate attention mechanism is the use of separate pooling in the horizontal and vertical directions instead of the general global pooling. This is due to the fact that in the context of a small target recognition task, the precise location information of the target is of great importance. Global pooling, represents a global encoding of spatial information, which compresses global spatial information into channel descriptors and outputs only a single feature vector. This makes it difficult to retain location information. Pooling along the horizontal and vertical directions separately generates a pair of orientation-aware feature maps, which allows the model to learn long-range correlation information along one spatial direction while retaining accurate position information along the other spatial direction. The output $z_c^h(h)$ denotes the vertical direction pooling at height h in the cth channel can be expressed as

$$z_c^h(h)=\frac{1}{W}x_c(h,i), \tag{1}$$

where $x_c(h,i)$ denotes the input feature map's value at the cth channel, row h, column i. In the horizontal direction, the output $z_c^w(w)$ at the cth channel with width w can be expressed as

$$z_c^w(w)=\frac{1}{H}x_c(j,w). \tag{2}$$

Similarly, $x_c(j,w)$ denotes the input feature map's value at the cth channel, row j, column w.

1.3.2 Introducing context-aware light-weight label assignment detection head

In the YOLOv5 target detection algorithm, the primary function of the detection head is to transform the feature map into the final detection result, which typically comprises a regression branch and a classification branch for predicting the target's size and location within the image[26]. The original detection head in YOLOv5 treats the object detection task as a regression problem, dividing the entire image into multiple grid cells and predicting the bounding box and the confidence level for each bounding box within each grid cell. Finally, the prediction results of multiple grid cells are combined into a tensor, thereby simultaneously accomplishing object localization and classification during the forward propagation of the network.

In order to enhance the precision of target detection, it is essential to leverage contextual features and integrate a more comprehensive array of scene data. For this purpose, a CLLAHead detection head in conjunction with a distributed focus loss (DFL) function is employed to substitute for the detection head in the original model. The method is distinguished by its integration of contextual information[27] and use of distributed focus loss function[28] through multi-level feature extraction, which enhances the recognition and localisation of targets in images. It is particularly suited to complex scenarios comprising multiple targets of varying sizes. Furthermore, the parameter count of this detection head is minimal, making it well-suited to environments with limited computing resources, such as embedded devices.

In contrast to the conventional approach of feature fusion, which employs direct summing or splicing, the CLLAHead detection head utilizes a local feature weighted fusion of feature maps at varying scales. The

overall flow of the detection head is illustrated in Fig. 9. The notation "k1n256s1" refers to a convolutional layer characterized by having 256 output channels, a kernel size of 1×1, and a stride of 1. The feature maps preprocessed by this convolutional layer are fed into the CLLA feature fusion module for local feature extraction and weighted fusion. The feature fusion module performs local feature extraction on a region of $L \times W$ size, and attention computation is conducted between feature maps of varying levels to enhance the feature representation[29]. For the input feature vectors, the query vector ($\boldsymbol{Q}$), key vector ($\boldsymbol{K}$) and value vector (V) are computed by linear transformation. The feature-weighted fusion component calculates the weights corresponding to different features based on the similarity between the query and key vectors, and then weighs the sum. The fused features are subjected to target detection by the CLLASegment module, and a segmentation mask is generated for each detected target. This allows the position of the target to be accurately determined and the shape of the target to be precisely delineated[30]. The DFL module is employed for the implementation of DFL, which serves the primary function of facilitating the acquisition of a distributional representation through tensor transformations for the purpose of learning predictive probability distributions in target detection tasks. This enables the model to gain a more nuanced understanding of target uncertainty[31].

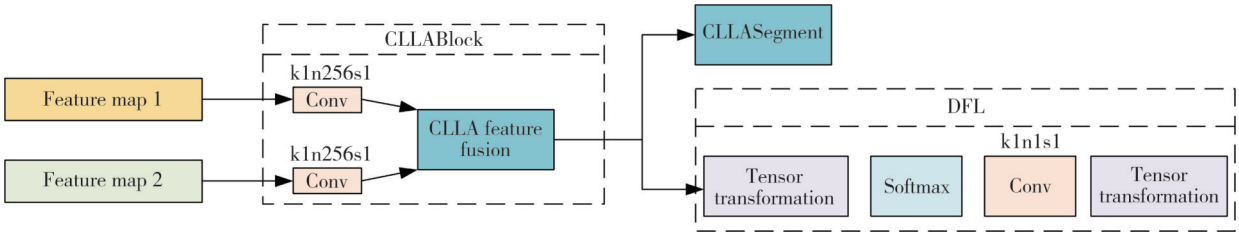


Fig. 9 Overall process of CLLAHead

1.4 Automatic reporting process

The automatic reporting process relies on frame difference motion detection and the aforementioned improved YOLOv5 algorithm. Frame difference motion detection is designed to identify moving targets in video. The algorithm performs a difference operation on two or three frames that are consecutive in time, subtracting the pixel values at corresponding positions across different frames. It then judges the absolute value of the greyscale difference and determines that there is a motion target in the frame if the absolute value exceeds a certain threshold. The particular methodology for implementing this function is as follows: transforming the image data into a greyscale map, and calculating the inter-frame difference between the $(N-1)$th frame and the Nth frame in accordance with

$$diff(i,j)=\left| frame_N(i,j)-frame_{N-1}(i,j)\right|, \quad (3)$$

where $frame(i,j)$ denotes the pixel value at the coordinates (i,j) in the original images, while $diff(i,j)$ represents the pixel value in the difference image. Afterwards, each pixel value of the difference image is compared with the threshold value specified in the procedure, and the pixel value $P(i,j)$ is set to a predetermined maximum or minimum value according to the comparison result as

$$P(i,j)=\begin{cases}255, & diff(i,j)>\text{threshold},\\ 0, & diff(i,j)\leqslant\text{threshold}.\end{cases} \quad (4)$$

In the event that the differential pixel value exceeds the specified threshold, the pixel value of the point is set to 255, which corresponds to the color white. Conversely, if the differential pixel value is less than or equal to the aforementioned threshold, the pixel value of the point is set to 0, which represents the color black. Subsequently, the result after thresholding is denoised. Ultimately, the contour of the moving object is determined by the white area on the image. When the contour is sufficiently large, it is determined that a moving object is present. The algorithmic flow and practical results are illustrated in Fig. 10.

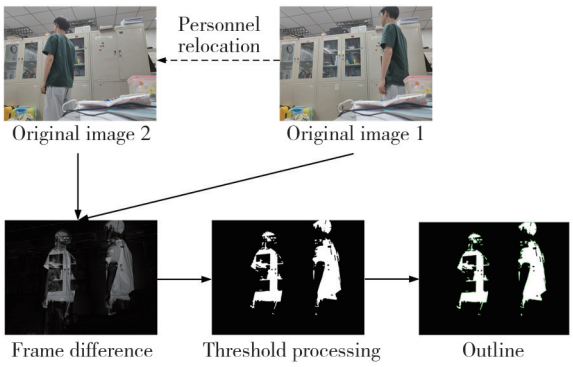


Fig. 10 Motion detection algorithm

Motion detection algorithm is used for photographing individuals or vehicle intruding in an unattended environment in automatic mode. The overall automatic reporting process is to capture images when the motion detection algorithm determines the presence of moving

objects. This is followed by the application of the improved YOLOv5 algorithm for target recognition. In the event that preset abnormal objects, such as people and vehicles, are present within the target, the remote camera is networked and the captured image is returned to the client via email. Conversely, if no abnormal objects are present within the target, the camera does not return the image.

2 Experiment and analysis

2.1 Result of image exposure correction process

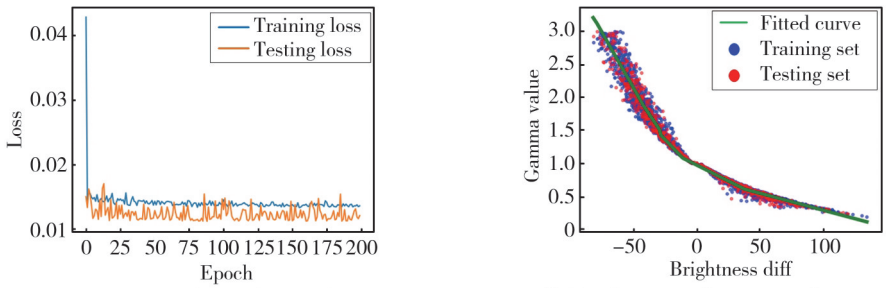
This part mainly includes the results of both DNN and BP neural networks, as well as the final results of the entire image exposure correction process. The effect of image exposure correction is demonstrated in Fig. 11, the subfigures (a) and (b) present the restoration results of overexposed image, while subfigures (c) and (d) illustrate the correction outcomes for underexposed image. Although the restored images exhibit significant improvements in terms of color and brightness, they become noticeably more blurred compared to the original images. This limitation is more evident in the restoration of overexposed images.

(a) Overexposed image (b) Overexposed image restoration effect

(c) Underexposed image (d) Underexposed image restoration effect

Fig. 11 DNN exposure correction effect

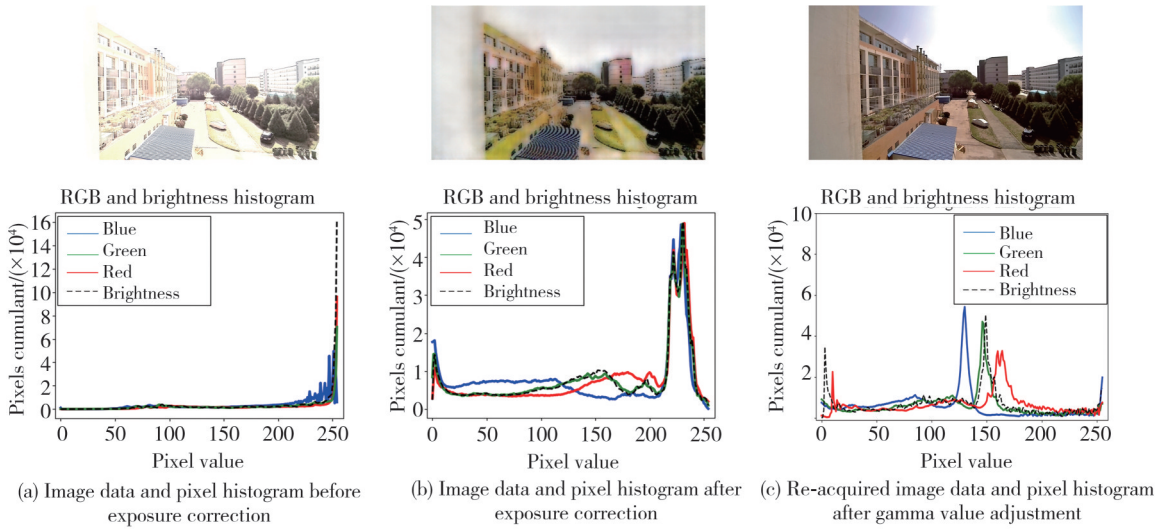
The fitted curve is illustrated in Fig. 12. The average testing loss of the network is 0.013, while the average training loss is 0.015. The fitted curve exhibits a high degree of consistency with the dataset in Fig. 12(b), indicating the model's strong fitting capability.



(a) Training loss and testing loss (b) Fitted curve performance on dataset

Fig. 12 Fitted results of BP neural network

The actual testing results of the exposure correction process are shown in Fig. 13. The figure presents the image data and corresponding pixel histograms before and after the DNN exposure correction, as well as after gamma value adjustment.



(a) Image data and pixel histogram before exposure correction (b) Image data and pixel histogram after exposure correction (c) Re-acquired image data and pixel histogram after gamma value adjustment

Fig. 13 Actual testing results after gamma value adjustment

Histograms quantitatively analyze the color distribution and brightness of image data, with the horizontal axis representing pixel values ranging from 0 to 255 and the vertical axis representing the frequency of each pixel value in the image. In the histogram before exposure correction, brightness values are concentrated near 255, indicating an overall bright image, and the pixel values across all three color channels are also predominantly distributed near 255, indicating that the image has an overall white color bias. After exposure correction, the pixel value distribution tends to be more balanced; however, the DNN exposure correction network causes image blurriness, as Fig. 13(b). The final images, Fig. 13(c), obtained after gamma value adjustment using the fitted curve, exhibit balanced color distribution, the majority of pixel values fall within the range of 100 to 200, surpassing the original images in both subjective quality and quantitative histogram analysis.

2.2 Result of improved YOLOv5 algorithm

2.2.1 Dataset and experimental environment

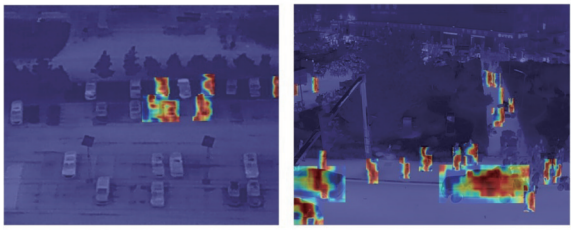
The experimental configuration is based on the x86_64 architecture of the Ubuntu 18.04 system with an NVIDIA RTX 3060 GPU and 32 GB running memory. Python 3.9 was used as the programming environment and torch-1.12.0 as the deep learning framework.

The infrared dataset used in this research totals 10 820 images. The tag types are a variety of large and small targets under infrared light conditions, including trucks, bus, cars, walking personnel, and cyclist, which are common in industrial environments. Trucks and bus are basically large targets, most of cars are small targets, and personnel are basically small targets. The images were captured at a height of between 5 m and 60 m above the ground, in alignment with the installation location of the outdoor industrial environment surveillance camera. One-tenth of the images in the dataset were randomly selected to form the testing set, one-tenth of the remaining images were selected as the validation set, and the other images were the training set. The whole process used stochastic gradient descent method to train 160 epochs, batch size is set to 16, learning rate is set to 0.01, cosine annealing hyperparameter is set to 0.01, learning rate momentum is set to 0.937, and weight decay coefficient is set to 0.000 5.

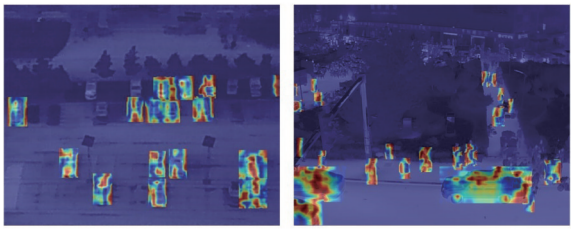
2.2.2 Comparative experiments before and after model improvement

To evaluate the efficacy of the model enhancement, a comparative analysis was conducted between the algorithms prior to and subsequent to the improvement.

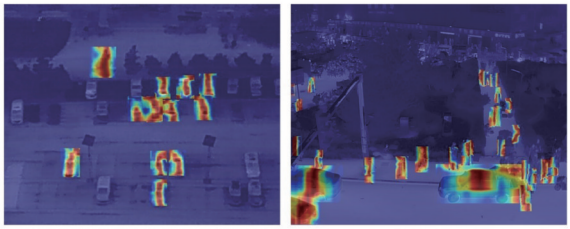
Heatmap is a visual representation of the potential locations of targets within an image. By mapping the confidence levels across different regions of an image, heatmaps offer insights into the decision-making process of the model. The intensity of color in a heatmap corresponds to the likelihood of a target's presence, with the brighter the color, the greater the probability that the model considers the presence of a target in that region. Fig. 14 illustrates the comparison of heatmap results between the original YOLOv5 algorithm and the modifications made by separately introducing the coordinate attention mechanism and the CLLAHead detection head. Result shows the incorporation of the coordinate attention mechanism enhances the ability of small infrared target localization, identifying more targets and delineating the boundaries between adjacent targets more distinctly. And CLLAHead detection head leads to a notable enhancement in the model's capacity to recognize distant minute targets.



(a) Heatmap of original model



(b) Heatmap of model with CA attention mechanism



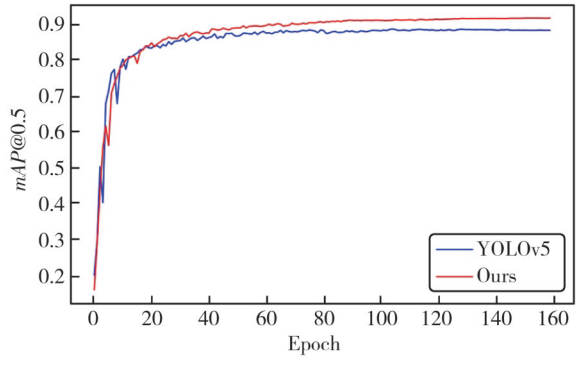
(c) Heatmap of model with CLLAHead detection head

Fig. 14 Heatmaps of module impacts on YOLOv5 algorithm

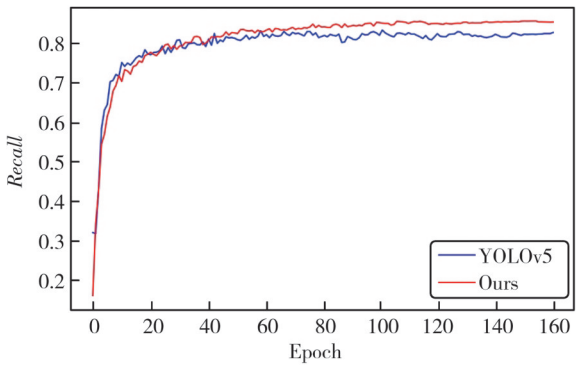
The *mAP*@0.5 and *Recall* of two algorithms during the training process were compared, and the results are shown in Fig. 15. The metric *mAP*@0.5 is the mean average precision across all categories, calculated at an

intersection over union (IoU) threshold of 0.5, and *Recall* is the ratio of correctly identified positive instances to the total actual positive instances.

Fig. 15(a) illustrates the change curve of the model *mAP*@0.5 before and after improvement. During the initial 20 training sessions, the model *mAP*@0.5 exhibits significant fluctuations before and after improvement due to model underfitting.



(a) *mAP*@0.5 curve



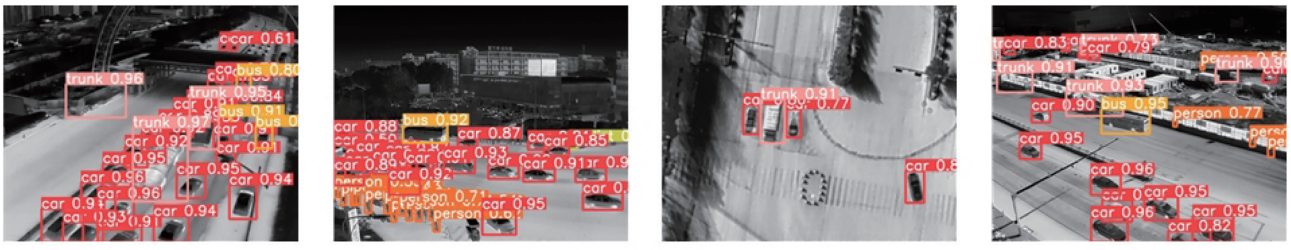
(b) *Recall* curve

Fig. 15 *mAP*@0.5 and *Recall* curve comparison before and after model improvement

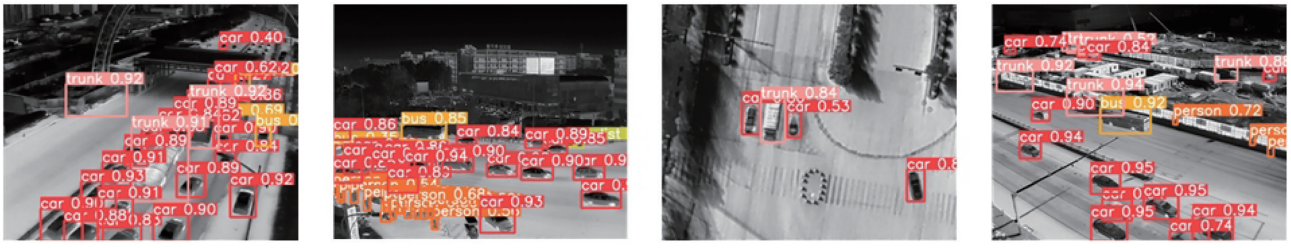
Following this period, the value of the improved model exceeds that of the pre-improved model. After the 80 training sessions, the model *mAP*@0.5 value demonstrated a tendency towards stabilization, and the *mAP*@0.5 of the improved model reaches approximately 91.5%, representing a notable enhancement in comparison to the 86.8% achieved by the pre-improved model. This suggests that the improved model exhibits superior detection accuracy relative to the pre-improved model.

Fig. 15(b) illustrates the change curve of the model *Recall* before and after the implementation of improvements. The improved model exhibits a higher recall rate than the pre-improved model. After stabilisation, the improved model's *recall* rate reaches 85.5%, exceeding the original model's 82.9%. This indicates that the improved model exhibits a reduced leakage rate and a heightened probability of detecting the target in the same environment, while simultaneously ensuring the ability to discriminate between similar targets and image noise.

Fig.16 illustrates the test plot of the results obtained prior to and following the implementation of the algorithmic improvement. Fig.16(a) depicts the enhanced algorithmic detection outcomes, whereas Fig. 16(b) illustrates the original algorithmic detection outcomes. The original algorithm's target recognition accuracy is relatively low due to the distance of the target, the small proportion of the overall picture, or the mutual occlusion of the targets. The improved model significantly addresses these shortcomings, enhancing target recognition accuracy and reducing target omission.



(a) Improved algorithm detection results



(b) Original algorithm detection results

Fig. 16 Comparison between improved algorithm and original algorithm detection results

2.2.3 Ablation experiments

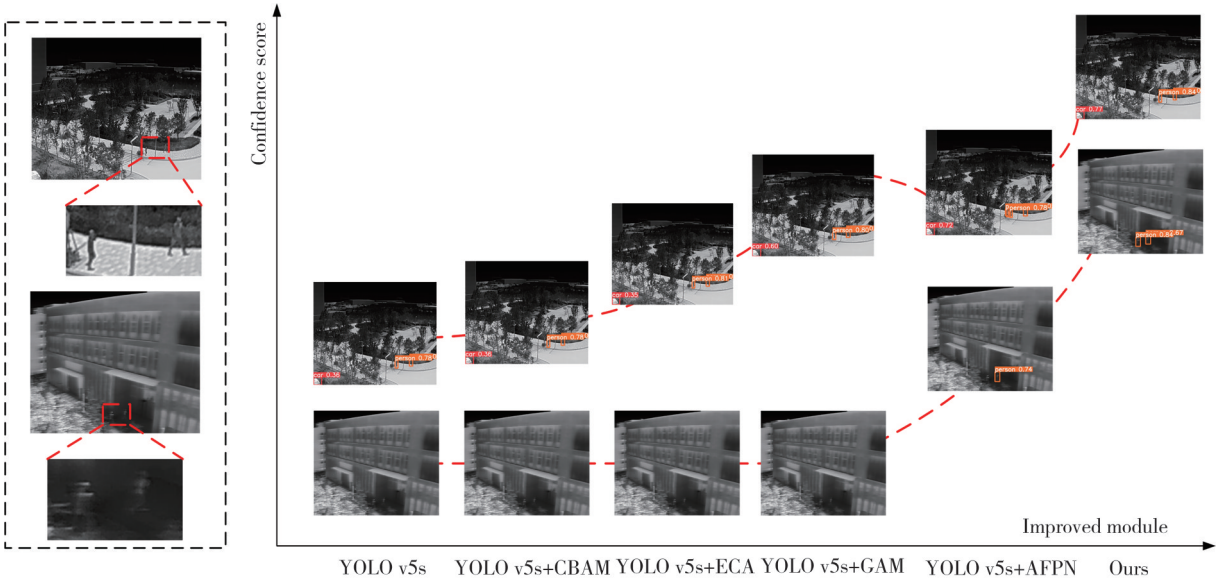
The CLLAHead detection head, the coordinate attention mechanism, and the distribution focus loss function are incorporated into the original model, and the final results of the ablation experiment are presented in Table 1. The overall evaluation index of the model is based on *mAP*@0.5 and *Recall*, while *mAP*@0.5 of the small targets (Person and Cyclist) is employed to assess the model's capacity to detect small targets. The results of the experiments demonstrate that the introduction of each module in this study has led to an enhancement in the model's accuracy in recognising the target. Of these, the CLLAHead detection head has the most pronounced effect on the model's performance, with an improvement of 2.4% in *mAP*@0.5 for the overall target and 2.5% and 3.6% for the Person and Cyclist mini-targets, respectively. The complete improved algorithm demonstrates an enhancement in *mAP*@0.5 by 4.7% and in *Recall* by 2.6% in comparison to the original algorithm for the overall target. Furthermore, the *mAP*@0.5 improvement is 6.8% and 6.1% for the two small targets, respectively. The experimental results demonstrate that the method proposed in this research can effectively enhance the *mAP* and *Recall* of the network model, thereby improving the precision of infrared small target detection.

Tabel 1 Results of ablation experiments

Improved modules			Performance metrics			
CLLAHead	DFL	CA	*mAP*@0.5/%	*Recall*/%	*mAP*@0.5-Person/%	*mAP*@0.5-Cyclist/%
×	×	×	86.8	82.9	81.9	82.1
√	×	×	89.2	84.1	84.4	85.7
√	√	×	90.1	84.4	85.8	86.6
√	√	√	91.5	85.5	88.7	88.2

2.2.4 Comparative experiments

The improved YOLOv5 algorithm respectively incorporating efficient channel attention (ECA), global attention module (GAM), convolutional block attention module (CBAM), and adaptive feature pyramid network (AFPN) modules is compared with the improved algorithm in this research for small object detection on infrared dataset, as illustrated in Fig. 17. The results show that compared with the original YOLOv5s model, the above improvements all lead to an increase in target recognition accuracy. However, the ECA and CBAM predominantly influence the recognition accuracy of large targets, exhibiting a relatively weaker impact on small targets. The AFPN detection head, improves the recognition accuracy of small targets but demonstrates a limited capacity to differentiate between similar targets. Additionally, all the aforementioned improvement methods exhibit a tendency for small target missed detection.



(a) Original images (b) Comparison of detection results with diffferent improved modules

Fig. 17 Comparison of recognition results with different improvement modules

2.3 Physical equipment

The camera utilizes email for the transmission and receipt of information. This is achieved by parsing the subject line of the email to ascertain the requisite control commands, which are then executed to facilitate the

desired functions. The actual email is illustrated in Fig.18. The sender of the return email is "Device 1", the subject and content of the email are its corresponding functions, and the acquired image data are added to the return email through attachments.

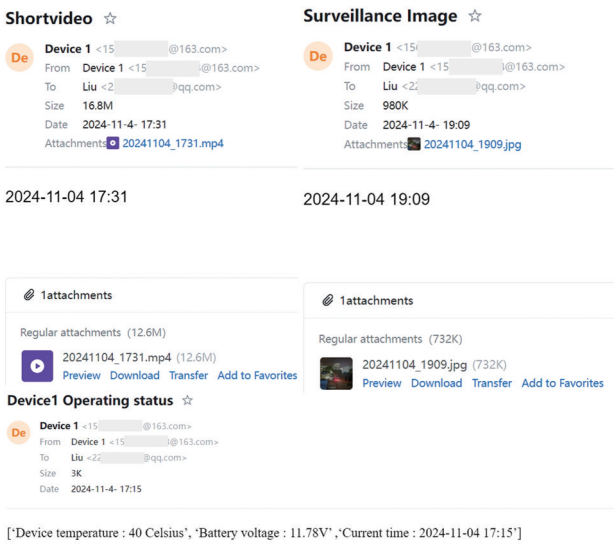


Fig. 18 Email for the actual return of the camera

3 Conclusions

This paper presents the design and implementation of a remote intelligent surveillance camera based on the RK3566 chip, which employs email control and data transmission. Following testing, the average time required for transmitting a single image is less than 10 s, while the average time for transmitting a short video is less than 60 s, which meets the demand of remote surveillance under unattended environment.

Furthermore, an image exposure correction process has been developed, which utilises a DNN for the preprocessing of images acquired by the surveillance camera, and adjust the camera's gamma value based on BP neural network fitted result. Thereby ensuring the quality of the captured images.

In order to enhance the detection capacity for anomalies during nocturnal periods, improvements are made to the YOLOv5 algorithm, including the incorporation of a coordinate attention mechanism in the backbone network to enhance the model's perceptual capacity and the use of the CLLAHead detection head to obtain richer feature representations, thereby increasing the model's detection accuracy. The experimental results demonstrate that the enhanced algorithm attains a *mAP*@0.5 of 91.5%, representing a 4.7% improvement over the original algorithm. Additionally, the *Recall* reaches 85.5%, an increase of 2.6%, while markedly enhancing the detection capability for small infrared targets.

In future research, more precise power consumption control methods should be explored. Additionally, the use of dedicated email accounts or email platforms can be considered to enhance upload bandwidth and improve device transmission speed.

Acknowledgement

This work was Funded by Open Foundation of the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University, China (No.GZKF-202219).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] LIU X T, HE J F, GAO P, et al. Denoising of 3D Magnetic resonance images based on balanced low-rank tensor and nonlocal self-similarity. Biomedical Signal Processing and Control, 2024, 96: 106588.

[2] XU Z, YANG G Q, ZHANG Y Y. Road target detection algorithm based on improved YOLOv5//2023 IEEE 7th Information Technology and Mechatronics Engineering Conference, September 15-17, 2023, Chongqing, China. New York: IEEE, 2023: 787-791.

[3] RATTHI K I, YOGAMEENA B, PERUMAAL S S. Human height estimation using AI-assisted computer vision for intelligent video surveillance system. Measurement, 2024, 236: 115133.

[4] MAGNO M, TOMBARI F, BRUNELLI D, et al. Multimodal video analysis on self-powered resource-limited wireless smart camera. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2013, 3(2): 223-235.

[5] LIU Q, HIRATSUKA S, GOTO S, et al. A 41mW VGA@30fps quadtree video encoder for video surveillance systems//2007 7th International Conference on ASIC, October 22-25, 2007, Guilin, China. New York: IEEE, 2007: 758-761.

[6] MOSTAFA S A, RAVI S, ASAAD ZEBARI D, et al. A YOLO-based deep learning model for real-time face mask detection via drone surveillance in public spaces. Information Sciences, 2024, 676: 120865.

[7] YANG H M, WANG S, CHEN Y F, et al. Improving power transmission tower state recognition in remote sensing images using cooperative Adaboost-MobileNet. Remote Sensing Letters, 2023, 14(2): 124-134.

[8] BAI Z Y, WANG Y, ZHANG A C, et al. Road surface condition monitoring in extreme weather using a feature-learning enhanced mask-RCNN. Journal of Transportation Engineering, Part B: Pavements, 2024, 150(3): 04024030.

[9] LIU X Y, WANG T, YANG J M, et al. MPQ-YOLO: Ultra low mixed-precision quantization of YOLO for edge devices deployment. Neurocomputing, 2022, 574: 127210.
[10] SUN Q Y, LI P B, HE C T, et al. A lightweight and high-precision passion fruit YOLO detection model for deployment in embedded devices. Sensors, 2024, 24(15): 4942.
[11] WEI X, WEI Y, LU X B. HD-YOLO: Using radius-aware loss function for head detection in top-view fisheye images. Journal of Visual Communication and Image Representation, 2023, 90: 103715.
[12] ABDUSALOMOV A, BARATOV N, KUTLIMURATOV A, et al. An improvement of the fire detection and classification method using YOLOv3 for surveillance systems. Sensors, 2021, 21(19): 6519.
[13] KHAZUKOV K, SHEPELEV V, KARPETA T, et al. Real-time monitoring of traffic parameters. Journal of Big Data, 2020, 7(1): 84.
[14] DARMADI, DONI H N. Traffic counting using YOLO Version-5 (a case study of Jakarta-Cikampek toll road). IOP Conference Series: Earth and Environmental Science, 2024, 1321(1): 012015.
[15] RANI N G, PRIYA N H, AHILAN A, et al. LV-YOLO: logistic vehicle speed detection and counting using deep learning based YOLO network. Signal, Image and Video Processing, 2024, 18(10): 7419-7429.
[16] TIAN Y, YU C Z, XIE F, et al. Research on video detection method of moving target oriented to substation. IOP Conference Series: Earth and Environmental Science, 2021, 804(3): 032011.
[17] GAO Z W, ZHANG N B. Object detection based on RetinaNet and CBAM attention mechanism // International Conference on Internet of Things and Machine Learning (IoTML 2022), December 16-18, 2022, Harbin, China. Society of Photo-Optical Instrumentation Engineers, 2023: 12640.
[18] MA X J, JI K F, XIONG B L, et al. Light-YOLOv4: an edge-device oriented target detection method for remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 10808-10820.
[19] ORTIZ R F, LIN W M. Improving performance of simultaneous multithreading CPUs using autonomous control of speculative traces. Microprocessors and Microsystems, 2024, 108: 105073.
[20] PIEDRAHITA CASTILLO D, REGIDOR F M, HIGUERA J B, et al. A new mail system for secure data transmission in cyber physical systems. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2020, 28(S2): 23-48.
[21] ZHANG X Y, TANG X J, YU L P, et al. Automated camera exposure control for accuracy-enhanced stereo-digital image correlation measurement. Sensors, 2022, 22(24): 9641.
[22] KARAIMER H C, BROWN M S. A software platform for manipulating the camera imaging pipeline//European Conference on Computer Vision, October 11-14, 2016, Amsterdam, The Netherlands. Cham: Springer International Publishing, 2016: 429-444.
[23] AFIFI M, DERPANIS K G, OMMER B, et al. Learning multi-scale photo exposure correction//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, Nashville, TN, USA. New York: IEEE, 2021: 9153-9163.
[24] YI X M, ZHOU Y, WU P, et al. U-Net with coordinate attention and VGGNet: a grape image segmentation algorithm based on fusion pyramid pooling and the dual-attention mechanism. Agronomy, 2024, 14(5): 925.
[25] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, Nashville, TN, USA. New York: IEEE, 2021: 13708-13717.
[26] WEI C Y, TAN Z, QING Q X, et al. Fast helmet and license plate detection based on lightweight YOLOv5. Sensors, 2023, 23(9): 4335.
[27] TIAN Z T, CUI J Q, JIANG L, et al. Learning context-aware classifier for semantic segmentation//The 37th AAAI Conference on Artificial Intelligence, February 7-14, 2023, Washing. DC, USA. New York: ACM, 2023: 2438-2446.
[28] LI X, WANG W H, WU L J, et al. Generalized focal loss//34th International Conference on Neural Information Processing Systems, December 6-12, 2020, Vancouver, BC, Canada. New York: ACM, 2020: 21002-21012.
[29] ZHAO M, YANG R, HU M, et al. Deep learning-based technique for remote sensing image enhancement using multiscale feature fusion. Sensors, 2024, 24(2): 673-688.
[30] LI J, SUN H C, ZHANG Z Y. A multi-scale-enhanced YOLO-V5 model for detecting small objects in remote sensing image information. Sensors, 2024, 24(13): 4347.
[31] VAN AMERSFOORT J, SMITH L, TEH Y W, et al. Uncertainty estimation using a single deep deterministic neural network//The 37th International Conference on Machine Learning, July 12-18, 2020, Vienna, Austria. New York: ACM, 2020: 9690-9700.

一种基于电子邮件平台的新型智能远程监拍相机

刘昕昊[1], 孟令军[1*], 刘　峰[1], 周小彤[2], 王嘉诚[2]

1. 中北大学 仪器与电子学院, 山西 太原 030051;
2. 上海大学 机电工程与自动化学院, 上海 200444

摘　要： 针对监拍系统服务器部署和接入困难、运行维护成本高等问题，基于RK3566芯片设计了一种通过电子邮件平台控制和数据回传的远程监拍相机。首先，为应对天气、光照等自然条件对获取图像质量造成的影响，利用一种DNN图像曝光修正网络对曝光异常的图像进行修复，并根据BP神经网络对曝光修复前后图像亮度差和gamma值的数据拟合曲线来调节相机gamma值。其次，为提高YOLOv5算法对夜间异常情况判别的准确率，采用红外图像数据并在模型中引入轻量级上下文感知检测头和坐标注意力机制以提高模型对小目标的检测精度和召回率。同时，为满足无人值守环境下的异常情况自动上报需求，设计了一种结合YOLOv5算法和帧差分运动检测算法的异常目标自动识别上报流程。经测试，该远程监拍相机对当前主流商用电子邮件平台均适用，相机传输单张图片数据的平均耗时在10 s以内，传输短视频的平均耗时小于60 s。BP网络平均训练损失为0.015，平均测试损失为0.013，拟合结果基本满足gamma调节精度要求。改进后的YOLOv5算法 $mAP@0.5$ 为91.5%，召回率为85.5%，有效提升了小目标识别的准确率。

关键词： 邮件传输；曝光修正；反向传播神经网络；gamma值；YOLOv5

引用格式： LIU Xinhao, MENG Lingjun, LIU Feng, et al. A novel email-based smart remote image surveillance camera. Journal of Measurement Science and Instrumentation, 2025, 16(1): 128-141. DOI: 10.62756/jmsi.1674-8042.2025013