Highly Differentiated Target Detection Method Based on YOLOv3 Model Under Extremely Low Light Conditions

Chenxu Yan

Jiangsu University of Science and Technology

Haijian Shao (✉ jsj_shj@just.edu.cn)

Jiangsu University of Science and Technology <https://orcid.org/0000-0001-5360-0788>

Xing Deng

Jiangsu University of Science and Technology

Research Article

Keywords: Target detection, Low light, Wavelet transformation, Highly differentiated features, YOLOv3

Posted Date: January 20th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1212268/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Highly differentiated target detection method based on YOLOv3 model under extremely low light conditions

Chenxu Yan¹, Haijian Shao^{1,*}, and Xing Deng^{1,2}

¹ School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China

² School of Automation, Key Laboratory of Measurement and Control for CSE, Ministry of Education, Southeast University, Jiangsu, 210096, Nanjing, China

Abstract. Because of the dark atmosphere created by YOLO v3's poor detection accuracy, the photographs obtained in reality usually considerably improve the difficulty of target detection tasks in extremely low light settings. Under extremely low light conditions, the YOLOv3 model is used to suggest a high-discrimination target detection approach. Target detection is an essential task in the field of computer vision, in which the image is the source of input, but the image obtained in reality may make the target detection task in extremely low light conditions because of the dark environment and poor quality imaging equipment, which increases the difficulty of target detection and affects the accuracy of detection. Therefore, based on the above reasons, this paper proposes a high-discrimination target detection method based on the YOLOv3 model under extremely low light conditions. The low light enhancement algorithm based on dehazing algorithm is improved for restraining the part of the transmission rate less than 0.5 in the pixel, to enhance the contrast of the processed images. The improved low light enhancement algorithm is 5.32% higher based on the evaluation index of target detection than the original images. Coiflet wavelet transform is applied to achieve high-discrimination feature extraction of images and decompose the low-frequency features of image information at different resolutions to obtain high-frequency features in the horizontal, vertical, and diagonal directions. The eventually constructed YOLOv3 network has reduced the negative impact of the detected target on the detection accuracy under the illumination change. The results reveal that the target detection method established in this paper has been enhanced by 40.6% on low light photos when compared to the traditional YOLOv3 model, and it still has considerable advantages in comparison to other benchmark approaches in target detection models.

Keywords: Target detection · Low light · Wavelet transformation · Highly differentiated features · YOLOv3

* Corresponding author: jsj_shj@just.edu.cn

1 Introduction

Target detection is a method for detecting and locating multiple distinct targets in images. It's widely used in image retrieval [1], image classification [2, 3], semantic segmentation [4, 5], visual tracking [6], super pixels [7], robot navigation [8], and other fields. Extracting significant things from complex backgrounds with a single characteristic can be difficult since natural images feature a variety of intricate backgrounds. Traditional target detection approaches, such as Hog feature + SVM algorithm [9], Haar feature + Adaboost [10, 11], and DPM algorithm [12, 13], are mostly dependent on human feature annotation methods [11, 12]. These target identification strategies manually extract features, but because the extraction procedure is difficult and portability is weak, the algorithm needs to be changed to match the goal [14]. In a growing number of fields, convolutional neural networks have lately surpassed classic pattern recognition and machine learning methods in terms of performance and accuracy [15], the second-order detection model and the first-order detection model appear one after another.

The detection accuracy is considerably improved by using a second-order model. Girshick [16], for example, introduced the R-CNN model's target identification method in 2014 and used the sliding window concept. The PASCAL VOC target detection test [17] revealed that mAP could be boosted by about 20% above standard methods, to 62.4 percent. The revised Faster R-CNN [18] improves test speed hundreds of times while also boosting accuracy.

The detection time has been considerably reduced due to first-order detection models. For example, Facebook's artificial intelligence lab's YOLO [19] target detection model includes only one training network, and the test speed on a single Titan X approaches 45 frames per second. Faster R-CNN is nine times faster than this. Following the YOLO and YOLOv2 [15] target detection network, the YOLOv3 [20] target detection network proposed by Joseph Redmon and Ali Farhadi can complete the detection of 320x320 size images in 22 milliseconds, and in the Titan X environment, the detection accuracy reaches 57.9%.

Insufficient illumination will substantially impair the visual quality of the image, resulting in loss of image details and low image contrast, which will damage not only people's subjective feelings, but also the performance of various target identification models. This is because most models are trained based on normal image data sets and cannot adapt to extremely low light conditions, naturally, it also affects the accuracy of target detection. As a result, whether processing an image captured in a low light setting or with a low-end equipment, an effective low light image enhancement solution is necessary to restore as much detailed information as possible and improve image quality.

Researchers have previously investigated low light image enhancement and proposed numerous techniques. For example, M. Tanaka et al. used a gradient-based low light image enhancement algorithm, combined with intensity-range constraints, to effectively enhance the low light image [21]. S. A. Priyanka et al. used a principal component analysis framework to enhance low light images with decomposed luminance-chrominance components [22]. And X. Guo et al. found the maximum value in the R, G, and B channels of the low light image

to estimate the illumination of each pixel individually. They refined the initial lighting by applying a structure on the initial illumination map as the final illumination map, realize the corresponding image enhancement [23].

Target detection for low light photos has also seen a lot more success. Kniaz et al., for example, built a novel convolutional neural network for detecting encoded targets automatically [24]. T. Guo et al. used an improved MSRCP-based image enhancement algorithm for the question of insufficient underwater light, combined with YOLOv3 for rapid detection of marine organisms [25]. Ziteng Cui et al. proposed a new multi-task automatic coding conversion model for the effect of insufficient photons and noise in the image, which improved the performance of target detection [26].

Xuan Dong et al. developed a low light improvement strategy based on a dehazing algorithm in their study [27]. The program uses a dehazing model to improve an inverted low light image, which has a unique effect. This algorithm's low light enhanced image, on the other hand, lacks an effective manner of processing light transmittance, leading to an uneven distribution of light and dark in the image, poor image contrast, and noticeable blurring of the bright section of the image. The above impacts both subjective perception and the target detection effect. As a result, this paper enhances the image processing technique for extremely low light circumstances, merges it with the YOLOv3 model, and exploits the strong discrimination of the Coiflet wavelet feature extraction to boost detection accuracy.

2 Design of high-discrimination target detection method under extremely low light conditions

Figure 1 shows the design of the high-discrimination target detection method in severe low light conditions, which is made up of four modules. The first part is the design and improvement of a low light enhancement algorithm based on dehazing algorithm, in which a more reasonable and effective restraint method is primarily used for the transmittance, resulting in an enhanced image with a stronger contrast between light and dark, which improves target detection accuracy. Considering that, directly putting the pictures in the data set into the deep learning network for training, will lose most of the low-frequency and high-frequency information. Hence, the second part designs a high-discrimination feature extraction method through the Coiflet wavelet transform. More precise information can be collected from each image in the horizontal, vertical, and diagonal directions, and also can be stored in the third part (YOLOv3 deep learning model) for training. The training process will pay more attention to the detailed features of the image, which is conducive to target detection Increased accuracy. The fourth part evaluates the improved low light enhancement algorithm and target detection model proposed in this paper from both subjective and objective aspects.

The following are the primary contributions of this article in general:

(1) The low light enhancement algorithm based on dehazing algorithm has been modified to confine the picture transmission rate more appropriately, increase image quality, and aid in target identification accuracy.

(2) It is proposed to use the Coiflet wavelet to extract the image's highly differentiated characteristics, allowing the convolutional neural network to pay more attention to its horizontal, vertical, and diagonal details, improving target recognition accuracy.

Fig. 1. Highly differentiated target detection method process under extremely low light conditions

2.1 Low light image enhancement algorithm based on dehazing model

According to paper [27], the pixels in the sky and distant background area always have high intensity in all color (RGB) channels after inverting a low light image. In contrast, the close-up part of the non-sky area has at least one color channel with a pixel value less than 180. As shown in Figure 2, the pixel distribution of the low light image after inversion is strikingly similar to the pixel distribution of the fog image, both of which have high values of more than 80% pixels, and all of these pixels exclusively apply to the perspective component.

As shown in the above images, a low light image with 0-255 inversion is quite similar to the foggy image in the real scene in terms of the naked eye, and the histograms of figures (b) and (c) are very similar. In other words, the two photos have a relatively similar distribution, with the gray value of the pixel with a gray level of roughly 200 being higher.

The inversion picture of the low light image looks visually similar to the fog images, so the low light image can be created by applying the atmospheric dehazing model to the inversion image of the low light image and then inverting the processed image. Make the image more appealing. Paper [27] describes a low light picture improvement algorithm based on a dehazing algorithm, similar to the scenario of images acquired in foggy weather situations. The algorithm first inverts the input low light image, and then uses the atmospheric dehazing model to dehaze the inverted image.

Fig. 2. Low light inverted image and foggy image comparison

First, input a low light picture, and then invert the three channels of this picture from 0-255:

$$R^c(x) = 255 - I^c(x) \quad (1)$$

The c in the above formula represents the three channels of the image in RGB, $I^c(x)$ represents the pixel value of the input image on each channel (I is the set of all pixel values, x is the single-pixel value), $R^c(x)$ represents the pixel of the output image on each channel Value (R is the set of all pixel values, x is the value of a single-pixel).

Then perform dehazing according to the atmospheric dehazing model:

$$R(x) = J(x)t(x) + A(1 - t(x)) \quad (2)$$

where $R(x)$ is the brightness of the input image, $J(x)$ is the image obtained after dehazing, A is the brightness of the original image or scene, that is, the ambient light, $t(x)$ is the transmittance. The existing condition is only $R(x)$, so the ambient light A and the transmittance $t(x)$ must be calculated. For the estimation of ambient light, the method is to traverse all of the pixels in the image, sort them in descending order according to the minimum value of each pixel on the RGB three channels, then select the first 100 pixels, and finally select the ambient light with the largest sum of the three channels among the 100 pixels. The estimation of $t(x)$ is as follows:

$$t(x) = 1 - w \min_{c \in \{r, g, b\}} \left(\min \left(\frac{R^c(y)}{A^c} \right) \right) \quad (3)$$

Obtained by equation conversion:

$$J(x) = \frac{R(x) - A}{t(x)} + A \quad (4)$$

However, the enhancing impact for low light images is relatively poor when employing this approach directly. Therefore, it is considered that the region of interest can be enhanced without effecting the region of non-interest. Therefore, a constraint term $P(x)$ is introduced:

$$P(x) = \begin{cases} 2t(x), & 0 < t(x) < 0.5, \\ 1, & 0.5 < t(x) < 1 \end{cases} \quad (5)$$

Thus the previous enhanced equation becomes:

$$J(x) = \frac{R(x) - A}{t(x)P(x)} + A \quad (6)$$

This formula means that when $t(x)$ is less than 0.5, the corresponding pixel needs to be enhanced, so assign a small value to $t(x)$ to make $t(x)P(x)$ smaller to increase the RGB intensity of the pixel. On the other hand, when $t(x)$ is greater than 0.5, the original value is retained to avoid excessively increasing the corresponding pixel intensity.

In the case of $0 < t(x) < 0.5$, to make the dark place darker, paper [28] added the constraint term $P(x)$, and reduced $t(x)$ to $2t^2(x)$, reducing the transmittance in the range of 0-0.5, to make the transmittance $t(x)$ in the range of 0-0.5 lower. However, it simply updates the original data to a square value and uses ‘2’ as the coefficient to make the constrained transmittance closer to the original transmittance, to make the transmittance $t(x)$ in the range of 0-0.5 lower. This method cannot make a small value converge smaller, that is, as $t(x)$ decreases in the same range, $2t^2(x)$ does not converge more in each same range where $t(x)$ decreases. Namely, it is not really realized to make the smaller value in the range of 0-0.5 smaller.

Thus, the convergence method of $t(x)$ is changed from $2t^2(x)$ to $\ln(t(x)+1)/2$. The derivative of $\ln(t(x)+1)/2$ is nonlinear, so shows the same reduction range of $t(x)$, the value of $\ln(t(x)+1)/2$ actually has a greater reduction. In order to reflect the effect of the improvement, table 1 demonstrates the changes before and after the improvement.

Table 1. Value changes before and after improvements

$t(x)$	$2t^2(x)$	$\ln(t(x)+1)/2$	$2(t(x)^2 - (t(x) - 0.1)^2)$	$(\ln(t(x)+1) - \ln(t(x)+0.9))/2$
0.5	0.500	0.203	0.180	0.035
0.4	0.320	0.168	0.140	0.037
0.3	0.180	0.131	0.100	0.040
0.2	0.080	0.091	0.060	0.044
0.1	0.020	0.048	0.020	0.048

It can be seen from the above table that as $t(x)$ approaches from 0.5 to 0.1, the difference of $2t^2(x)$ gradually decreases during the process of decreasing from 0.500 to 0.020, and the drop rate is 0.040. However, the improved $\ln(t(x) + 1)/2$ shows that the difference gradually increases during the decrease from 0.203 to 0.048 within the same decrease of $t(x)$, and the magnitude of each decrease increases. In addition, it can be seen that if $t(x)$ is close to 0.5, the effect of the coefficient $P(x)$ constraint will no longer be obvious. The more prominent ones in the figure are the two values of 0.4 and 0.5. After $2t^2(x)$ constraints, they are 0.50 and 0.32 respectively, which is not much different from the original value. The improved constraint not only ensures that $t(x)$ less than 0.5 is effectively constrained, but also ensures that the smaller the value, the greater the magnitude of the constraint, which makes the final image more layered and closer to reality.

In fact, the characteristics of $2t^2(x)$ and $\ln(t(x) + 1)/2$ are not only reflected in these special values, the five values in the table are just to illustrate the different constraints of the two strategies. On the image of the function $\ln(t(x) + 1)/2$, the characteristic of this decline is continuous, so all changes in $t(x)$ within the range of 0-0.5 can be constrained. The visualization effect is shown in the figure 3 shown.

The left side of the dotted line in the above figure represents the processing method of $t(x)$ in paper[27], and the right side of the dotted line represents the processing method of $t(x)$ in this article. The blue lines in figures (a) and (b) describe the functions $2t^2(x)$ and $\ln(t(x) + 1)/2$, respectively, and the yellow lines represent their derivatives, respectively. The images of the two functions, as well as their increasing and decreasing trends show that, while the original $2t^2(x)$ in the independent variable diminishes while the dependent variable shrinks as well, the amount of the reduction remains the same. This is due to the fact that its derivative is linearly related, i.e., decreasing $t(x)$ in the same continuous interval does not ensure that the value obtained after applying the $2t^2(x)$ constraint each time will converge more in the same interval than the previous same interval value. Instead, it goes in the opposite direction: as $t(x)$ decreases, the $2t^2(x)$ decreases linearly, which does not play a better constraining effect on smaller $t(x)$ values. Observing the $\ln(t(x) + 1)/2$ function itself and its derivative image again, it can be seen that if $t(x)$ gradually becomes smaller from 0.5, $\ln(t(x) + 1)/2$ still keeps gradually decreasing, and the magnitude of each decrease is larger. This constraint strategy is exactly the opposite of the original $2t^2(x)$, which ensures that when $t(x)$ becomes smaller, the value after the constraint can become smaller.

Figures (c) and (d) respectively show the effects of low light image processing in paper[27] and the method in this paper, and the corresponding histograms. The image originally constrained by $2t^2(x)$ performs worse in the contrast of light. Especially in bright places, it is easy to integrate with the surrounding brightness, the contrast is not obvious, and the clarity is worse. However, the image after changing to $\ln(t(x) + 1)/2$ presents a better sensory effect, which

Fig. 3. (a): The $t(x)$ function before the improvement and its derivative (b): Improved $t(x)$ function and its derivative (c): Pre-improved low light image enhancement (d): Improved low light image enhancement

makes the contrast effect of the whole image more prominent. Moreover, the information displayed in the histogram is more abundant.

After the modification, the entire process can be described as follows:

- (1) Enter a low light image;
- (2) Reverse the three channels of the low light image in the range of 0-255 respectively;
- (3) Use atmospheric dehazing model for dehazing;
- (4) Judge the transmittance $t(x)$ obtained in (3), if $0 < t(x) < 0.5$, $t(x) = \ln(t(x) + 1) / 2$, otherwise $t(x)$ keep the original value;
- (5) Reverse the processed image to get a low light enhanced image.

2.2 Design of high-discrimination feature extraction method

The Coiflet wavelet is a biorthogonal wavelet with a lot of support $6N-1$, which is close to symmetry, is the support range. The wavelet function has a vanishing moment of $2N$, while the scale function has a vanishing moment of $2N-1$. With a higher compression ratio, the high-frequency coefficient will be smaller, the filter flatter, and the image energy after wavelet decomposition will be more concentrated, so that as many wavelet coefficients as possible are zero or as few non-zero wavelet coefficients as possible, which is conducive to data compression and noise elimination. This is also known as the vanishing moment's magnitude determining the image's vibration level following decomposition. Longer support lengths necessitate more calculation time and yield more high-amplitude wavelet coefficients, which is the polar opposite of the vanishing moment. If the support length is too long, boundary difficulties will arise; if the support length is too short, the vanishing moment will be too low, preventing signal energy concentration. As a result, the vanishing moment and the support length are two mutually exclusive concepts. This is better reflected in the Coiflet wavelet, which has a nice vanishing moment and tight support in both the frequency and time domains. Furthermore, the Coiflet wavelet outperforms the general gaussian function in terms of orthogonality and biorthogonality, as well as spectrum usage rate. The two vector spaces V_j and W_j formed by the scale function of the Coiflet wavelet and the wavelet function are as follows:

$$V_j = s^{i/2} \phi(s^i x - k) | j, k \in Z \quad (7)$$

$$W_j = S^{i/2} \psi(s^i x - k) | j, k \in Z \quad (8)$$

In 2D space, scale-space $V_j(x1, x2)$ and wavelet space $W_j(x1, x2)$ have the following relationships:

$$V_{j-1}(x1, x2) = V_j(x1, x2) \oplus W_j(x1, x2) \quad (9)$$

$$W_j(x1, x2) = V_{j-1}(x1, x2) / V_j(x1, x2) \quad (10)$$

The various subspaces of the scale-space V_j have a nested relationship, and the role of the wavelet space W_j is the information between the wedding scale

subspaces V_{j-1} and V_j , it captures the information lost when V_{j-1} approaches V_j , that is to say, the vector space V_j and the vector space W_j are orthogonal, and the vector space W_j can represent the part that cannot be expressed in the vector space V_j , then the information can be expressed completely.

For any function $f(x_1, x_2)$ in the scale-space, according to equation (9), we have:

$$P_{j-1}f(x_1, x_2) = P_jf(x_1, x_2) + D_jf(x_1, x_2) \quad (11)$$

Since $P_{j-1}f(x_1, x_2)$ represents the projection of function $f(x_1, x_2)$ on space $V_{j-1}(x_1, x_2)$, so $P_{j-1}f(x_1, x_2)$ can be represented by the components $V_j(x_1, x_2)$ and $W_j(x_1, x_2)$ in $V_{j-1}(x_1, x_2)$. Since two-dimensional space can be decomposed into one-dimensional space, $V_{j-1}(x_1, x_2)$ can be decomposed into the following formula:

$$\begin{aligned} V_{j-1}(x_1, x_2) &= V_{j-1}(x_1) \otimes V_{j-1}(x_2) \\ &= [V_j(x_1) \oplus W_j(x_1)] \otimes [V_j(x_2) \oplus W_j(x_2)] \\ &= [V_j(x_1) \otimes V_j(x_2)] \oplus [V_j(x_1) \otimes W_j(x_2)] \oplus [W_j(x_1) \otimes V_j(x_2)] \oplus [W_j(x_1) \otimes W_j(x_2)] \end{aligned} \quad (12)$$

According to the correspondence between equation (9) and the upper formula:

$$V_j(x_1, x_2) = V_j(x_1) \otimes V_j(x_2) \quad (13)$$

The orthogonal normalization basis of the above formula is:

$$\phi_{ik1}(x_1)\phi_{ik2}(x_2) = s^{i/2}\phi(s^i x_1 - k_1)s^{i/2}\phi(s^i x_2 - k_2) \quad (14)$$

The $\phi_{ik1}(x_1)$ and $\phi_{ik2}(x_2)$ in the above formula are low-pass scale functions, so the space $V_j(x_1, x_2)$ expressed by this formula represents the low-frequency characteristics of the original space.

Another component used to express $P_{j-1}f(x_1, x_2)$ is $W_j(x_1, x_2)$. According to the corresponding relationship between equation (9) and equation (12), $W_j(x_1, x_2)$ is expressed as:

$$W_j(x_1, x_2) = [V_j(x_1) \otimes W_j(x_2)] \oplus [W_j(x_1) \otimes V_j(x_2)] \oplus [W_j(x_1) \otimes W_j(x_2)] \quad (15)$$

The upper equation can be divided into three parts, each part of the orthogonal base is:

$$\begin{aligned} \phi_{ik1}(x_1)\psi_{ik2}(x_2) &= s^{i/2}\phi(s^i x_1 - k_1)s^{i/2}\psi(s^i x_2 - k_2) \\ \psi_{ik1}(x_1)\phi_{ik2}(x_2) &= s^{i/2}\psi(s^i x_1 - k_1)s^{i/2}\phi(s^i x_2 - k_2) \\ \psi_{ik1}(x_1)\psi_{ik2}(x_2) &= s^{i/2}\psi(s^i x_1 - k_1)s^{i/2}\psi(s^i x_2 - k_2) \end{aligned} \quad (16)$$

These three parts represent high-frequency characteristics in horizontal, vertical, and diagonal directions.

Therefore, $P_{j-1}f(x_1, x_2)$ is represented by components in $V_j(x_1, x_2)$ and $W_j(x_1, x_2)$ as:

$$\begin{aligned} P_{j-1}f(x_1, x_2) &= P_jf(x_1, x_2) + D_jf(x_1, x_2) \\ &= \sum_{k_1 k_2} x_{k_1 k_2}^{(j)} \phi_{jk1}(x_1)\phi_{jk2}(x_2) + \sum_{k_1 k_2} \alpha_{k_1 k_2}^{(j)} \phi_{jk1}(x_1)\psi_{jk2}(x_2) \\ &+ \sum_{k_1 k_2} \beta_{k_1 k_2}^{(j)} \psi_{jk1}(x_1)\phi_{jk2}(x_2) + \sum_{k_1 k_2} \gamma_{k_1 k_2}^{(j)} \psi_{jk1}(x_1)\psi_{jk2}(x_2) \end{aligned} \quad (17)$$

As a result, the function $f(x_1, x_2)$ is extracted in three directions using low-frequency and high-frequency characteristics. In the two-dimensional image of target detection, one sample can be decomposed into four samples, which are used to represent the approximate characteristics of the original image as well as the high frequency characteristics in the horizontal, vertical, and diagonal directions, as shown in figure 4 below:

Fig. 4. Highly differentiated feature extraction method effect

2.3 Deep learning model analysis and structure design

The target detection method suggested in this paper for extremely low light situations is based on the YOLOv3 model and leverages Darknet-53 as the backbone network. For the core network, Darknet-53 has 52 convolutional layers and 1 fully linked layer. A 32-channel convolutional layer and 5 repeating residual units comprise the 52 convolutional layers. Each of the 5 residual units has an individual convolutional layer and a set of repeated convolutional layers, with repetitions of 1, 2, 8, 8, 4, and a 1*1 convolution (half the number of channels) and a 3*3 convolution in the repeated convolutional layer (recover the number of channels).

The input size of the image is 416*416, since the step size of the first independent convolutional layer of each residual unit is 2, the three feature layers of its output have gone through 8 times, 16 times and 32 times respectively,

and a total of three times downsampling. The size of the corresponding image input is changed from 416*416 to 52*52, 26*26 and 13*13, the output feature of each size has 3 priori boxes, and each priori box contains 25 parameters, these 25 parameters cover the length, width, abscissa, ordinate, confidence parameter and category of the detected object (there are 20 types of target objects in the PASCAL VOC data set), so the number of predictions reached 3247 ($13*13*3 + 26*26*3 + 52*52*3$). The output feature map is shown in figure 5 below, the x, y, w, h of the detection frame can be obtained by decoding the a priori frame, the calculation method is as follows:

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \\ b_w &= pwe^{t_w} \\ b_h &= phe^{t_h} \end{aligned} \quad (18)$$

Among them, $\sigma(t_x)$ and $\sigma(t_y)$ are the offsets based on the grid point coordinates of the upper left corner of the center point of the rectangular box, σ is the sigmoid activation function, pw and ph are the width and height of the a priori box, and b_w and b_h are the width and height of the actual detection box.

The 3 feature maps of different scales generated by each input image are respectively predicted using 9 scales of a priori boxes, which ensures a good detection effect for different sizes of receptive fields. Taking the VOC data set as an example, the 9 priori boxes are (10x13), (16x30), (33x23), (30x61), (62x45), (59x119), (116x90), (156x198), (373x326), since the smaller the feature map, the larger the receptive field, so we adhere to the principle of using a larger a priori box on the small feature map, and use the smaller priori box (10x13), (16x30), (33x23) on the 52*52 feature map, use the medium priori box (30x61), (62x45), (59x119) on the 26*26 feature map, and use the larger priori box (116x90), (156x198), (373x326) on the 13*13 feature map, as shown in table 2 below. The overall structure of the deep learning model in this article is shown in Figure 6. The YOLOv3 target detection model is the main body of the convolutional neural network, which also incorporates the improved low light enhancement algorithm based on the dehazing algorithm and the high-discrimination feature extraction method proposed in this paper.

Table 2. Comparison table of characteristic graph and prior box

The feature chart scale	52*52	26*26	13*13
The feature chart scale	Small	Medium	Max
	(10x13)	(30x61)	(116x90)
Priori box	(16x30)	(62x45)	(156x198)
	(33x23)	(59x119)	(373x326)

Fig. 5. YOLOv3 output feature diagram

2.4 Performance evaluation

The improved low light enhancement algorithm in this paper and other low light enhancement algorithms have been objectively evaluated for performance on the five indicators of information entropy (E), peak signal-to-noise ratio (PSNR), spectral angle (SAM), mean square error (RMSE), average gradient (G), and the calculation methods of the five evaluation indicators are as follows:

$$E = - \sum_{i=0}^{255} \sum_{j=0}^{255} p_{ij} \log_2 p_{ij} \quad (19)$$

$P_{i,j}$ represents the proportion of pixels in the image with grayscale values and neighborhood grayscale means in the range of 0-255.

$$PSNR = 10 \times \log \left(\frac{255^2}{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \frac{(f(i,j) - g(i,j))^2}{MN}} \right) \quad (20)$$

M and N represent the rows and columns of the image, respectively, and $f(i,j)$ and $g(i,j)$ represent the pixel values of the i-th row and j-th column

Fig. 6. The overall structure of the deep learning model

before and after the image enhancement, respectively.

$$SAM = \cos^{-1} \frac{d^T x}{(d^T d)^{1/2} (x^T x)^{1/2}} \quad (21)$$

Where d is the two-dimensional matrix of the image before enhancement, and x is the two-dimensional matrix of the image after enhancement.

$$RMSE = \sqrt{\frac{1}{M \times N} \sum_i^{M \times N} (y_i - \hat{y}_i)^2} \quad (22)$$

M and N represent the rows and columns of the image, respectively, and y_i and $\hat{y}_i$ represent the pixel values of the i -th row and j -th column before and after the image enhancement, respectively.

$$G = \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N \sqrt{\frac{(\frac{\partial f}{\partial x})^2 + (\frac{\partial f}{\partial y})^2}{2}} \quad (23)$$

M and N represent the rows and columns of the image, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ respectively represent the gradient of the image in the horizontal and vertical directions.

The target detection model's advantages and disadvantages must be tested on a large number of test sets. Neural network models' general assessment criteria include speed and accuracy, as well as low memory use. The main criterion in the target detection model is mAP (Mean Average Accuracy, used to evaluate accuracy). mAP calculates the average value of each type of target's detection accuracy, which requires calculating the Recall and Precision values for each type of detection target.

Recall value calculation formula:

$$Recall = TP / (TP + FP) \quad (24)$$

Precision value calculation formula:

$$Precision = TP / (TP + FN) \quad (25)$$

T represents correct detection, F represents error detection, P represents positive sample, N represents negative sample, TP represents positive sample detected correctly, FP represents positive sample detected incorrectly, and FN represents Detect false negative samples in (24) and (25) respectively. In target detection, the IoU threshold is an intersection ratio that measures the degree of overlap between the formal bounding box and the anticipated bounding box. The method for determining if the detection is correct is determined by the IoU threshold. It can be used to see if the target in an image can be spotted properly using the intersection and union ratio.

It is important to traverse the accurate information of the target in the image, that is, traverse the real position specified in the label and actually detecting a picture. Then, using the detector, detect the check box and sort the check box's confidence from high to low. Finally, compare the check box with the highest confidence to the true position, and set it as TP if the Iou value is greater than the set threshold. The rest of the check boxes are FP, and the object is marked as detected.

Based on the Recall and Precision values, a series of Precision-Recall curves can be constructed. The curve has Recall as the abscissa and Precision as the ordinate, with the area under the curve representing the accuracy AP of this type and mAP representing the average accuracy rate of 20 categories of items in the VOC data set.

$$mAP = \left(\sum_{i=1}^c AP_i \right) / C \quad (26)$$

Where C represents the object category of the data set, and AP_i represents the AP of each category.

3 Experiment

3.1 Data set

The data sets used in the experiment are PASCAL VOC2007 and PASCAL VOC2012. The VOC data sets are divided into 4 categories: vehicle, household, animal, and person. There are 20 sub-categories after subdividing, as shown in table 3 below.

The data sets used for training and verification are the trainval data sets of PASCAL VOC 2007 and PASCAL VOC 2012, of which 90% are used for training and 10% are used for verification. The data set has a total of 16,551 images (5011 for VOC2007 and 11540 for VOC2012) and 40058 detection objects (12608 for VOC2007 and 27450 for VOC2012). The test data set is the test data set of VOC2007. This data set There are 4952 pictures and 12032 detection objects in total, as shown in Table 4 below:

Table 3. Distribution of the PASCAL VOC dataset

The main category	Subcategories
vehicle	car,bus,bicycle,motor,plane,boat,train
household	chair,sofa,table,tv,bottle,plant
animal	cat,dog,cow,horse,sheep,bird
person	person

Table 4. Experiment data set

-	Train	Validate	Test
Data set	VOC2007+VOC2012	VOC2007+VOC2012	VOC2007+VOC2012
Pictures	14896	1655	4952
Objects	36052	4006	12032

3.2 Low light enhancement algorithm analysis

In figure 7, the red scatter diagram represents pixels with transmittance less than 0.5 without constraints, the blue scatter diagram represents pixels constrained by $2t^2(x)$ in the original text, and the green scatter plot represents the scatter plot after the constraint is improved to $\ln(t(x) + 1)/2$ (all $t(x)$ values are magnified by 100 times to facilitate observation of the difference, which has no effect on the real effect).

In figure 7, the red scatter diagram represents pixels with transmittance less than 0.5 without constraints, the blue scatter diagram represents pixels constrained by $2t^2(x)$ in the original text, and the green scatter plot represents the scatter plot after the constraint is improved to $\ln(t(x) + 1)/2$ (in order to facilitate the observation of the difference, all $t(x)$ values are magnified by 100 times, which will not affect the real effect). It can be seen that the original constraint method of $2t^2(x)$ does not produce effective constraints on the larger value of 0-0.5, nor does it produce greater constraints on the smaller value, that is, the scattered points at the bottom of $t(x)$ are not significantly increased. Through the improvement of the constraint strategy, the pixels with transmittance less than 0.5 in the image are reduced in a more reasonable way, which not only reduces the transmittance of all pixels in the range of 0-0.5, but also makes the $t(x)$ scattered points become denser and denser from high to low.

This paper compares subjectively with seven low light enhancement algorithms based on Retinex [29] theory, such as SSR, MSR, MSRCR, MSRCR, Gimp, FM, and paper algorithm [27]. From the standpoint of sensory effects, the six algorithms of SSR, MSR, MSRCR, MSRCR, GIMP, and FM have substantial overall distortion, poor contrast, and sharpness, as demonstrated in figure 8. In comparison to existing methods, the improved algorithm in this study has a better overall sensory experience and more coordinated light-dark contrast.

The five indicators of information entropy (E), peak signal-to-noise ratio (PSNR), spectral angle (SAM), mean square error (RMSE), and average gradient (G) are used in Table 5 to perform an objective performance evaluation of the revised algorithm in this study and the algorithm discussed previously. The

Fig. 7. (a): Pixels with transmittance less than 0.5 in the image without constraints. (b): Pixels with transmittance less than 0.5 in the image before $P(x)$ improvement. (c): Pixels with transmittance less than 0.5 in the image after $P(x)$ improvement

image obtained in this paper has a higher information entropy and peak signal-to-noise ratio. A higher information entropy indicates that the image contains more information, while a higher peak signal-to-noise ratio indicates that the image quality obtained by the improved algorithm is better, and implies that the image is less distorted. The algorithm proposed in this paper has a lower spectral angle index, which means that, assuming that the spectrum of each pixel in the image is regarded as a high-dimensional vector, the angle between the processed image and the original image is smaller, the two spectra are more similar, and the likelihood of belonging to similar objects is higher. The mean square error is a pixel-based objective picture quality evaluation metric that compares the fused image to an ideal reference image. This article's picture value is lower, indicating that the smaller the difference between the output and original image, the better the image quality after processing. The average gradient is the average value of the gray change rate, which reflects the rate of change in the contrast of the image's microscopic features, i.e. the rate of density change in the image's multi-dimensional direction, and characterizes the image's relative clarity. Although the average gradient index values of the MSRCR, MSRCP, Gimp, and FM algorithms are higher, meaning that the picture is clearer, this does not imply that they have stronger imaging effects, and they are not in other assessment indicators. There is no superior performance, which means the images produced by these four algorithms do not contain more information and are accompanied by more serious distortion and bigger mistakes when compared to the original image. Table 5 shows that the overall performance of the algorithm proposed in this paper is better on the five indicators, ensuring not only that the processed image maintains better consistency with the original image, but also that the image has good clarity and distortion is less. Most importantly, that the image obtained by this algorithm has greater information entropy, the image contrast is stronger, and the information preset is higher.

Table 5. Objective comparison of several algorithms

-	E	PSNR	SAM	RMSE	G
SSR	7.3265	6.4193	0.8799	121.03	4.5637
MSR	6.4932	7.0757	0.9767	112.91	4.2225
MSRCR	7.5419	6.5554	0.8993	119.88	10.4752
MSRCP	7.5950	6.4766	0.8945	120.97	11.3002
Gimp	7.6698	6.1437	0.8985	125.70	11.9929
FM	7.7125	7.7785	0.8715	105.13	9.1544
Paper[27]	7.7069	7.5162	0.8260	108.07	6.9699
Ours	7.7424	7.7816	0.7790	104.72	7.2279

3.3 Analysis of image high-discrimination feature extraction method

Putting the photographs in the data set directly into the deep learning network for training in the original YOLOv3 will lose the majority of the high-frequency

Fig. 8. Subjective comparison of several algorithms

information. If this data can be entered into a training network, the training process will pay more attention to the specific information in each picture in the horizontal, vertical, and diagonal directions, improving target detection accuracy. As a result, the 16551 images in the training data set are separated into different image information directions. To analyze the image, the original image is decomposed into smooth approximation portions, horizontal, vertical, and oblique parts., as well as a highlight section with a diagonal detail, the scale function and wavelet function of the Coiflet wavelet.

After extracting the low-frequency and high-frequency features from the training data set, the amplitude of the recovered feature picture is visually examined. It can be noticed in figure 9 that an image is wavelet from left to right. The frequency distribution in three directions: horizontal, vertical, and diagonal are shown in Figure 9.

Fig. 9. Frequency distribution after feature extraction

3.4 Performance analysis of target detection methods

The environment used in the experiment is tensorflow-gpu1.14 and keras2.1.5, 100 batches are trained under Nvidia 2080ti graphics card, batch size is set to 8, and the maximum learning rate is 1e-3.

First, the low light improvement technique reduces the negative effects of low light on target detection, and the wavelet decomposition approach for extracting picture features considerably increases the detection effect of the YOLOv3 detection model. The detection frame locates the position of the detected object in the image, and the number on the detection frame represents the correct score of the target object being detected. Figure 10 shows the actual detection comparison between the original YOLOv3 model and the target detection model proposed in this paper. Table 6 shows the AP of our target detection models, YOLOv3, YOLOv4[30], RFBnet[31], Mobilenet-SSD[32], Faster RCNN[28], and

M2det[33], on various types of objects after using our method to enhance the 20 types of low light images in the VOC dataset. We can see that our proposed method has many advantages.

In addition, the mAP of these target detection models is calculated in three situations (low light environment, low light image enhancement using the method given in [27], and low light image enhancement using our method), with the comparison results presented in Table 7. It can be seen that the enhancement approach in paper[28] does not guarantee an increase in target identification accuracy, while the target detection model we suggested does.

Figure 12 shows a polar axis pie chart corresponding to the detection accuracy of each model after employing our enhanced low light enhancement algorithm. The data on each axis in the figure indicates the detection accuracy of the model on the item represented by that axis. The minimum value outward from the center point is 0, and the maximum value is 100, with a 10% interval. The blue coverage area reflects the model's overall accuracy, and it can be observed that this method's detection result is stronger than other models.

Fig. 10. Comparison of target detection effects before and after low light image enhancement

4 Conclusion

Under extremely low light conditions, the proposed high-discrimination target detection method based on the YOLOv3 model compensates for the low light impact on target detection tasks in two ways: an improved low light enhancement algorithm based on the dehazing algorithm and a high-discrimination feature extraction method realized by wavelet transform.

Fig. 11. P-R curve of the test object

Table 6. The accuracy of several models on the detected object

-	Ours	YOLOv3	YOLOv4	RFBnet	Mobilenet-SSD	Faster-RCNN	M2det
plane	82.62	28.42	58.51	69.53	39.01	62.90	70.18
bicycle	85.68	60.75	70.85	79.91	41.79	80.88	79.86
bird	74.01	37.16	42.82	62.46	33.40	52.60	58.63
boat	70.57	19.46	39.52	49.86	28.40	50.67	53.16
bottle	62.17	24.45	32.33	32.38	13.50	29.32	27.00
bus	85.42	49.74	70.00	75.90	42.81	76.51	73.54
car	88.62	62.28	66.43	76.14	40.71	74.60	76.87
cat	84.31	46.54	44.44	71.92	46.76	72.61	64.65
chair	60.68	16.06	27.18	43.91	22.06	41.73	38.50
cow	80.10	18.86	21.62	66.87	33.35	55.74	61.84
table	76.65	42.33	48.89	68.20	44.03	71.90	66.00
dog	80.25	37.70	48.87	67.08	44.20	65.33	61.95
horse	86.64	65.84	62.59	77.63	46.73	78.88	77.53
motor	86.81	47.33	55.62	71.91	42.96	74.58	74.44
person	81.35	47.19	56.32	66.10	35.02	68.23	65.80
plant	46.39	14.80	25.19	39.32	19.39	37.90	38.43
sheep	76.40	1.07	37.15	56.43	31.09	56.12	56.80
sofa	81.80	40.74	33.64	64.42	43.33	70.30	58.75
train	83.89	58.70	67.05	79.62	47.12	70.88	80.70
tv	75.01	40.04	46.02	62.16	34.77	62.62	59.37

Table 7. The detection accuracy of several models of low light images under different enhancement methods

-	Ours	YOLOv3	YOLOv4	RFBnet	Mobilenet-SSD	Faster-RCNN	M2det
Dark	70.96	36.87	54.96	64.06	53.91	65.19	65.09
Paper[27]	72.15	37.97	47.75	64.09	36.52	62.71	62.20
Ours	77.47	39.10	69.83	73.79	63.26	62.75	72.32

Fig. 12. Polar axis pie chart of accuracy comparison between our model and other models on the VOC dataset

Target detection accuracy will be considerably reduced in extremely low light settings. Because the inverted low light image and the foggy image have comparable features, the inverted image is dehazed using the atmospheric dehazing model in the improved low light enhancement algorithm based on dehazing algorithm. The calculation of ambient light and transmittance is the most important part of the dehazing procedure. However, the paper [27] has an unreasonable place in the calculation of the transmittance. When the penalty factor $P(x)$ is used to constrain the transmittance $t(x)$ in the range of 0-0.5, it is simply the transmittance $t(x)$ becomes $2t^2(x)$, which makes the smaller transmittance smaller to a certain extent, but it is not considered that the magnitude of the reduction should also decrease with the decrease of $t(x)$. Therefore, the final imaging contrast of the original algorithm is worse, especially when there is blurring in bright places, which is great for target detection tasks, this has caused an inevitable impact. However, the improved low light enhancement algorithm proposed in this article has a higher image quality. The bright areas in the picture and the surrounding objects will have better contrast, and compared with other algorithms in this paper, the image obtained by the algorithm in this paper has a more comprehensive performance in indicators such as information entropy, peak signal-to-noise ratio, spectral angle, mean square error and average gradient, which provides a guarantee for the accuracy of target detection under extremely low light conditions.

This research suggests the use of the Coiflet wavelet for high-discrimination feature extraction, in addition to applying low light augmentation methods to assist convolutional neural networks in obtaining picture features. In the frequency and time domains, the Coiflet wavelet maintains a good vanishing moment and tight support, and it performs better in terms of orthogonality and biorthogonality. At varying resolutions, the low-frequency information of each picture can be further divided into high-frequency information using the two-dimensional wavelet transform. The convolutional neural network pays more attention to these detailed features during training because of the high-frequency information that harvests details in many directions. So that the resolution of these images with high-frequency details may be improved even more, ensuring that various fuzzy features can be identified and recovered successfully. The improved low light enhancement algorithm based on the dehazing algorithm ensures that the image features can be better extracted by the target detection model, and the high-discrimination feature extraction method extracts the image detail features that play a large role in the target detection task, and this is combined with the first-order target detection model YOLOv3 to ensure that the detection task has higher accuracy. As a result, under extremely low light situations, the high-discrimination target recognition approach based on the YOLOv3 model suggested in this paper has superior environmental adaptability and is more suitable for actual scenes than other target detection models.

Author Contributions Chenxu Yan: Conceptualization, Data curation, Formal analysis, Investigation, Writing-original draft;

Haijian Shao: Conceptualization, Project administration, Resources, Supervision;

Xing Deng: Validation, Visualization, Writing-original draft;

Funding This research was supported by National Sciences Foundation of China grants (No.61902158 No.61806087).

Declarations

Conflict of interest The authors certify that the submission is an original work that is not currently being considered for publication elsewhere. The authors declare that they have no known competing financial interests or personal relationship that could have appeared into influence the work reported in this paper.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study

References

1. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3-d object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing* **21**(9), 4290–4303 (2012)
2. Zhang, K., Guo, Y., Wang, X., Yuan, J., Ma, Z., Zhao, Z.: Channel-wise and feature-points reweights densenet for image classification. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 410–414. IEEE (2019)
3. Benjilali, W., Guicquero, W., Jacques, L., Sicard, G.: Hardware-friendly compressive imaging based on random modulations & permutations for image acquisition and classification. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 2085–2089. IEEE (2019)
4. Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J.: Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4106–4115 (2019)
5. Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G.: Semantic correlation promoted shape-variant context for segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8885–8894 (2019)
6. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. In: International conference on machine learning. pp. 597–606. PMLR (2015)
7. Lin, D., Ji, Y., Lischinski, D., Cohen-Or, D., Huang, H.: Multi-scale context intertwining for semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 603–619 (2018)

8. Craye, C., Filliat, D., Goudou, J.F.: Environment exploration for object-based visual saliency learning. In: 2016 IEEE international conference on robotics and automation (ICRA). pp. 2303–2309. IEEE (2016)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 886–893. Ieee (2005)
10. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: Proceedings. international conference on image processing. vol. 1, pp. I–I. IEEE (2002)
11. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. vol. 1, pp. I–I. Ieee (2001)
12. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE conference on computer vision and pattern recognition. pp. 1–8. Ieee (2008)
13. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2009)
14. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4733–4742 (2016)
15. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
16. Girshick, R., Donahue, J., Darrell, T., Malik, J.: 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 (2014)
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
20. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
21. Tanaka, M., Shibata, T., Okutomi, M.: Gradient-based low-light image enhancement. In: 2019 IEEE International Conference on Consumer Electronics (ICCE). pp. 1–2. IEEE (2019)
22. Priyanka, S.A., Wang, Y.K., Huang, S.Y.: Low-light image enhancement by principal component analysis. IEEE Access **7**, 3082–3092 (2018)
23. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. IEEE Transactions on image processing **26**(2), 982–993 (2016)
24. Kniaz, V., Grodzitskiy, L., Knyaz, V.: Deep learning for coded target detection. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences (2021)
25. Guo, T., Wei, Y., Shao, H., Ma, B.: Research on underwater target detection method based on improved msrnp and yolov3. In: 2021 IEEE International Conference on Mechatronics and Automation (ICMA). pp. 1158–1163. IEEE (2021)

26. Cui, Z., Qi, G.J., Gu, L., You, S., Zhang, Z., Harada, T.: Multitask aet with orthogonal tangent regularity for dark object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2553–2562 (2021)
27. Dong, X., Wang, G., Pang, Y., Li, W., Wen, J., Meng, W., Lu, Y.: Fast efficient algorithm for enhancement of low lighting video. In: 2011 IEEE International Conference on Multimedia and Expo. pp. 1–6. IEEE (2011)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28**, 91–99 (2015)
29. Land, E.H., McCann, J.J.: Lightness and retinex theory. *Josa* **61**(1), 1–11 (1971)
30. Bochkovskiy, A., Wang, C.Y., YOLO, H.Y.M.L.: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
31. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 385–400 (2018)
32. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
33. Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H.: M2det: A single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9259–9266 (2019)