sensors

MDPI

Article

Safety Helmet Detection Based on YOLOv5 Driven by Super-Resolution Reconstruction

Ju Han [1], Yicheng Liu [2,*], Zhipeng Li [2], Yan Liu [2] and Bixiong Zhan [1]

[1] China Construction First Group Construction & Development Co., Ltd., Beijing 100102, China
[2] College of Electrical Engineering, Sichuan University, Chengdu 610065, China
[*] Correspondence: liuyicheng@scu.edu.cn

Abstract: High-resolution image transmission is required in safety helmet detection problems in the construction industry, which makes it difficult for existing image detection methods to achieve high-speed detection. To overcome this problem, a novel super-resolution (SR) reconstruction module is designed to improve the resolution of images before the detection module. In the super-resolution reconstruction module, the multichannel attention mechanism module is used to improve the breadth of feature capture. Furthermore, a novel CSP (Cross Stage Partial) module of YOLO (You Only Look Once) v5 is presented to reduce information loss and gradient confusion. Experiments are performed to validate the proposed algorithm. The PSNR (peak signal-to-noise ratio) of the proposed module is 29.420, and the SSIM (structural similarity) reaches 0.855. These results show that the proposed model works well for safety helmet detection in construction industries.

Keywords: deep learning; real-time detection; safety helmet detection; super-resolution reconstruction; YOLOv5

1. Introduction

The construction industry is one of the most prone to safety accidents. Therefore, it is of great practical significance to study safety guarantees in this field. Over the past 20 years, it has experienced a decline in accident rates [1].

Head injuries can easily lead to a disability [2]. Reducing head injuries is the primary problem to ensure personnel security in the industry, and safety helmets are widely used to do so. The impact resistance of safety helmets can disperse the impact of rocks. Thus, many industrial regulations require workers to wear safety helmets during working. However, workers do not wear safety helmets as required due to the lack of safety awareness. Because of these reasons, safety accidents are frequent. According to relevant research statistics, 47.3% of people with head injuries in construction site accidents did not wear safety helmets [3]. Thus, it is very important to strengthen the supervision and management of workers. At present, the management of safety helmet wearing in most construction sites still requires manual monitoring. However, the efficiency of manual monitoring is very low due to the large working area and large flow of personnel. With the development of science and technology, video surveillance has become more and more popular. It is a vital part of safety helmet detection. Hence, the optimization of monitoring systems is widely studied [4,5]. Traditional video surveillance is mainly used for continuous monitoring. However, the final judgement still relies on humans' decisions, and the degree of automation is not enough. Intelligence algorithms are a method to enhance automation. They are widely used in image processing, prediction, robotics and so on [6–11]. Currently, evolutionary algorithms and deep learning are two important intelligent systems [12–14]. Among them, deep learning is widely used in image processing because of its strong learning ability [15–17], which can be combined with video surveillance to solve the problems of traditional methods [18].

check for updates

Citation: Han, J.; Liu, Y.; Li, Z.; Liu, Y.; Zhan, B. Safety Helmet Detection Based on YOLOv5 Driven by Super-Resolution Reconstruction. *Sensors* **2023**, *23*, 1822. https://doi.org/10.3390/s23041822

Academic Editor: Zahir M. Hussain

Received: 25 October 2022
Revised: 3 February 2023
Accepted: 4 February 2023
Published: 6 February 2023

Slow transmission processes are caused by large image data collected by cameras. Therefore, the obtained original images need to be compressed before transmitting them so that the process can be accelerated. After compressed image data is transferred to the terminal main control unit, they are input into a YOLOv5 target detection network. Since the image compression usually reduces the image resolution, it can be restored by super-resolution (SR) reconstruction.

Many scholars apply deep-learning algorithms to the field of image SR reconstruction. Convolutional neural networks were first introduced in SR reconstruction [19]. Researchers added residual structure into the network to ensure convergence [20]. The attention mechanism structure has been added to further improve the performance of reconstruction [21]. However, in the detection of safety helmets, there exist the following problems to be solved. First, we use cameras to acquire images, and they are placed all over the construction sites. However, the tasks of image detection need computer terminals to complete, which are placed in the control room. Hence, image acquisition and detection are asynchronous in construction sites. Many studies have focused on detection; however, data transmission has not seen adequate research. In addition, previous popular datasets collected on the web have not considered construction sites, such as COCO128 and so on. In other words, the pictures in these datasets are taken from other environments. This problem leads to poor detection results of the model in construction sites.

To solve those problems, a safety helmet detection model is proposed, which is driven by an SR reconstruction network based on YOLOv5. To enhance the learning ability of the network, the double residual channel structure [22] is applied to an SR reconstruction network in the proposed model. PCAB and MPAB are introduced into the depth feature-extraction module of the main channel to improve the result of overall reconstruction. In summary, the contributions of this paper are as follows:

(1) A novel super-resolution (SR) reconstruction module is designed to improve the resolution of the image before the detection module. Compared with existing methods [23,24], this method reduces the influence of high-resolution image transmission on detection speed in the construction industry.

(2) A novel CSP (Cross Stage Partial) module of YOLO (You Only Look Once) v5 is presented to reduce the information loss and the gradient confusion.

(3) Based on the proposed SR reconstruction network and YOLOv5 network, a novel end-to-end safety helmet detection model is proposed to make the proposed model reach an average precision (AP) of 79.1%.

(4) More than 13,000 images are collected for safety helmet detection in construction sites.

The organization of the rest of this paper is as follows. Section 2 introduces the work related to target detection and SR reconstruction. Section 3 shows the structure and the details of the Dual-Channel Residual SR reconstruction model and the improved YOLOv5 model. The experimental details and results are given in Section 4. The conclusions are given in Section 5.

2. Related Work

Traditional safety helmet detection methods choose features manually for target detection. They have strong subjectivity, poor generalization ability and limitations in engineering applications. With the continuous development of deep-learning algorithms, researchers are applying deep-learning algorithms to target detection and image SR reconstruction.

2.1. Target Detection

At present, research on target detection algorithms include two-stage and one- stage algorithms. Two-stage detection algorithms generate a series of candidate boxes as samples and then classify samples through a convolutional neural network. This kind of detection method has higher task accuracy but slower speed. Girshick et al. [25] proposed the region convolutional neural network, fast regions with CNN [26] and faster regions with CNN [27] algorithms. A one-stage detection algorithm directly regresses the category probability and position coordinate values of objects through a backbone network without using a region proposal network (RPN). This kind of detection method sacrifices detection precision but improves detection speed. In 2016, Liu et al. [28] introduced the multiscaledetection method and proposed the SSD (single shot multibox detection) detection algorithm, which improved the detection accuracy. Redmon et al. [29–31] proposed YOLOv1, YOLOv2 and YOLOv3. The YOLOv1 network model abstracted the target detection task into a regression problem for the first time, which greatly sped up the target recognition speed. The YOLOv2 network model introduced a new basic model named darknet-19 based on YOLOv1 to realize end-to-end training. Compared with YOLOv1, the YOLOv2 network model realizes more accurate, faster and more target categories. YOLOv3 introduced the feature pyramid network (FPN) algorithm, promoted the new basic model darknet-53 and integrated three feature layers of different sizes for detection tasks. It improved detection speed and accuracy, especially the detection performance of small targets. Bochkovskiy et al. [32] proposed YOLOv4. This detection network takes CSP darknet-53 as the backbone network and uses the PANET path aggregation algorithm. As a result, it improved the detection accuracy of the model. In 2020, Jocher et al. [33] proposed YOLOv5. This network model adds a focus structure to the backbone network of YOLOv4 to obtain a balance between detection speed and accuracy. Carion et al. [23] proposed DETR for end-to-end object detection and brought transformers into the object detection fields. Recently, Wang et al. [34] proposed YOLOv7, which has achieved better accuracy and speed than YOLOv5.

2.2. SR Reconstruction

The image SR reconstruction algorithm is used to recover high-resolution images from one or more low-resolution images. Dong et al. [19] proposed SR Convolution Neural Networks (SRCNNs). SRCNNs effectively improve the results of image SR reconstruction compared with traditional image SR algorithms. However, the network is relatively simple, and the convergence speed is slow during the execution of the algorithm. In subsequent research, researchers added a residual structure to the convolution network to effectively solve the above problems. Kim et al. [23] proposed the VDSR network and increased the number of layers of the CNN to 20. The residual structure and CNN are embedded into image SR reconstruction, and the image reconstruction result is improved. Li et al. [20] proposed a multiscale residual network (MSRN). This network includes image multiscale features in the residual structure to further improve the image reconstruction result. Zhang et al. [35] proposed the residual channel attention network SRCAN. This network applies a channel attention mechanism to the image SR problem and achieves a better reconstruction effect than previous algorithms. Lu et al. [36] presented a novel recursive unit for SR reconstruction fields to force models to learn more details by learnable up-sampling methods. Liu et al. [37] proposed an attention-based approach to discriminate between texture areas and smooth areas.

3. Materials and Methods

In this paper, the proposed safety helmet detection model is designed based on a Dual-Channel Residual SR reconstruction module and an improved YOLOv5 module. The overall architecture is given in Figure 1. I_{LR} means the input image features and I_{SR} means the reconstructed image features. The two submodules in this figure are addressed as follows.

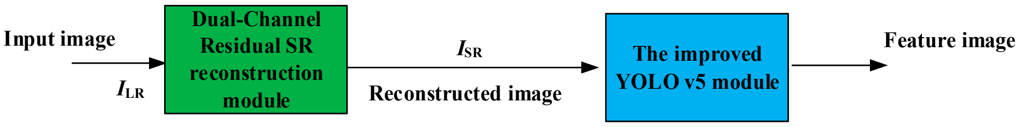
Input image
I_{LR}
Dual-Channel Residual SR reconstruction module
I_{SR}
Reconstructed image
The improved YOLO v5 module
Feature image

Figure 1. The overall architecture of the proposed module.

3.1. Dual-Channel Residual SR Reconstruction Module

The SR reconstruction module consists of three modules: a shallow feature extraction module, a depth nonlinear feature mapping module and an up-sampling reconstruction module. The specific structure is shown in Figure 2.

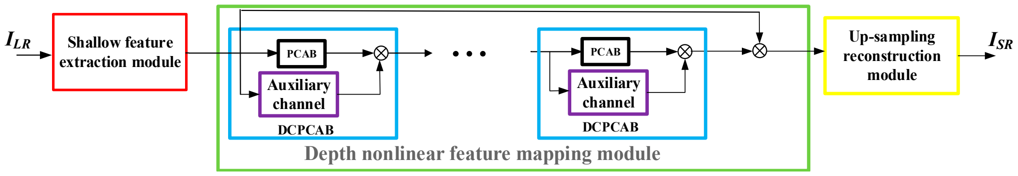
I_{LR}
Shallow feature extraction module
PCAB
Auxiliary channel
DCPCAB
PCAB
Auxiliary channel
DCPCAB
Depth nonlinear feature mapping module
Up-sampling reconstruction module
I_{SR}

Figure 2. The structure of the Dual-Channel Residual SR reconstruction module.

The shallow feature-extraction module is an ordinary convolution layer. As shown in Figure 2, the shallow extracted feature F_L is obtained from the input original low-resolution image I_{LR} through this module.

The depth nonlinear feature-mapping module is composed of several dual-channel pixel-channel attention blocks (DCPCABs). Each DCPCAB is shown in Figure 3. In this figure, the main channel of the DCPCAB module is composed of several pixel attention blocks (PABs), a channel attention (CA) block and a convolution layer. In this paper, the number of PABs in DCPCAB is two. The auxiliary channel is composed of two convolution layers and an adaptive structured convolution block.

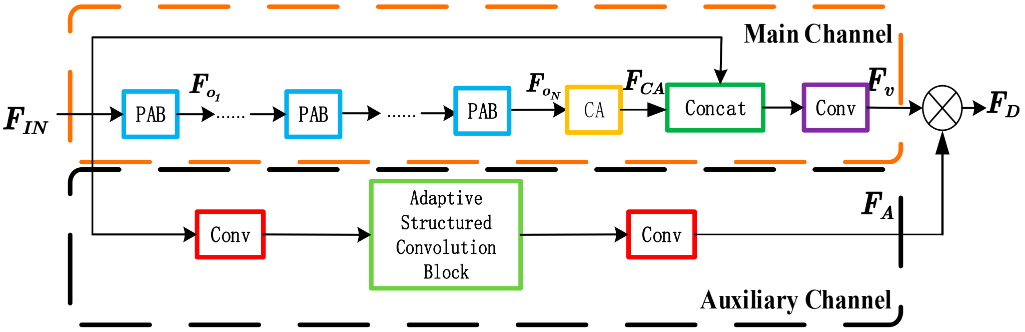

Figure 3. The DCPCAB architecture.

The architecture of PAB is shown in Figure 4. In this figure, F_{in} is the input feature, and it is put into three branches x, y and z. One convolution layer with a 1 × 1 kernel size is adopted in branch x to reduce the input feature F_{in} to the output feature Fx. In branch y, the input features are first fed into a convolution layer with a 1 × 1 kernel size for dimension reduction and then put into a convolution layer with a 3 × 3 kernel size for feature extraction to obtain the output feature Fy. In branch z, the input feature is first fed into a convolution layer for dimension reduction. The reduced dimension feature is input into the pixel attention (PA) mechanism network for pixel-level feature weighting. A convolution layer is adopted for feature extraction to obtain the output feature given as

$$\boldsymbol{F}_z = conv_{3\times3}(\boldsymbol{F}_{PA}(conv_{1\times1}(\boldsymbol{F}_{in}))) \tag{1}$$

where $conv_{3\times3}$ represents the convolution operation with a 3 × 3 kernel size. F_{PA} can be given by

$$F_{PA}(conv(F_{in})) = conv_{1\times1}(conv_{1\times1}(F_{in})) * \delta(conv_{1\times1}(conv_{1\times1}(F_{in}))) \tag{2}$$

where δ is the sigmoid activation function.

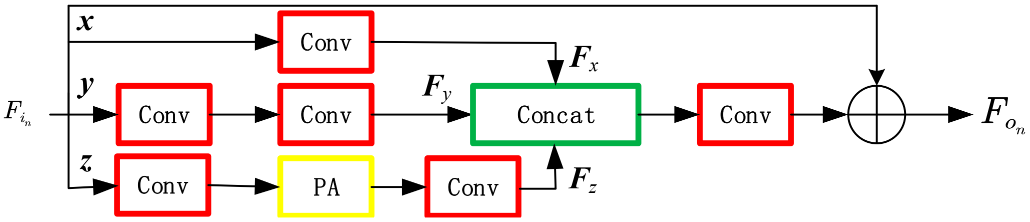

Figure 4. The PAB architecture.

The output features of the three branches are combined through a *concat* operation given by

$$F_o = conv_{1\times1}(concat(F_x, F_y, F_z)) + F_{in} \tag{3}$$

where *concat* means the operation of channel merging, which is used to merge the three features.

There are multiple PAB blocks in a DCPCAB module in Figure 3. The output F_{on} of each PAB can be iteratively calculated as

$$F_{on} = conv_{1\times1}(concat(F_{xn} + F_{yn} + F_{zn})) + F_{o(n-1)} \tag{4}$$

F_{CA} in Figure 3 is the CA output and is obtained by Equation (5). The output of **CA** can be given by

$$F_{CA} = conv_{1\times1}(F_{on}) \otimes sigmoid(conv_{1\times1}(F_{GAP}(F_{on}))) \tag{5}$$

where FGAP stands for the global average pooling operation and $\otimes$ means the pointwise multiplication operation.

The main channel output Fv in Figure 3 can be obtained by

$$F_v = conv_{1\times1}(concat(F_{CA} + F_{IN})) \tag{6}$$

where F_{IN} is the input of the DCPCAB.

The SR reconstruction operation always makes the edge information of the original images blurred or even deformed. The auxiliary channel module is introduced to broaden the width of the whole network to solve these problems. Adaptive structured convolution blocks are added in the auxiliary channel. The modules are adaptive to different expansion rates according to different image sizes. They can make the whole depth nonlinear feature-mapping submodule focus on the extraction of high-frequency features of the image. The operation of the auxiliary channel is given by

$$F_A = conv(F_{DC}(conv(F_{IN}), rate)) \tag{7}$$

where F_{DC} refers to the expansion convolution operation and *rate* refers to the expansion rate.

The final output FD in Figure 3 is given by

$$F_D = F_A \otimes F_v \tag{8}$$

The output of module *Fout* in Figure 2 can be calculated by

$$F_{out} = conv(F_{DN}) + F_L \tag{9}$$

where F_{DN} means the output of the final DCPCAB.

The up-sampling reconstruction module in Figure 2 can be given by

$$I_{SR} = conv(F_{up}(F_{out})) \tag{10}$$

where F_{up} represents the up-sampling operation and I_{SR} represents the results of the up-sampling reconstruction module.

To optimize the proposed **SR** reconstruction network, a loss function is adopted as

$$L1 = \frac{1}{k}\sum_{i=1}^{k}||\boldsymbol{I}_{SR} - \boldsymbol{I}_{HR}||_1 \tag{11}$$

where $|| \, ||_1$ means an $L1$ norm, k represents the number of training pictures and I_{HR} represents the corresponding high-resolution image of the I_{SR}. In this paper, the final loss of the best results is 0.0034, and the meaning of it is the MAE of the resolution of the reconstructed image and high resolution image.

The receptive field and the speed of the SR reconstruction models can be improved by the dual-channel residual structure. The number of parameters in the SR reconstruction model can be reduced by the PCAB structure. In other words, the PCAB structure can make the model more lightweight.

Remark 1: Here, the proposed Dual-Channel Residual SR reconstruction model is compared with SRCNN and SRGAN. Neither of these models consider the receptive field or the speed of the SR reconstruction model. SRCNN effectively improves the results of image SR reconstruction compared with traditional image SR algorithms. However, the network is relatively simple, and the convergence speed is slow during the execution of the algorithm. SRGAN [38] considers restoring fine-grained texture details. To improve the receptive field and the speed of the SR reconstruction model, the dual-channel is used in the proposed model. The number of parameters in the SR reconstruction model can be reduced by the PCAB structure. When other architectures were used instead of the dual-channel structure, the number of parameters we need to train must be the product of multiple dimensions. But the dual-channel can halve the number of parameters by introducing dual channels. In other words, the PCAB structure can make the model more lightweight.

3.2. The Improved YOLOv5 Module

YOLOv5 is an improved version algorithm based on YOLOv4 proposed by the Ultralytics LLC company. It is a network with excellent detection accuracy and speed in a single-stage detection network. YOLOv5 has a good detection effect on Pascal visual object classes (Pascal VOC) and common objects in context (COCO) target detection tasks, so YOLOv5 is selected as the detection network.

The YOLOv5 network structure is divided into four parts: the input port, backbone network, neck part and prediction part. The structure is shown in Figure 5 [33].

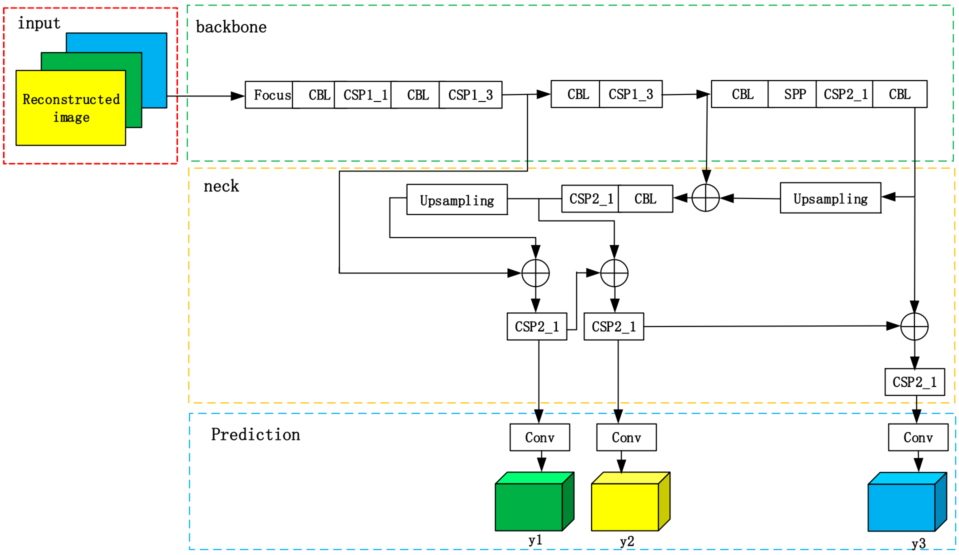

Figure 5. The structure of the improved YOLOv5 module.

The input port is used to mosaic random images to enrich the datasets, calculate the adaptive anchor frame and zoom images adaptively. The backbone mainly adopts a focus structure and cross-stage partial (CSP) structure to obtain features. The focus in the backbone is used to slice the input image data. The structure combining three multiscale pooling layers is used to improve the receptive field of the network while minimizing the loss of speed. It is helpful for the network to extract the important image features, reduce the image loss caused by early image processing and further improve the detection accuracy of the model. The structures of CSP and CBL are shown in Figure 6.

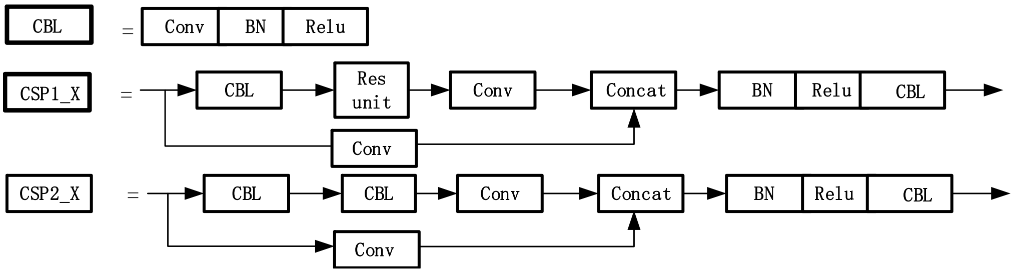

Figure 6. The structures of CBL, CSP1_X and CSP2_X.

The CSP1_*X* module is improved in two parts in Figure 7. The original CSP structure of YOLOv5 can lead to problems such as information loss and gradient confusion. Therefore, we use the LSandGlass module to replace the Res unit residual module in YOLOv5 and the 3 × 3 depth space convolution layer. The LSandGlass is different from the bottleneck structure with deep spatial convolution in China construction, 3 × 3 deep space convolution Dwise layers are moved to both ends of the residual path with high dimensional representation and the CBL blocks are stated in the mid. Two-deep convolution can encode more spatial information and make more gradients propagate across multiple layers, thus reducing information loss.

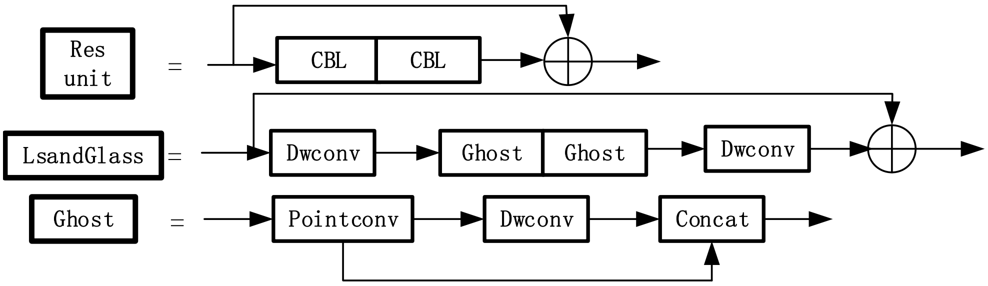

Figure 7. The structure of the Res unit, LSandGlass and Ghost.

Dwconv is moved to both ends of the residual path with high-dimensional representation to realize gradient propagation across multiple layers and reduce the loss of information. Considering such processing can increase the overall computation, the Ghost module is used to replace the CBL module of the bottleneck module in CSP. This scheme is adopted to reduce the computation and to compress the model size compared with the original 3 × 3 standard convolution.

The neck module of YOLOv5 uses the structure of feature pyramid networks (FPN) and pyramid attention networks (PAN). The prediction module contains the bounding box loss function and the non-maximum suppression (NMS) function. YOLOv5 uses the binary cross entropy loss function to calculate the loss of category probability and target confidence score. In the experiment, CIOU loss is selected as the bounding box loss function. The related formulas are given as [39]

$$CIOU\ Loss = 1 - \left(IOU - \frac{{d_1}^2}{{d_2}^2} - \beta\theta \right) \tag{12}$$

$$\beta = \frac{\theta}{(1 - IOU) + \theta} \tag{13}$$

$$\theta = \frac{4}{\pi^2} \left(tan^{-1} \frac{W^g}{h^g} - tan^{-1} \frac{W}{h} \right)^2 \tag{14}$$

where d_1 represents the Euclidean distance between the prediction box and the centre point of the target box and d_2 represents the diagonal distance of the minimum circumscribed matrix. $\frac{W^g}{h^g}$ represents the aspect ratio of the target frame, and $\frac{W}{h}$ represents the aspect ratio of the predicted frame.

***Remark 2*:** Here, the improved YOLOv5 model is compared with the original YOLOv5 model. The original YOLOv5 model did not consider problems such as information loss and gradient confusion. The proposed YOLOv5 model uses the LSandGlass module to replace the Res unit residual module in YOLOv5 to solve these problems. Considering that such processing can increase the overall computation, the Ghost module is used to replace the CBL module to solve it.

4. Results

4.1. Experimental Setup

First, we collected the image datasets by ourselves, which all come from construction sites and depict safety helmets. We first obtained the videos from construction sites and then used the VOTT to get images from the videos. The time of each video is about 14s, and we cropped at 7 frames per second to obtain the experimental images. The number of the images in the datasets is about 13,000, and the resolution of each image is 610 × 480. Then, we used pure interpolation to resize the input images to get 2× low-resolution.

To realize fast and reliable results, the entire method was implemented on a workstation equipped with two NVIDIA TITANRTX GPUs and an Intel i9 CPU.

All coding work was based on Python 3.7 and PyTorch 1.7.

The initial learning rates of the SR and detection were 0.0001 and 0.01, respectively.

The training epoch times of SR and detection were all 100.

The prediction times of SR and detection were all 100.

The kernels of the SRCNN were 1 × 1, 5 × 5 and 9 × 9.

The number of residual block layers for the generator in SRGAN was 16, and the weights of the loss function for SRGAN was given as 1, 1 and 1.

The Adam optimizer was applied with a momentum of 0.9, and the batch size was 32.

The training and test datasets were collected by CSCEC-2020Z-10.

4.2. Metrics

The structural similarity values (SSIM) and the peak signal-to-noise ratio (PSNR) are used to measure the quality of the reconstructed images. Among them, the former is adopted to measure the difference between the original image and the SR reconstructed image. The latter is used to measure the difference between the original image and the SR reconstructed image.

PSNR is given by

$$PSNR = 10log_{10} \frac{255}{\frac{1}{3} \sum (MSE)_C} \tag{15}$$

where **MSE** is the indicator of the square error for the image.

SSIM is defined as

$$SSIM(I_0, I_1) = \frac{(2m_0 m_1 + c_1)(2\sigma_0 \sigma_1 + c_2)}{\left(m_0^2 + m_1^2 + c_1\right)\left(\sigma_0^2 + \sigma_1^2 + c_2\right)} \tag{16}$$

where I_0 and I_1 are the original and the reconstructed high-resolution images. m is the indicator of the mean, and σ is the variance. c_1 and c_2 are both constants. In this paper, c_1 is set as 0.01×255^2, and c_2 is set as 0.03×255^2.

Precision (**P**), Recall (**R**) and Average Precision (**AP**) are used to measure the detection tasks. Among them, Precision is used to describe the ratio of predicted positive examples to all positive examples, and it is calculated by

$$P = \frac{TP}{TP + FP} \times 100\% \tag{17}$$

where P, TP and FP indicate the precision, true positives and false positives, respectively.

Recall is used to describe how many of the positive samples were detected in the prediction, and it is given by

$$R = \frac{TP}{TP + FN} \times 100\% \tag{18}$$

where FN indicates false negatives.

Average Precision synthesizes P and R, which is calculated from the area under the precision–recall curve and can be given by

$$AP = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{19}$$

where TN indicates true negatives.

4.3. SR Reconstruction Experiments

The SR reconstruction experiments are trained on approximately 80% of the 13,000 original images. They are validated on about 10% of the images and tested on the remaining 10%. Examples of the input images are shown in Figure 8. The right three subfigures show the low-resolution input images, and the left three subfigures are the output images.

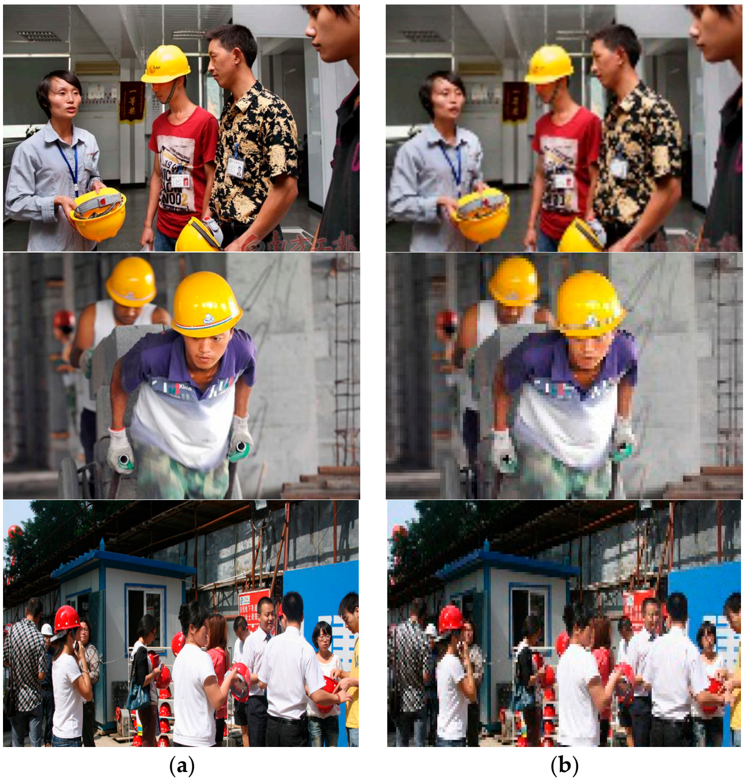

Figure 8. Examples of input images in the SR reconstruction module. (**a**) The output of the training datasets; (**b**) the input of the training datasets.

The different super-resolution reconstruction structures are shown in Table 1, and the experiments' results are shown in Table 2. Examples of the results are shown in Figure 9. It can be observed that the PSNR of the proposed method improved by 16.22% compared with SRCNN and by 5.36% compared with SRGAN. This means that the reconstructed images using the proposed method are closer to the original images. The SSIM of the proposed method improved by 9.76% compared with SRCNN and by 4.27% compared with SRGAN. These results mean that the proposed method can extract more image-structural information for human eyes.

Table 1. The difference of the three SR reconstruction methods.

Model	Structure
SRCNN [19]	Three parts: input, non-linear mapping and output
SRGAN [35]	Two parts: generator network and discriminator network
Dual-Channel Residual SR Reconstruction model	Three parts: input, Dual-channel module and output

Table 2. The results achieved by different SR reconstruction methods.

Model / Metrics	SRCNN [19]	SRGAN [35]	Dual-Channel Residual SR Reconstruction Model
PSNR	25.313	27.923	29.420
SSIM	0.779	0.820	0.855
Parameters' number	7,235,377	73,478,945	3,525,431

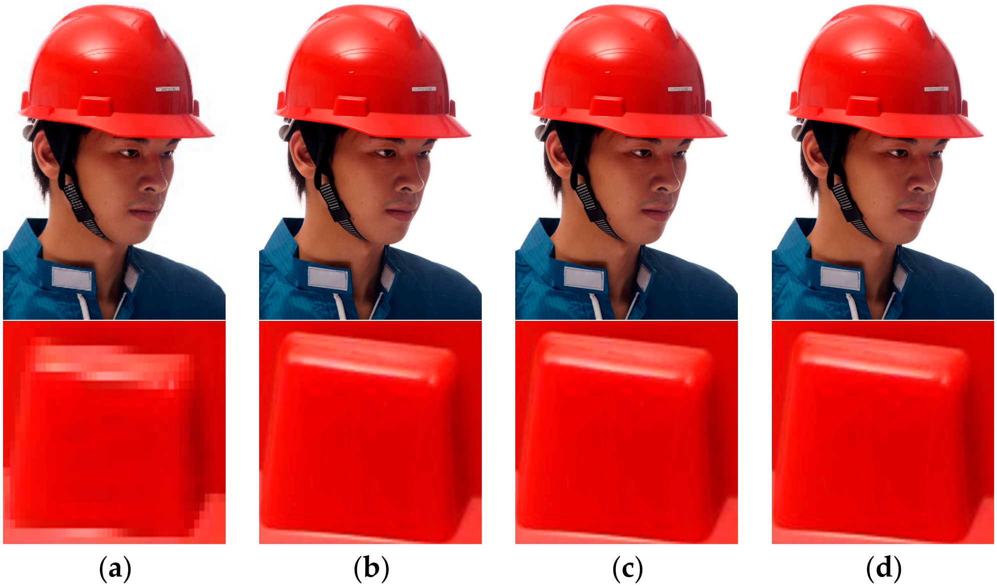

Figure 9. Examples of the results. (**a**) The input of the SR reconstruction module; (**b**) the output of SRCNN; (**c**) the output of SRGAN; (**d**) the output of the Dual-Channel Residual SR reconstruction module.

4.4. Safety Helmet Detection Experiments

To verify the advantage of the proposed model in safety helmet detection, we compare the Precision, Recall and AP with those of other models. Each model employs the improved YOLOv5 and the original YOLOv5. The datasets are trained by different SR reconstruction methods first and different YOLO methods second.

From Table 3, it can be observed that the proposed model obtains a larger value of Precision compared with other models. This means that the proposed model has a better ability to identify safety helmets. It can also be observed that the proposed model obtains the largest values of Recall compared with the other models. This means the proposed model's ability to find all the safety helmets is the best. The AP value of the proposed model compared with SRCNN+YOLOv5 improved by 25.96% and by 11.10% compared with SRGAN+YOLOv5. Furthermore, when both use the Dual-Channel Residual SR reconstruction module, the AP value of the improved YOLOv5 is approximately 0.64% higher than that of YOLOv5. It is obvious that the LSandGlass module can realize better detection results than the res module. Moreover, the original YOLOv5 is more affected by image resolution. The proposed SR reconstruction module with improved YOLOv5 improved by 23.07% compared with SRCNN with improved YOLOv5 and by 9.25% compared with SRGAN with improved YOLOv5. The proposed SR reconstruction module with the original YOLOv5 improved by 25.16% compared with SRCNN with the original YOLOv5 and by 10.39% compared with SRGAN with the original YOLOv5. These results show that the improved YOLOv5 has better robustness in detection tasks.

Table 3. Comparison with other models in safety helmet detection.

Model / Metrics	Precision (%)	Recall (%)	AP (%)
SRCNN [19]+YOLOv5	73.2	55.1	62.8
SRCNN+improved YOLOv5	74.1	58.3	64.3
SRGAN [35]+YOLOv5	81.3	61.1	71.2
SRGAN+improved YOLOv5	84.3	60.3	72.4
Dual-Channel Residual SR reconstruction module+YOLOv5	88.4	71.5	78.6
Dual-Channel Residual SR reconstruction module+improved YOLOv5	88.6	71.5	79.1

As shown in Figure 10, when the features of safety helmets are obvious in the image, the proposed model has very good recognition of the safety helmets. Comparing (b) with (d) and (f), these results are obtained by the same detection method and different SR reconstruction methods. The number of valid detection boxes in (b) is much greater than that in (d) and (f). The Precision in (b) is also larger than those in (d) and (f). These results mean that the proposed SR reconstruction method obtains better performance than SRCNN and SRGAN. Comparing (a) with (b), (c) and (d), (e) and (f), the valid detection boxes in (b), (d) and (f) are more than those in (a), (c) and (e). The above results show that the improved YOLOv5 achieves a higher recognition ratio of safety helmets; furthermore, the proposed method can obtain more accurate positioning and higher recognition precision for safety helmet detection. This means that the improved YOLOv5 is superior to the original YOLOv5. The above results indicate that the proposed model can achieve better detection results than other models.

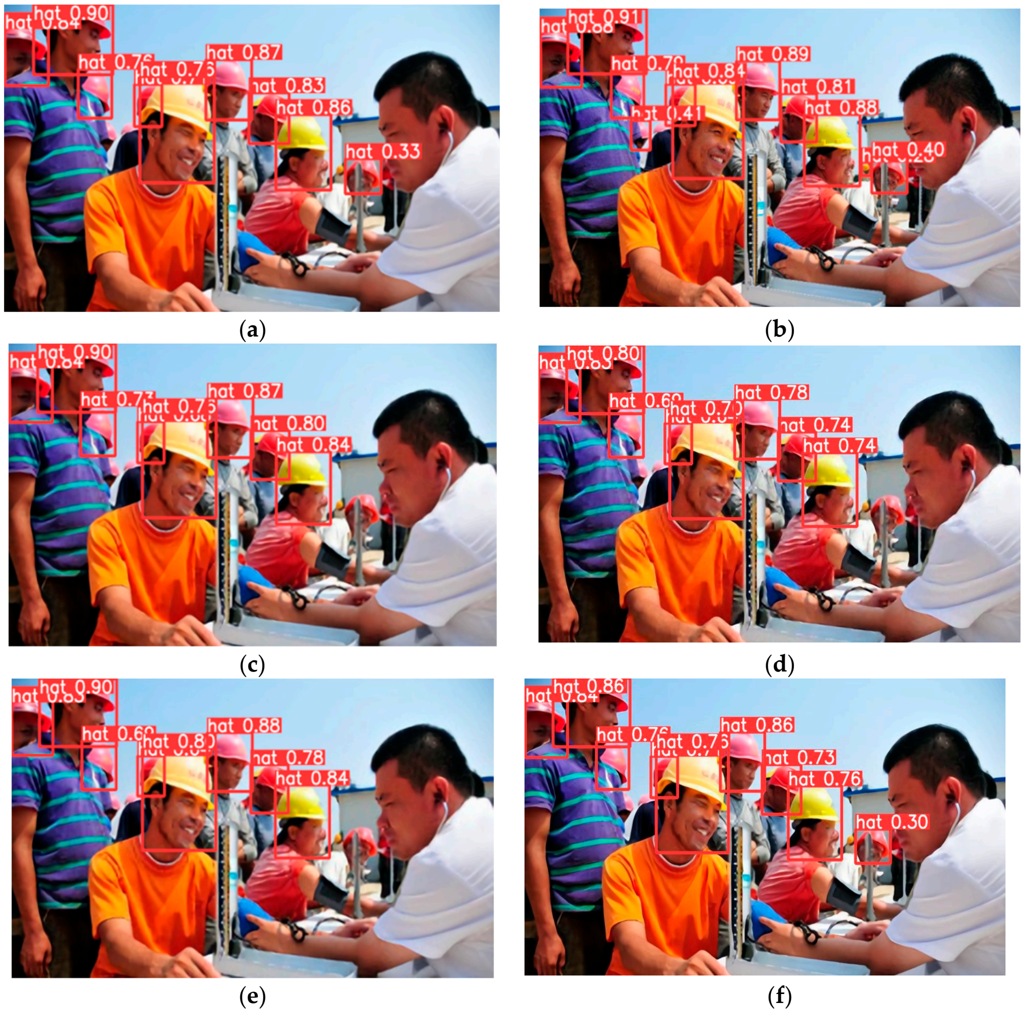

Figure 10. Target detection results obtained using different methods. (**a**) The result obtained using the proposed SR resolution module with the original YOLOv5 network; (**b**) the result obtained using the proposed SR resolution module with the improved YOLOv5 network; (**c**) the result obtained using SRGAN with the original YOLOv5 network; (**d**) the result obtained using SRGAN with the improved YOLOv5 network; (**e**) the result obtained using SRCNN with the original YOLOv5 network; and (**f**) the result obtained using SRCNN with the improved YOLOv5 network.

5. Discussion and Conclusions

A novel safety helmet detection model is presented to implement super-resolution reconstruction-driven safety helmet detection. At construction sites, the images collected need to be transmitted to the terminal for detection. The resolution of images is lowered to make it faster. This can lead to a reduction in the detection accuracy. A novel detection model is proposed to overcome this problem. It consists of two modules. First, the SR reconstruction module is used to improve the image quality. Then, to finish the helmet detection, a novel YOLOv5 module is used as the detection module. They are trained separately but tested by the proposed datasets together. The experimental results show that the proposed SR module can increase the PSNR value while maintaining a consistent SSIM value compared with some existing SR reconstruction methods. It demonstrates the superiority of the proposed model. Based on the current results, the proposed model is a feasible tool for safety helmet detection. It can be easily used in construction monitoring or traffic safety monitoring. This paper mainly uses the individual models on specific tasks and combines the models to achieve the whole task. In the future, we will continue to realize the integrated design of SR reconstruction and YOLOv5 to reduce design redundancy. At the same time, we will implement a lightweight model and improve the computational effectiveness. Besides that, we will consider the noise in images coming from industrial sites in the future research.

Author Contributions: Conceptualization, J.H., Y.L. (Yicheng Liu) and Y.L. (Yan Liu); methodology, J.H. and Y.L. (Yicheng Liu); software, Z.L.; validation, Y.L. (Yicheng Liu), Z.L. and Y.L. (Yan Liu); data curation, B.Z.; writing—original draft preparation, J.H. and Z.L.; writing—review and editing, Y.L.; visualization, B.Z.; supervision, J.H.; project administration, B.Z. and J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the research on automatic and intelligent safety management technology at construction sites, grant number CSCEC-2020Z-10 and the research on intelligent construction site management based on Internet of Things and image recognition technology, grant number KJYF-2019-12.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kurien, M.; Kim, M.K.; Kopsida, M.; Brilakis, I. Real-time simulation of construction workers using combined human body and hand tracking for robotic construction worker system. *Autom. Constr.* **2018**, *86*, 125–137. [CrossRef]
2. Hao, Z.; Wei, Y. 448 cases of construction standard statistical characteristic analysis of inductrial injury accident. *Stand. China* **2017**, *2*, 245–247.
3. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
4. Zhao, H.; Liu, J.; Chen, H.; Chen, J.; Li, Y.; Xu, J.; Deng, W. Intelligent diagnosis using continuous wavelet transform and gauss convolutional deep belief network. *IEEE Trans. Reliab.* **2022**, 1–11. [CrossRef]
5. Huang, C.; Zhou, X.; Ran, X.; Liu, Y.; Deng, W.; Deng, W. Co-evolutionary competitive swarm optimizer with three-phase for large-scale complex optimization problem. *Inf. Sci.* **2023**, *619*, 2–18. [CrossRef]
6. Luo, X.; Wang, G.; Song, T.; Zhang, J.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; Zhang, S. MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Med. Image Anal.* **2021**, *72*, 102102. [CrossRef]
7. Yan, W.; Liu, Y.; Lan, Q.; Zhang, T.; Tu, H. Trajectory planning and low-chattering fixed-time nonsingular terminal sliding mode control for a dual-arm free-floating space robot. *Robotica* **2022**, *40*, 625–645. [CrossRef]
8. Yao, J.; Yan, W.; Lan, Q.; Liu, Y.; Zhao, Y. Parameter optimization of dsRNA splicing evolutionary algorithm based fixed-time obstacle-avoidance trajectory planning for space robot. *Appl. Sci.* **2021**, *11*, 8839. [CrossRef]
9. Wu, D.; Luo, X.; Wang, G.; Shang, M.; Yuan, Y.; Yan, H. A highly accurate framework for self-labeled semisupervised classification in industrial applications. *IEEE Trans. Ind. Inform.* **2018**, *14*, 909–920. [CrossRef]
10. Wu, D.; Luo, X.; He, Y.; Zhou, M. A prediction-sampling-based multilayer-structured latent factor model for accurate representation to high-dimensional and sparse data. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–14. [CrossRef]
11. Zhao, H.; Zhang, P.; Zhang, R.; Yao, R.; Deng, W. A novel performance trend prediction approach using ENBLS with GWO. *Meas. Sci. Technol.* **2023**, *34*, 025018. [CrossRef]
12. Deng, W.; Zhang, L.; Zhou, X.; Zhou, Y.; Sun, Y.; Zhu, W.; Chen, H.; Deng, W.; Chen, H.; Zhao, H. Multi-strategy particle swarm and ant colony hybrid optimization for airport taxiway planning problem. *Inf. Sci.* **2022**, *612*, 576–593. [CrossRef]
13. Deng, W.; Liu, H.; Xu, J.; Zhao, H.; Song, Y. An improved quantum-inspired differential evolution algorithm for deep belief network. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 7319–7327. [CrossRef]
14. Song, Y.; Cai, X.; Zhou, X.; Zhang, B.; Chen, H.; Li, Y.; Deng, W.; Deng, W. Dynamic hybrid mechanism-based differential evolution algorithm and its application. *Expert Syst. Appl.* **2023**, *213*, 118834. [CrossRef]
15. Luo, X.; Chen, J.; Song, T.; Wang, G. Semi-supervised medical image segmentation through dual-task consistency. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 8801–8809. [CrossRef]
16. Luo, X.; Wang, G.; Liao, W.; Chen, J.; Song, T.; Chen, Y.; Zhang, S.; Metaxas, D.N.; Zhang, S. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Med. Image Anal.* **2022**, *80*, 102517. [CrossRef]
17. Li, J.; Qin, H.; Wang, J.; Li, J. OpenStreetMap-based autonomous navigation for the four wheel-legged robot via 3D-lidar and CCD camera. *IEEE Trans. Ind. Electron.* **2021**, *69*, 2708–2717. [CrossRef]
18. Hale, A.R.; Heming, B.H.J.; Carthey, J.; Kirwan, B. Modelling of safety management systems. *Saf. Sci.* **1997**, *26*, 121–140. [CrossRef]
19. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 184–199.
20. Li, J.; Fang, F.; Mei, K.; Zhang, G. Multi-scale residual network for image super-resolution. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 527–542.

21. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
22. Long, Z.; Peng, Z. Based on dual channel residual network image super-resolution algorithm. *J. Xi'an Jiaotong Univ.* **2021**, *1*, 1–8.
23. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
24. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
25. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef] [PubMed]
26. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
27. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
28. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
29. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
30. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
31. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
32. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
33. Jocher, G.; Nishimura, K.; Mineeva, T.; Vilariño, R. Yolov5. [EB/OL]. Available online: https://github.com/ultralyc-s/yolov5 (accessed on 9 August 2020).
34. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
35. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
36. Lu, Y.; Zhou, Y.; Jiang, Z.; Guo, X.; Yang, Z. Channel attention and multi-level features fusion for single image super-resolution. In Proceedings of the 2018 IEEE Visual Communications and Image Processing (VCIP), Taichung, Taiwan, 9–12 December 2018; pp. 1–4.
37. Liu, Y.; Wang, Y.; Li, N.; Cheng, X.; Zhang, Y.; Huang, Y.; Lu, G. An attention-based approach for single image super resolution. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2777–2784.
38. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 294–310.
39. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.