Hindawi
Security and Communication Networks
Volume 2021, Article ID 9999398, 14 pages
https://doi.org/10.1155/2021/9999398

Research Article

Towards Efficient Video Detection Object Super-Resolution with Deep Fusion Network for Public Safety

Sheng Ren, Jianqi Li, Tianyi Tu, Yibo Peng, and Jian Jiang

School of Computer and Electrical Engineering, Hunan University of Arts and Science, Changde 415000, China

Correspondence should be addressed to Jianqi Li; hnwlpyb@huas.edu.cn

Received 22 March 2021; Revised 14 April 2021; Accepted 14 May 2021; Published 24 May 2021

Academic Editor: David Megías

Video surveillance plays an increasingly important role in public security and is a technical foundation for constructing safe and smart cities. The traditional video surveillance systems can only provide real-time monitoring or manually analyze cases by reviewing the surveillance video. So, it is difficult to use the data sampled from the surveillance video effectively. In this paper, we proposed an efficient video detection object super-resolution with a deep fusion network for public security. Firstly, we designed a super-resolution framework for video detection objects. By fusing object detection algorithms, video keyframe selection algorithms, and super-resolution reconstruction algorithms, we proposed a deep learning-based intelligent video detection object super-resolution (SR) method. Secondly, we designed a regression-based object detection algorithm and a key video frame selection algorithm. The object detection algorithm is used to assist police and security personnel to track suspicious objects in real time. The keyframe selection algorithm can select key information from a large amount of redundant information, which helps to improve the efficiency of video content analysis and reduce labor costs. Finally, we designed an asymmetric depth recursive back-projection network for super-resolution reconstruction. By combining the advantages of the pixel-based super-resolution algorithm and the feature space-based super-resolution algorithm, we improved the resolution and the visual perception clarity of the key objects. Extensive experimental evaluations show the efficiency and effectiveness of our method.

1. Introduction

Video surveillance systems are widely distributed in urban streets and roads, commercial places, residential areas, bank outlets, stations, terminals, airports, and other public places, playing an increasingly important role in public security. Through the video surveillance system, suspicious signs and objects can be found in time and monitored closely, to avoid the occurrence of criminal harm effectively. The police can obtain information about criminals by surveillance videos and inquire about the locations of suspected vehicles and personnel. In the interrogation stage of case investigation, the surveillance video can be used as objective litigation evidence. Video surveillance has become the fourth major field of investigation technology after criminal science and technology, action technology, and network investigation technology. It plays an irreplaceable role in the construction of a safe and smart city. In 2015, the Ministry of Public Security, the Ministry of Science and Technology, and other nine ministries and commissions put forward "several opinions on strengthening the networking application of public security video monitoring construction". And, they pointed out that constructing the network application of public security video surveillance helps to maintain national security and social stability and to prevent and crack down on violent terrorist crimes under the new situation. It is of great significance to improve the urban and rural management and innovate the social governance system.

The traditional video surveillance system can only provide real-time monitoring or manually analyze cases by reviewing the surveillance video. It leads to a low usage rate of surveillance video data. With artificial intelligence and machine learning technologies, video surveillance systems can intelligently analyze video content, detect abnormal behaviors, and discover potentially harmful behaviors [1]. Besides, these technologies can assist police and security

personnel in investigating cases, thereby providing more accurate and safer surveillance. Alibaba Cloud's intelligent video surveillance platform can identify fireworks and inspectors wearing helmets, detect intruders in the area, and escort safe production. Baidu's video surveillance development platform EasyMonitor has a wealth of AI business skills, including electronic fences, smoke detection, safety cap detection, and departure detection [2]. Large Internet companies have released many intelligent video surveillance products. However, on the one hand, limited by the cost of the software product purchase, operation, and maintenance, the surveillance video field lacks simple and effective auxiliary tools. On the other hand, limited by hardware cost, hardware technology, and shooting environment, the current surveillance video suffers from low-resolution and unclear visual perception. In summary, there are two problems in the field of public security based on surveillance video: (1) in most cases, surveillance video viewers need to manually identify the object and manually select video keyframes for analysis, making it easy to lose the object in addition to being inefficient. (2) It is difficult to use the low-resolution video frames since it is easy to lose high-frequency information when zooming in to view the object, resulting in blurring and hardness of recognition.

To solve the above problems, we proposed a super-resolution method for video detection objects. We use object detection algorithms to assist surveillance video viewers to track objects in real time and use super-resolution algorithms to reconstruct high-resolution video frames with clear visual perception. The traditional object detection algorithm OpenCV cascade classifier uses a sliding window to select the region, then uses HOG + SVM and other methods for feature extraction, and finally uses the classifier to classify the detection area [3, 4]. Object detection algorithms based on deep learning can be divided into two types: object detection and recognition algorithms based on region recommendations and object detection; recognition algorithms based on regression. R-CNN, object detection and recognition algorithm based on region recommendations, first generates object candidate frames based on region recommendations, then filters the generated object candidate frames, and finally refines the size and position of the candidate frames [5]. The regression-based object detection and recognition algorithm YOLO regards object detection as a regression problem. The training phase aims to train a set of weights and directly call the trained weights for object positioning during testing [6]. Traditional super-resolution methods are mainly based on interpolation (such as zero-order interpolation, bilinear interpolation, and bicubic interpolation) and examples. The instance-based sparse representation method establishes the mapping relationship between low-resolution and super-resolution images by learning the sparse associations between image blocks to achieve super-resolution reconstruction of images [7]. Super-resolution algorithms based on deep learning can be divided into pixel-space-based methods and feature-space-based methods. A super-resolution method based on pixel space SRCNN first uses a $9 * 9$ convolutional layer to extract the initial features of the image, then uses a $1 * 1$ convolutional layer to learn the nonlinear mapping from low resolution (LR) to high resolution (HR), and finally uses a $5 * 5$ convolutional layers reconstruct super-resolution images [8]. SRGAN, a super-resolution method based on feature space, learns the nonlinear mapping from LR to HR through a generator, and then the discriminator constrains the generated super-resolution image in terms of semantics and style [9]. These object detection algorithms and super-resolution algorithms have achieved better and better results in their respective fields, but they lack integration, unification, and optimization for specific application scenarios and are not suitable for direct use in the field of public security based on video surveillance [10].

In this paper, we proposed a super-resolution method based on a deep fusion network for surveillance video object detection. Firstly, we designed a comprehensive surveillance video analysis framework that integrates object detection algorithms, keyframe selection algorithms, and super-resolution algorithms. It solves the problem of large workload, easy object loss, and low resolution in public security video data analysis. Secondly, we used regression-based object detection and recognition algorithms to identify video objects in real time, which is convenient for surveillance video viewers to track objects. Besides, we employed the keyframe selection algorithm to select the frames with significant changes in the scene of the surveillance video to reduce the workload of video analysis. Finally, a super-resolution algorithm combining pixel space and feature space is used to reconstruct the object in the keyframe. It is beneficial to improve the resolution of key objects and the visual perception quality of objects detected by surveillance video and surveillance video viewer's investigation and handling cases. The main contributions of this paper are summarized as follows: (1) We designed a novel comprehensive analysis framework for surveillance video. It improves the efficiency and accuracy of video analysis by combining object detection, keyframe selection, and super-resolution algorithms. (2) We proposed a keyframe selection algorithm and use regression-based object detection and recognition algorithm to recognize video objects in real time. (3) We proposed a super-resolution approach that deeply integrates the advantages of pixel space and feature space to improve the resolution of surveillance video detection objects.

The rest of this paper is organized as follows: Section 2 explains the related works, Section 3 describes the work process of our approach, Section 4 explains our approach in detail, and Section 5 provides experimental results. Section 6 concludes our work.

2. Related Work

2.1. Object Detection Algorithms. Object detection is the basic task of computer vision and can be widely used in object tracking, crowd counting, face recognition, and other fields. It is an important algorithm in public safety. Object detection algorithms based on candidate regions, also known as two-stage object detection algorithms, mainly include region-based convolution neural networks (R-CNN), Fast R-CNN, Faster R-CNN, and other models

[11–13]. Regression-based object detection algorithms, also known as one-stage object detection algorithms, mainly include YOLO series algorithms and SSDs [14, 15]. The two-stage object detection algorithm consists of two steps. We first generate object candidate frames based on regional suggestions, then filter the candidate frames, and classify the objects. Faster R-CNN first uses the Region Proposal Network (RPN) to generate a candidate frame and then uses softmax to determine whether the candidate frame is in the foreground or the background. And, it employs bounding box regression to correct the candidate frame to obtain a more accurate candidate frame, called proposals. Then RPN uses the region of interest (ROI) layer pool proposals of different sizes into the same size and uses a fully connected layer for object classification and position adjustment regression [5]. The two-stage object detection algorithm has high recognition accuracy, but the generation and selection of candidate frames will consume a lot of computing power and time, and it is difficult to reach the detection speed of 24FPS, which is not suitable for real-time object detection. YOLO v5 directly returns the position and category of the bounding box in the output layer. It first uses Mosaic data enhancement to solve the problem of many small objects in the data set and uneven distribution on the input side. Then, it introduces the focus and cross stage partial (CSP) structure into backbone to improve the feature extraction capabilities. Besides, in neck, it uses feature pyramid network (FPN) and pyramid attention network (PAN) structures to fuse semantic features and positioning features of feature maps at different scales and aggregates parameters for different detection layers. Finally, on the prediction side, it uses generalized intersection over union (GIoU) loss to handle the noncoincidence of bounding boxes. YOLO v5 has a faster running speed and can meet the requirements of real-time object detection. The object detection algorithm will generate a lot of redundant information in the surveillance video frames. Selecting keyframes from abundant video frames for efficient object tracking is an effective method to improve detection efficiency and reduce the calculation and workload.

2.2. Super-Resolution Algorithms. The super-resolution algorithm can reconstruct the corresponding high-resolution image from a single or a series of low-resolution images. It is a low-level computer vision algorithm and the basis of a high-level computer vision algorithm. The super-resolution algorithms based on deep convolutional neural networks can reconstruct high-quality super-resolution images with rich high-frequency information and clear texture features, which have become the mainstream research methods. The super-resolution convolutional neural networks (SRCNN) use convolutional neural networks for the first time in the field of super-resolution and only use three-layer convolutional neural networks to surpass most traditional super-resolution methods [8]. Compared with the example-based super-resolution method, SRCNN does not require complicated preprocessing and can optimize the super-resolution reconstruction results several times through backpropagation, which not only improves the quality of reconstructed results but also improves the efficiency. The main disadvantage of SRCNN is that it is difficult to build deep convolutional neural networks to improve feature extraction and characterization capabilities. The residual network ResNet connects the input side information to the output side via a shortcut connection [16]. The convolutional layer only needs to learn the residuals of the input and output, which is the basis for building a deep neural network. The key to super-resolution reconstruction is to optimize high-frequency information. The residuals learned by the residual network contain rich high-frequency information, and since most of the residuals are close to 0, the learning efficiency is very high. Therefore, the residual network plays a very important role in improving the efficiency and quality of super-resolution reconstruction and becomes the backbone network of the super-resolution methods. Super-resolution using very deep convolutional networks (VDSR) introduces the residual network into the super-resolution field for the first time and builds a super-resolution network with a depth of 20 layers, which expands the receptive field and improves the convergence speed [17]. The speed and quality of VDSR are significantly better than SRCNN, and it uses a single network to reconstruct super-resolution images of different multiples. Megvii Research Institute proposed a method that can achieve super-resolution reconstruction at any multiple through a single model [18]. Zhang et al. of Harbin Institute of Technology proposed a plug-and-play super-resolution method based on regular depth to process low-resolution images with arbitrary blur kernels [19]. SenseTime proposed a super-resolution method of real scenes with original images [20]. Li et al. of Sichuan University proposed an image super-resolution feedback network to improve the low-level representation with high-level information [21]. Li et al. of Wuhan University proposed a fast spatiotemporal residual network for video super-resolution [22]. Gu et al. of the Chinese University of Hong Kong proposed a fuzzy kernel estimation method in the blind oversegmentation problem [23]. Dai et al. of Tsinghua University proposed a second-order attention network for image super-resolution [24], and Ma et al. proposed a method of constraining the super-resolution image structure using a gradient map [25]. Xu et al. of China Taiwan MediaTek Inc. proposed a dynamic convolutional network to realize super-resolution restoration of multiple combinations of blur kernels and noisy images [26]. The super-resolution method based on pixel space produces more photo-realistic images, while the high-frequency information generated by the super-resolution method based on feature space is rich. By combining the advantages of the two methods in a deep fusion network, we can reconstruct high-quality key objects.

To solve the problem that video surveillance system in the field of public security lacks basic and intelligent management and analysis algorithms, we proposed a super-resolution method of surveillance video objects based on a deep fusion convolutional neural network. It assists police and other surveillance video analysts to track, identify, and analyze objects and investigate cases. By optimizing the

regression-based one-stage object detection algorithm, we can identify the objects in the surveillance video in real time, which assists the viewers to track the video objects and analyze the video content. We designed a keyframe selection algorithm for surveillance video. By analyzing the object category, number, and confidence degree in the video frame, a small number of video frames with significant changes in the object are selected from a large number of videos to assist the surveillance video viewer to quickly locate the objects and reduce the workload. We used the super-resolution algorithm based on the deep fusion network to reconstruct the determined object. It helps to improve the resolution of the key object, assist the surveillance video viewer to carefully check the details of the key object, and improve the quality of the surveillance video content analysis. The super-resolution method of surveillance video objects based on a deep fusion convolutional neural network constructed in this paper can be effectively used in the field of public security to assist police and security personnel to track and analyze surveillance video objects. And, it also provides an effective auxiliary tool for preventing and cracking down on violent terrorist crimes.

3. Preliminary

3.1. Object Detection and Super-Resolution Process. The regression-based one-stage object detection method takes the video frame as the input of the network and returns the position and category of the bounding box (BBox) at the output end. For each grid of the video frame, it predicts n BBox and c category information. Among them, each BBox contains four location information items (w, y, w, h) and one confidence information item. The BBox position information (x, y) is used to calibrate the center point of the BBox, and (w, h) is used to calibrate the width and height of the BBox relative to the video frame. Confidence predicts the confidence of the objects contained in the BBox and the accuracy of the BBox compared to the real object box. The confidence is defined by (1), where C means confidence, O means object (if there is an object, $\Pr(O) = 1$; otherwise $\Pr(O) = 0$), and $\mathrm{IOU}_{\mathrm{pre}}^{\mathrm{tru}}$ predicts the intersection over union between the BBox and the real box. Define A as the predicted frame and B as the real frame; S is the set of all frames; and $A, B \subseteq S \in \mathbb{R}^n$. IoU can be described by (2). According to the predicted category information of each grid of the video frame and the confidence of the BBox prediction, we can obtain the class-specific confidence score of each BBox, which is defined in (3). According to the class-specific confidence score, the BBox with a low score is filtered out by setting a threshold, and the remaining BBox is processed by non-maximum suppression (NMS) to obtain the final detection result.

$$C = \Pr(O) * \mathrm{IOU}_{\mathrm{pre}}^{\mathrm{tru}}, \tag{1}$$

$$\mathrm{IOU} = \frac{|A \cap B|}{|A \cup B|}, \tag{2}$$

$$CC = \Pr(\mathrm{Class}_i) * \Pr(O) * \mathrm{IOU}_{\mathrm{pre}}^{\mathrm{tru}}, \tag{3}$$

$$I_{\mathrm{LR}} = (I_{\mathrm{HR}} * k)\downarrow_s + n, \tag{4}$$

$$I_{\mathrm{LR}} = \mathrm{bicubic}(I_{\mathrm{HR}})\downarrow_s, \tag{5}$$

$$I_{\mathrm{SR}} = \mathrm{upsample}(\mathrm{Res}(\mathrm{Feature}(I_{\mathrm{LR}}))). \tag{6}$$

The super-resolution reconstruction method reconstructs a corresponding high-resolution image or video based on one or a series of low-resolution images (video frames). Since it is difficult to obtain a series of low-resolution images of the same reconstruction object, current research mainly focuses mainly on single-image super-resolution. Video super-resolution reconstruction needs to comprehensively utilize information in both space and time dimensions and improve the reconstruction effect of the center frame by using the information of adjacent low-resolution frames. The degradation process of an image or video frame can be described by (4), where I_{LR} represents a low-resolution image, I_{HR} refers to a high-resolution image, $\downarrow_s$ represents a downsampling scale, k is a blur kernel, and n represents noise. In real-world scenes, there are many types of downsampling scales, blur kernels, and noises. To facilitate supervised deep learning, the SR method generally uses bicubic interpolation to downsample high-resolution images and obtains paired images (LR, HR) for learning the mapping relationship from low resolution to high resolution. The process of super-resolution reconstruction is shown in (6). It can be typically divided into three steps. Firstly, we use the convolutional layer (Feature) to extract the low-resolution initial features. Then, we employ the residual network (Res) to learn the nonlinear mapping from low resolution to high resolution. Finally, we use the upsampling layer (Upsample) to enlarge the low resolution to a specified scale and reconstruct the super-resolution image.

3.2. Work Process Overview. Video surveillance systems widely distributed in communities, roads, streets, and commercial places (supermarkets) are a powerful guarantee of public safety. The combination of traditional video surveillance systems and artificial intelligence technologies can effectively improve the accuracy of video content analysis and abnormal behavior warning, greatly improve the work efficiency of surveillance video viewers, and reduce labor costs. The super-resolution framework of the video detection object based on deep learning is shown in Figure 1. By fusing the object detection algorithm based on deep learning and the super-resolution algorithm, we can track the object of the surveillance video in real time, pick out the keyframes that change significantly, and perform super-resolution reconstruction of the key object. This provides the police and judges with clear and high-resolution objects, which can assist in the investigation and review of related cases. The method proposed in this paper can solve the problems of traditional video surveillance, provide effective assistance to surveillance video viewers, and effectively serve the public security field.

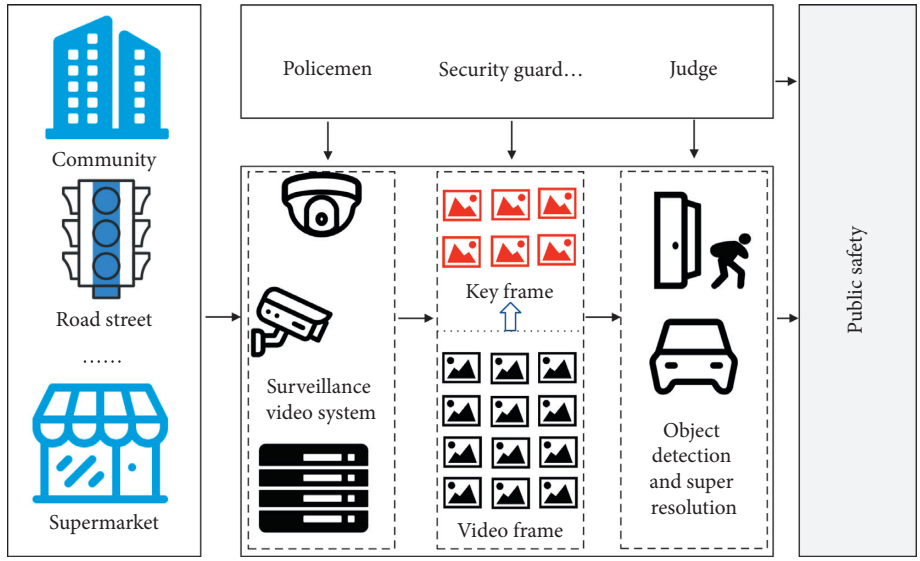

Figure 1: The work process of the video detection object super-resolution.

The super-resolution method of video detection object based on deep learning is mainly divided into three steps. Firstly, we use a regression-based object detection algorithm to perform real-time object detection on surveillance videos obtained from communities, roads, streets, supermarkets, and other places. Marking the scope, category, and confidence of the object in real time in the video frame assists police, security, and other video viewers to track the object. Then, we use a keyframe selection algorithm to select frames with significant changes from a large number of videos and select the key information from a large amount of redundant information to help viewers analyze video content quickly and efficiently. We finally use the super-resolution algorithm to reconstruct, improve the resolution of, and identify the key object.

4. Video Detection Object Super-Resolution

4.1. Video Detection Object Super-Resolution Network. To improve the safety of public places such as communities, supermarkets, roads, and streets and assist surveillance video viewers to quickly and effectively locate objects, we design a video object detection super-resolution algorithm that combines object detection and super-resolution reconstruction. As shown in Figure 2, we use a regression-based object detection algorithm for real-time object detection and a deep back-projection network for super-resolution reconstruction. It should be noted that, for the convenience of drawing, we did not show the details of each functional module in detail in Figure 2, such as the feature fusion module of different scales of object detection and the feature prediction module of video frames, and the asymmetric recursive structure of super-resolution.

The object detection network consists of three parts. Firstly, we use the backbone built by modules such as focus, CBL (conv, BN, Leaky ReLU), and CSP to extract video frame features. The focus module is used to slice the input image and transform the dimension of the feature map. For example, we can slice the $608 * 608 * 3$ image into a $304 * 304 * 12$ feature map according to RGB. Then, it passes through a conv layer containing 32 filters and transforms it into a $304 * 304 * 32$ feature map. The CBL module is composed of the conv layer, the BN layer, and the Leaky ReLU activation function, and CBL is the basic module unit of the object detection network. The CSP module is composed of CBL, residual unit, conv layer, BN layer, and Leaky ReLU activation function. It is the feature splitting module unit of the object detection network. Secondly, we use the multiscale fusion module constructed by CBL, CSP-1, CSP-2, and other modules to extract fusion features of different scales and strengthen the ability of network feature fusion (the structure of the CSP-1 and CSP-2 modules is shown in Figure 3, and the relevant descriptions are provided in Section 4.2). Finally, we use GIoU loss to solve the problem that the detection frame and the real frame do not intersect to optimize the detection frame and speed up the convergence of the model.

Before we perform super-resolution reconstruction, we will select the detection object to avoid wasting computing power to super-resolve a large number of worthless objects. The selection process of the object is described in detail in Section 4.2. Our super-resolution reconstruction algorithm uses a deep asymmetric recursive back-projection network to reconstruct key objects. The super-resolution network consists of three parts. Firstly, we employ two conv layers to extract the initial features of the image. Then, we use the upper and lower iterative sampling layers to learn the mapping relationship between the low- and high-resolution image several times to obtain upsampling features of different levels. We set up a recursive structure to deepen the network level and learn advanced semantic features without increasing parameters. By sharing the downsampling layer and setting an asymmetric structure, we reduce the parameters of the network model. We fuse homology residuals

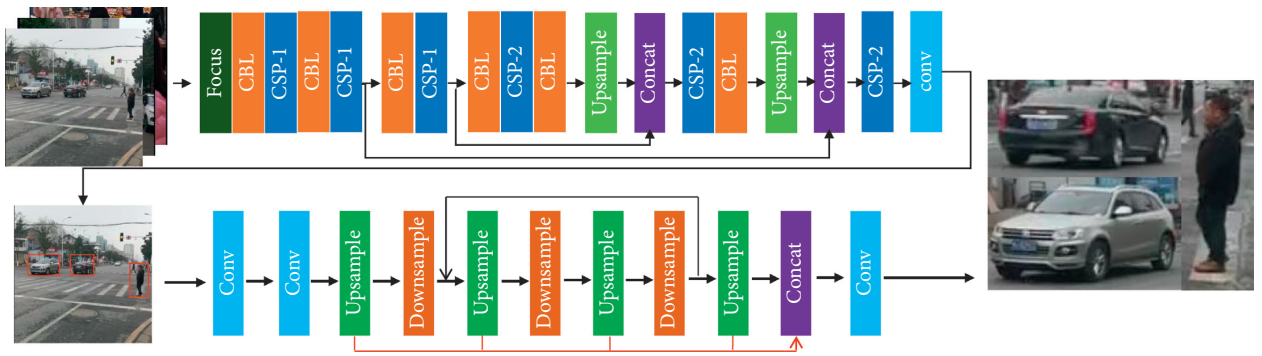

Figure 2: The video detection object super-resolution network.

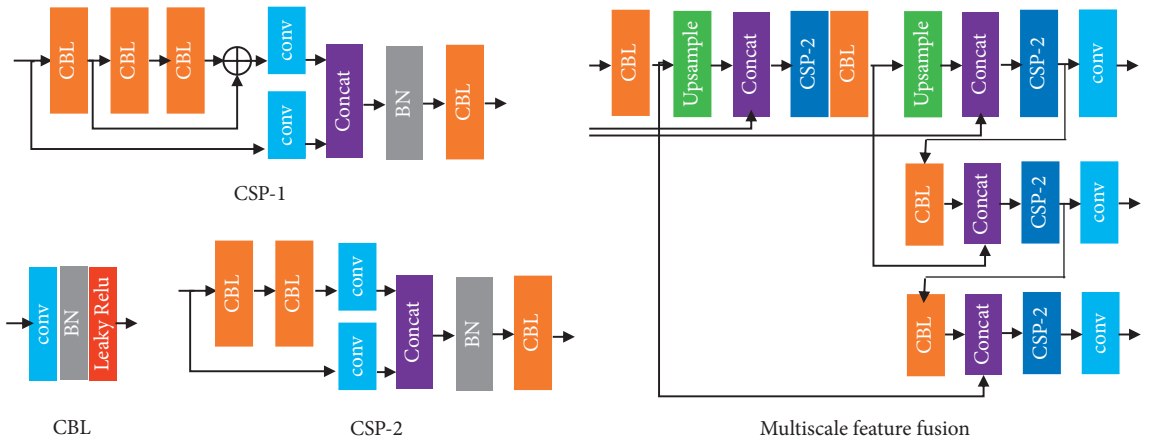

Figure 3: The core module of object detection network and multiscale feature fusion network.

and cascaded residuals in the upsampling unit to make full use of residual information while correcting projection errors. We concatenate cascade non-depth upsampling features to improve the high-frequency detail information of upsampling features [27, 28]. Finally, we use the conv layer to reconstruct the super-resolution video frame detection object.

4.2. Object Detection and Keyframe Selection Method. The object detection algorithm can assist police and security personnel in analyzing surveillance video objects and investigating security cases. It helps to track and monitor suspicious signs and objects in real time to avoid criminal harm. The object detection algorithm includes two core modules, CBL and CSP, in which CSP is composed of residual structure CSP-1 and convolution structure CSP-2, as shown in Figure 3. CBL is the basic module and consists of conv layer, BN layer, and Leaky ReLU activation function. The CSP-1 module includes two branches. The first branch includes a CBL feature extraction unit, two CBL residual mapping units, and a conv layer. The second branch directly uses a conv layer to extract the input low-level features, and then we fuse the output of the two branches through the concatenation layer. The CSP-2 module and the CSP-1 module share the same structure, and the difference is that the first branch of CSP-2 does not use the residual structure. The multiscale feature fusion module uses the top-down FPN structure to amplify the deep feature information and fuse it with the backbone features of different depths using the upsampling unit. We use the bottom-up PAN structure to downsample the upsampled features and merge them with neck features of different depths. The FPN structure integrates strong semantic features from top to bottom, and PAN integrates strong positioning features from bottom to top. The multiscale feature fusion module structure is shown in Figure 3.

Our object detection algorithm first uses backbone to extract the features of the input video frame; then uses the multiscale feature fusion module to strengthen the ability of network feature fusion; and finally uses GIoU loss to optimize the object frame, filter the object frame through DIOU_NMS, and output object frame information, classification information, and confidence information. The objective box optimization loss function is $\mathrm{GIoU} = \mathrm{IoU} - |A_c - U|/|A_c|$, where A_c is the smallest bounding box between the predicted box A and the real box B, and U is the union of the predicted box and the real box $U = A \cup B$. GIoU can distinguish different positional relationships between predicted frames and real frames with the same IoU and the same size and can optimize the situation where the prediction frame and the real frame do not intersect. DIOU_NMS also considers the overlap area and the center distance between the two boxes when filtering the object box.

DIOU_NMS will not delete two boxes with a far center point since they may be in two different objects. DIOU_NMS instead of NMS can optimize the detection accuracy of overlapping objects. It can improve the recognition accuracy of occluded objects without increasing the computational cost. Real-time object detection can assist surveillance video viewers to track objects. However, it still takes a lot of labor costs to face the massive amount of video information. Therefore, selecting keyframes with significant scene changes can effectively improve the efficiency of key object recognition. The keyframe selection algorithm we designed is shown in Algorithm 1.

According to the output of the surveillance video object detection algorithm, we can select frames from the video whose category, number, and detection frame have significant changes compared with the previous frame as keyframes. Static pictures often appear in surveillance videos (the object category and amount in the video have not changed). Detecting object super-scores for all static video frames will result in a large amount of redundant information, which not only wastes computing power but also is not conducive to video content analysis. As shown in Algorithm 1, our keyframe selection algorithm can use information such as the type, quantity, and confidence of the detection output to compare and analyze the difference between the current frame and the previous frame while detecting the object. If the surveillance video currently captures a static image, we believe that this frame has less valuable information and will not store keyframes. If any of the following three situations occurs, the keyframe selection algorithm can select a small number of frames with significant changes from the surveillance video: (1) The object category of the current frame changes relative to the previous frame; for example, the object of the previous frame has cars and trees, but the object of the current frame contains cars, trees, and people. (2) Compared with the previous frame, the amount of a certain category in the current frame has changed; for example, the previous frame has cars (5 cars), trees (3), and people (1), and the current frame has cars (5 cars), trees (3), and people (2). (3) Compared with the previous frame, the detection box of the current frame has changed. For example, the detection box of a car in the previous frame is ($b_x = 0.2$, $b_y = 0.7$, $b_w = 0.1$, $b_h = 0.15$), and the detection box of the same car in current frame is ($b_x = 0.5$, $b_y = 0.5$, $b_w = 0.25$, $b_h = 0.4$). Video viewers can understand the surveillance video content by viewing a few frames to improve work efficiency. The time complexity of the whole algorithm is $O(T_1 + N)$, where T_1 is the time of video object detection and N is the number of video frames.

4.3. Super-Resolution Method of Video Detection Object. The goal of the super-resolution method based on pixel space is to make the super-resolved image as close as possible to the real image at every pixel point in pixel space. These methods use L1 or L2 as the loss function and do not use adversarial training to generate video frames, and the reconstructed results are close to real-world video frames in pixel space. However, it is easy to lose high-frequency information, which leads to extremely smooth and fuzzy reconstructed video frames and poor visual perception quality. The feature space-based super-resolution method aims to make the feature space of the super-resolved image close to the real-world image. These approaches combine the perception loss, confrontation loss, and L1 and L2 as the loss function. They use adversarial training to generate video frames of high visual perception quality. However, these methods easily lead to deformation and distortion of super-resolution image structure. Therefore, we propose a super-resolution method based on the fusion of pixel space and feature space. The super-resolution method based on pixel space ensures the authenticity of video frames while the fusion of the loss function based on feature space improves the visual perception quality of video frames.

The super-resolution method based on the deep back-projection structure learns the mapping relationship between LR and SR several times in the reconstruction process through the up- and downsampling units placed in sequence. The structure of the up- and downsampling unit is shown in Figure 4. The upsampling unit includes 2 deconv layers and 1 conv layer. The first deconv layer enlarges the input LR feature map to a specified scale, and the conv layer transforms the enlarged feature map back to the original size and calculates the cascade projection error with the input LR of this unit and the homologous projection error with the original input LR. The second deconv layer amplifies the error information to a specified scale and adds it to the first deconv layer. It corrects the output of the upsampling unit by cascading and homologous projection errors. The downsampling unit includes 2 conv layers and 1 deconv layer. The structure and execution process of the downsampling units are similar to the upsampling unit. We did not use the homologous error in the downsampling unit and only used the cascaded error to correct the output of the downsampling unit.

To improve the effect of super-resolution reconstruction, we need to stack several upsampling and downsampling units to obtain high-level semantic features. However, stacking several upsampling units and downsampling units will result in a sharp increase of the model parameters, and the training and use of the model will become more difficult. As shown in Figure 4, we propose a recursive back-projection structure to increase the depth of the back-projection network without increasing the parameters. We input 2 sets of up- and downsampling units into the recursive loop structure and set the recursive hyperparameter to 8. We use the parameters of 2 upsampling and downsampling units to achieve the performance of 16 upsampling and downsampling units. Besides, we introduce an asymmetric structure based on recursive back-projection to further reduce the model parameters and improve the robustness of the model in practical applications. As shown in Figure 4, we concatenate the SR feature maps of upsampling units at different depths to reconstruct the output SR video frame to improve the effect of SR reconstruction. We use the downsampling units to transform the output of the upsampling units back to the original size. Therefore, we propose that all upsampling units share the same

Input: V# input video
Output: F_{KEY}# output keyframe
(1) **define** $\mathscr{B}$ **as** backbone of the object detection network
(2) **define** $\mathscr{N}$ **as** neck of the object detection network
(3) **define** P **as** prediction of the object detection network
(4) **define** $\mathscr{U}$ **as** output of the object detection network
(5) $\mathscr{U}$ = DIOU_NMS$(P(\mathscr{N}(\mathscr{B}(V))))$
(6) $\mathscr{U} \longrightarrow (C, N, B, P)$ # Choose category, quantity, BBox, confidence
(7) **for** $i = 1,\ldots\ n$ **do** #n is the number of video frames
(8) *if* $(C_i! = C_{i-1}$ *or* $N_i! = N_{i-1})$
(9) if $(P_i > 0.5)$
(10) *Storage* (F_i)
(11) **end for**
(12) $F_{KEY} \leftarrow (F_1, F_2, \ldots, F_k)$ #k is the number of key video frames
(13) **Return** F_{KEY};

Algorithm 1: Keyframe selection algorithm.

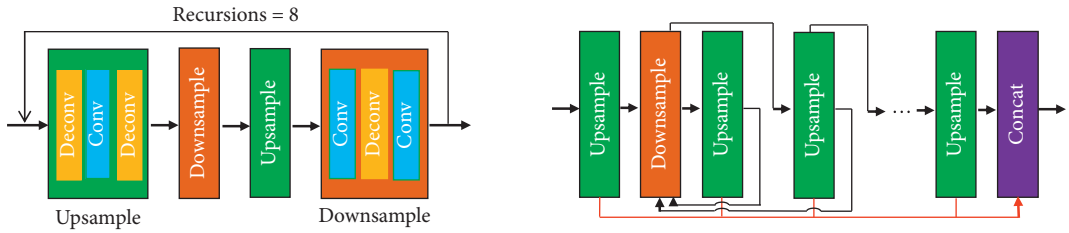

Figure 4: The sampling units, recursive structures, and asymmetric back-projection structure.

downsampling unit to construct a deep recursive back-projection network with an asymmetric structure.

$$L_T = \alpha L_1 + \beta L_{cx}, \tag{7}$$

$$L_1 = \frac{1}{N}\left\| \sum_{i=1}^{N} F_{\mathrm{DBPN}}\left(I_i^{\mathrm{LR}}\right) - I_i^{\mathrm{HR}} \right\|_1, \tag{8}$$

$$L_{cx}\left(\phi\left(I^{\mathrm{SR}}\right), \phi\left(I^{\mathrm{HR}}\right), l\right) = -\log\left(CX\left(\phi^l\left(I^{\mathrm{SR}}\right), \phi^l\left(I^{\mathrm{HR}}\right)\right)\right). \tag{9}$$

The image reconstructed by the super-resolution method based on pixel space lacks high-frequency detail information, and the texture features are not clear enough. We tried to directly generate super-resolution images using methods based on generative adversarial networks. Then, we found that these approaches can improve the clarity of visual perception. However, the generated SR image suffered from structural deformation and distortion, it is prone to mode collapse when processing real-world video frame super-resolution reconstruction, and the robustness of the model is not enough. Therefore, we propose to use the L1 loss function based on the pixel space and the contextual loss based on the feature space on the asymmetric depth recursive back-projection network. The contextual loss function is shown in (9), where $\phi(I^{\mathrm{SR}})$ represents the SR image feature map, $\phi(I^{\mathrm{HR}})$ represents the HR image feature map, and ϕ uses a pretrained 19-layer VGG network. We select the feature map of the 5th layer before max-pooling after the 4th layer of convolution. We have $CX(x, y) = (1/N)\sum_j \max_i CX_{ij}$, where CX_{ij} is the similarity of local features x_i and y_j. We decouple the video frame into a collection of local features and optimize the reconstruction of I^{SR} features by measuring the distance between the local feature distributions of I^{SR} and I^{HR}. Based on the asymmetric recursive back-projection SR network, we add contextual loss to optimize the reconstruction of high-frequency information in the feature space. Combining the advantages of the authenticity of the pixel space SR method and the high visual perception quality of the feature space SR method, we can reconstruct a high-quality video frame object with realistic visual perception.

5. Experimental Results

5.1. Implementation Details. We use PyTorch to build a super-resolution model for video detection objects. In the model training phase, we use a high-performance server to train and verify the model. In the model testing phase, we use a personal computer for testing to ensure the usability and robustness of the model in actual application scenarios. The high-performance server runs on Linux operating system, and the GPU is NVIDIA TITAN Xp. The operating system of the personal computer for testing is Windows 10 with the i5 CPU core. We have taken some videos that do not involve personal privacy in communities, roads, streets,

supermarkets, and schools. We then extracted the video frames for incremental training of the object detection module. We selected some high-resolution video frames from the shooting videos and added them to the DIV2K data set, and then we expanded the training set to 1,000 and the validation set and test set to 200.

$$\mathrm{PRE} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}, \tag{10}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}. \tag{11}$$

The evaluation metrics of the super-resolution method for video detection objects include the accuracy and speed of object detection, the peak signal-to-noise ratio, and the structural similarity of SR images. We use the mean Average Precision (mAP) to evaluate the object detection accuracy, which mainly includes the precision rate and the recall rate, as shown in (10) and (11). The numerator of precision is the number of detection frames with IoU greater than 0.5, and the denominator is the sum of the detection frames with IoU greater than 0.5 and less than 0.5. The numerator of recall is the number of detection frames with IoU greater than 0.5, and the denominator is the sum of detection frames with IoU greater than 0.5 and the true frames that are not detected. We use FPS (frames per second) to evaluate the object detection speed, which is the number of video frames that the model can process per second. We require the model to reach 30FPS on an ordinary PC. The PSNR is based on the sensitivity of pixel error, and we use it to measure the mean square error of the SR image and the real image. The unit of PSNR is dB. The larger the value, the smaller the distortion. The SSIM (structural similarity) measures the similarity of images from three aspects: brightness, contrast, and structure. The value range of SSIM is [0, 1]. The larger the value, the smaller the image distortion.

5.2. Object Detection Experiment Results. We first obtained some surveillance videos online, but we found that the scenes involved in these surveillance videos are relatively single. To verify the effectiveness of our method in the field of public safety, we have taken some surveillance videos that do not involve personal privacy in communities, roads, streets, supermarkets, schools, and other places. The video frame rate is 30FPS, the video resolution is 720p (16 : 9), and the 10-second video size is about 1.5 MB. We use regression-based object detection algorithms to detect objects in surveillance videos in real time. The object detection algorithm assists police and security personnel to track suspicious objects in real time and prevent crimes. On the other hand, the output information of object detection lays the foundation for the keyframe selection and key object super-resolution reconstruction. We use surveillance videos of three scenes: supermarkets, communities, and roads, to verify the effect of the object detection algorithm, as shown in Figure 5.

In the supermarket scene, the main detection object of surveillance video is people. Our object detection algorithm can accurately detect people in real time. We can monitor the valuables in the supermarket in real time by expanding the training data set. When the type, quantity, and confidence of the valuables change, we can promptly remind the supermarket management personnel. The real-time detection of people and valuables can assist supermarket managers to track suspicious persons in time and prevent the theft of valuables. In the community scenario, the main monitoring objects of surveillance video are people and cars, which helps in ensuring the safety of people and property. Since there are many small objects in community video surveillance and the occlusion between vehicles in the parking area is more serious, the object detection algorithm needs to recognize small objects and occluded objects well. We can see that our object detection algorithm can well identify small objects and obscured objects from Figure 5. It helps the community managers to track suspicious persons and vehicles in real time and maintain community safety. In road scenes, the main monitoring objects of surveillance video are people, vehicles, etc. We use object detection algorithms to assist traffic police to track suspicious people and vehicles in real time and maintain public safety.

5.3. Keyframe Selection Experiment Results. When investigating criminal cases, the police often need to find a very small number of frames with key information from a large number of videos. Selecting the keyframes manually requires a lot of costs and material resources and is inefficient. Based on the object detection algorithm, we designed a keyframe selection algorithm to pick out keyframes with significant changes in category, number, and confidence from a large amount of redundant video frame information. We selected 12 keyframes from the community surveillance video (540 frames), 14 keyframes from the supermarket surveillance video (450 frames), and 16 keyframes from the road surveillance video (510 frames). Our keyframe selection algorithm can filter a large number of redundant information frames generated by static pictures. Figure 6 shows some of the keyframes and key objects we selected from the scenes of community, supermarket, and road.

In the community scenario, the main monitoring object of the surveillance video is the ground parking lot. When there are no pedestrians and other objects on the ground parking lot and surrounding roads, the surveillance video is in a static image so no keyframe is extracted. As shown in Figure 6, when there are pedestrian objects in the community parking lot, the classes of surveillance video objects increase and the keyframes of the classes are extracted. When the pedestrian object is getting closer to the camera, the object detection frame and confidence become larger, and the confidence keyframe is extracted. When the number of pedestrians changes, the number keyframe is extracted. According to the extracted keyframes, we further extract key objects by setting the confidence threshold for the following super-resolution reconstruction. In the supermarket scene, the surveillance video mainly monitors the number of consumers before the product, and the frame that records the change in the number of consumers is the keyframe. By

Figure 5: Object detection results of supermarkets, communities, and roads.

setting the confidence threshold, we can monitor and record the change frame of the consumer before a certain valuable product in real time. In road scenes, surveillance video mainly monitors vehicles and pedestrians, and frames that record the change in the number, type, and confidence of vehicles and pedestrians are keyframes.

5.4. Keyframe Object SR Experimental Results. Surveillance videos suffer from low resolution due to hardware technology, hardware cost, shooting environment, network transmission, and so on. To observe the details of key information, police and security personnel often need to magnify the image 4 times, 8 times, or even higher. If you directly magnify an image, the magnified image will miss a lot of high-frequency information and suffer from low quality and poor visual perception, making it difficult to recognize. The pixel-space-based super-resolution method can reconstruct SR video frames close to the real image by optimizing the distance between the super-resolved image and the real image in the pixel space. However, the reconstructed frames may lack a lot of high-frequency detail information, and the visual perception of texture features is unclear. Therefore, on the basis of super-resolution method based on pixel space, we further integrate the contextual loss based on feature space. By optimizing the local features of the super-resolved image and the real image in the feature space, we can improve the high-frequency information of the reconstructed video frames and obtain high-quality SR images with clear visual perception. It should be noted that, on the basis of object detection and keyframe selection, we select key objects from the keyframes according to the confidence threshold. The resolution and storage capacity are often very small and lack high-frequency information.

Figure 6: Some keyframes and key objects selected by the keyframe selection algorithm.

Figure 7: Comparison of super-resolved key objects (tricycles) in road surveillance video.

ESRGAN is the champion solution of the PIRM2018-SR (region 3) super-resolution competition, which can generate natural and detailed SR images. SPSR adds gradient map branches based on ESRGAN to constrain the structure of the generated image, which improves the authenticity of the SR image and the quality of visual perception. Therefore, the SPSR and ESRGAN [29], which can generate rich high-frequency information, were selected as our comparison methods. Figures 7–10 show the SR reconstruction results of the key objects in the three scenarios of roads, communities, and supermarkets.

We designed an asymmetric deep recursive back-projection network. We first constructed a back-projection structure in which the upper and lower sampling units are stacked in sequence by simulating the human visual system. Then, we used the cascading projection error and homologous projection error to correct the loss of the upsampling and downsampling units. In addition, we improved the SR reconstruction effect by cascading the outputs of upsampling units of different depths. We designed a recursive loop and asymmetric structure to improve the SR reconstruction effect without increasing the parameters. In the road scene, we choose a tricycle and a taxi

Figure 8: Comparison of super-resolved key objects (taxies) in road surveillance video.

Figure 9: Comparison of super-resolved key objects (customers) in supermarket surveillance video.

Figure 10: Comparison of super-resolved key objects (children) in community surveillance video.

for reconstruction and make comparison with the SPSR and ESRGAN. From Figure 7, we can see that the SPSR method reconstructs the SR image and produces a lot of artifacts, and many lines do not exist in the LR. The effect of ESRGAN reconstruction is typically similar to our method, but the edge part of the SR video frame reconstructed by our method is sharper and the visual perception is clearer. The effect of the reconstructed taxi in Figure 8 is the same as that of Figure 7. The SR video frame reconstructed by SPSR has artifacts on the underside of the taxi. Our method not only is free of artifacts but also has better clarity. In the community and supermarket scenes, we selected a girl, a boy, and two customers in the supermarket as the key objects for SR reconstruction. In Figure 10, we can see that the SPSR reconstructs artifacts in many places, such as the little girl's face and arms. When using real-world video frames for reconstruction, SPSR lacks generalization ability and robustness, while ESRGAN suffers from a mass of parameters, difficult model training, and insufficient visual perception quality. Our model uses the least parameters and achieves the best reconstruction effect.

6. Conclusion

Maintaining and ensuring the safety of the public domain constitute the foundation of safe and smart cities. Surveillance equipment widely distributed in the public domain can detect suspicious signs and objects in time, can inquire information about suspected vehicles and personnel through

surveillance video, and can provide objective litigation evidence during the investigation and interrogation stage of the case. Traditional video surveillance systems rely on manual analysis of video content, which makes it difficult to effectively play the role of video surveillance. Therefore, we combine the traditional video surveillance system with artificial intelligence technology. Firstly, we fuse the object detection algorithm, keyframe selection algorithm, and super-resolution reconstruction algorithm to construct a super-resolution framework for video detection objects, which provides an effective auxiliary tool for the police personnel. Then, we employ the object detection algorithm to detect the object of the surveillance video in real time, which assists the police and other personnel to track the suspicious object. Besides, it helps select the type, quantity, and confidence of the surveillance video frames that have changed from a large amount of redundant information. Selecting key information from a large amount of redundant information can improve the efficiency of video analysis and reduce manual workload. Finally, to solve the problem that low-resolution surveillance video cannot be used effectively, we select the key object from the keyframe for super-resolution reconstruction. Combining the advantages of the pixel-based super-resolution method and the feature space loss function, we designed an asymmetric deep recursive back-projection network that can reconstruct the key objects with high resolution. The next step of our work is to realize super-resolution reconstruction of video detection objects under noise, blur, and other interference in the surveillance video.

Data Availability

The video, keyframe and key object images, and SR results data used to support the findings of this study have been deposited in the GitHub repository (https://github.com/yunfeiyoda/Video-Detection-Object-Super-resolution-.git).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the Hunan University of Arts and Science for providing them with high-performance servers. All the training and testing of the super-resolution model of the video detection object were completed on a high-performance server. This work was supported by the National Social Science Fund of China (no. 20&ZD120).

References

[1] K. Guo, B. Hu, J. Ma et al., "Toward anomaly behavior detection as an edge network service using a dual-task interactive guided neural network," *IEEE Internet of Things Journal*, vol. 99, 2020.

[2] L. Fang, Y. Li, X. Yun et al., "THP: a novel authentication scheme to prevent multiple attacks in SDN-based IoT network," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5745–5759, 2019.

[3] T. Mita, T. Kaneko, and O. Hori, "Joint haar-like features for face detection," in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)*, vol. 2, pp. 1619–1626, IEEE, Beijing, China, December 2005.

[4] L. Cuimei, Q. Zhiliang, J. Nan et al., "Human face detection algorithm via haar cascade classifier combined with three additional classifiers," in *Proceedings of the 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, pp. 483–487, IEEE, Yangzhou, China, October 2017.

[5] S. Ren, K. He, R. Girshick et al., "Faster r-cnn: towards real-time object detection with region proposal networks," 2015, http://arXiv.org/abs/1506.01497.

[6] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, http://arXiv.org/abs/1804.02767.

[7] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[8] C. Dong, C. C. Loy, K. He et al., "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[9] C. Ledig, L. Theis, F. Huszár et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, Honolulu, HI, USA, July 2017.

[10] K. Guo, N. Li, J. Kang et al., "Towards efficient federated learning-based scheme in medical cyber-physical systems for distributed data," *Software: Practice and Experience*, 2020.

[11] R. Girshick, J. Donahue, T. Darrell et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, IEEE, Columbus, OH, USA, June 2014.

[12] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, IEEE, Santiago, Chile, December 2015.

[13] K. Guo, Y. Wang, J. Kang et al., "Core dataset extraction from unlabeled medical big data for lesion localization," *Big Data Research*, vol. 24, Article ID 100185, 2021.

[14] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, http://arXiv.org/abs/2004.10934.

[15] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: single shot multibox detector," *European Conference on Computer Vision*, Springer, Berlin, Germany, pp. 21–37, 2016.

[16] K. He, X. Zhang, S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, IEEE, Las Vegas, NV, USA, June 2016.

[17] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654, IEEE, Las Vegas, NV, USA, June 2016.

[18] X. Hu, H. Mu, X. Zhang et al., "Meta-SR: a magnification-arbitrary network for super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1575–1584, IEEE, Long Beach, CA USA, June 2019.

[19] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1671–1681, IEEE, Long Beach, CA USA, June 2019.

[20] X. Xu, Y. Ma, and W. Sun, "Towards real scene super-resolution with raw images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1723–1731, IEEE, Long Beach, CA, USA, June 2019.

[21] Z. Li, J. Yang, Z. Liu et al., "Feedback network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3867–3876, Long Beach, CA, USA, June 2019.

[22] S. Li, F. He, B. Du et al., "Fast spatio-temporal residual network for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10522–10531, IEEE, Long Beach, CA, USA, June 2019.

[23] J. Gu, H. Lu, W. Zuo et al., "Blind super-resolution with iterative kernel correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1604–1613, IEEE, Long Beach, CA, USA, June 2019.

[24] T. Dai, J. Cai, Y. Zhang et al., "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11065–11074, IEEE, Long Beach, CA, USA, June 2019.

[25] C. Ma, Y. Rao, Y. Cheng et al., "Structure-preserving super resolution with gradient guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7769–7778, IEEE, Seattle, WA, USA, June 2020.

[26] Y. S. Xu, S. Y. R. Tseng, Y. Tseng et al., "Unified dynamic convolutional network for super-resolution with variational degradations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12496–12505, IEEE, Los Alamitos, CA, USA, June 2020.

[27] K. Guo, H. Guo, S. Ren, J. Zhang, and X. Li, "Towards efficient motion-blurred public security video super-resolution based on back-projection networks," *Journal of Network and Computer Applications*, vol. 166, p. 102691, 2020.

[28] S. Ren, J. Li, K. Guo, and F. Li, "Medical video super-resolution based on asymmetric back-projection network with multilevel error feedback," *IEEE Access*, vol. 9, pp. 17909–17920, 2021.

[29] X. Wang, K. Yu, S. Wu et al., "Esrgan: enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 1–16, Glasgow, UK, August 2018.