Hindawi
Security and Communication Networks
Volume 2022, Article ID 8080301, 11 pages
https://doi.org/10.1155/2022/8080301

Research Article

An Efficient Multiscale Pyramid Attention Network for Face Detection in Surveillance Images

Ming Liu,[1] **Ruijie Cai**,[1] **Lukai Li**,[1] **Jiafeng Wang**,[2] **and Qichao Yang**[1]

[1]*State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China*
[2]*School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450001, China*

Correspondence should be addressed to Qichao Yang; yangqichaoo@foxmail.com

Received 15 March 2022; Revised 18 June 2022; Accepted 27 June 2022; Published 9 August 2022

Academic Editor: Je Sen Teh

Although a large number of works have been done to explore efficient face detection in various scenes, practical face detection in unconstrained condition of varying lighting, pose, scale, and occlusion remains a challenging task. The primary limitation of existing solutions is that they are vulnerable to influence from the wild environment. Practical features extraction plays a crucial part in the face detection of low-quality images. Based on the EfficientNet, this paper builds a novel pyramid attention network to integrate multilevel features with rich context messages. Firstly, a context model is exploited to increase the receptive fields at the beginning of the network. Secondly, stacked pyramid feature attention modules and feature fusion simultaneously selectively integrate the contextual information and enable spatial details, thus enhancing the capacity to detect faces on hard images. In addition, hard samples augmentation of the training sets is conducted, which is beneficial for improving the accuracy. A thorough study on ablation verifies the effect of the proposed strategies. Moreover, extensive experiments on Wider Face and FDDB datasets, remarkably pushing the accuracy both of 96.3% (2%↑), demonstrate the performance of the proposed deep face detector which is superior and outperforms most of the preexisting methods. The method presented in this paper can perform the task of face detection in surveillance images well.

1. Introduction

During recent decades, the deployment of video surveillance systems has been increasingly intensive. Enormous video data are far beyond the ability of humans to process it manually. Continuing to rely on manual judgment and processing of video content are hard to meet the actual demands. However, the response of public security organs to control the social security situation needs to be further improved. Based on deep learning, surveillance video face recognition provides a piece of more powerful information means for the judiciary, public security, video patrol, and investigation. Although face detection technology has made adequate progress, the performance of face detection largely degrades in the environment of low image quality, which means the realistic face detection system still faces many challenges. These challenges mainly stem from the unpredictable changes that exist in face images, such as expressions, postures, lighting, resolution, and masking. These changes can cause severe data inconsistency between the training dataset and the test dataset [1]. Effectively dealing with these problems and improving recognition efficiency are still tricky problems in face detection systems.

Face detection is an essential task in many face-related visual works, such as face recognition and face editing. Face detection task can be considered as an advanced semantic feature detection problem. Due to the large distance between the surveillance camera and the objects, the captured faces are usually of low resolution. Uncontrolled attitudes and lighting conditions impact the performance of the face detection algorithm adversely. The off-the-shelf face recognition performances cannot meet the practical requirements.

The primary purpose of surveillance video is to deal with more and more complex security works. Especially after some large-scale terrorist attacks, it strongly stimulated the

world's demand on video surveillance. Recently, intelligent video surveillance technology has been dramatically developed and emerged as a more mature technology smart security system. Intelligent video surveillance technology has gradually entered the scale application stage. In video surveillance, the face feature, as a very critical biological feature, is always the object favoured by researchers Before the development of deep learning, Viola and Jones (V-J) [2] proposed cascading face classifiers trained by Haar features and AdaBoost methods as a landmark algorithm in the field of face detection. After that, many works focused on studying more sophisticated hand-crafted features to improve classifier performance. In addition to cascading structures, other works also developed formable part models to handle face detection tasks. However, the traditional face detection algorithms highly depend on the characteristics of the hand-crafted design. These characteristics are inadequate to meet the challenges of unconstraint conditions.

With the promotion of deep learning, face detection performances have been greatly improved. Cascade-CNN [3] continued V-J [2] framework, replacing hand-crafted features with CNN features. It developed CNN-based cascading network structures for face detection and achieved good detection accuracy. Qin et al. [4] proposed an overall training cascade-CNN for end-to-end optimization. Faceness [5] used a multitask CNN to train a series of face attribute classifications for detecting partially obscured faces. MTCNN [6] further expanded the idea of cascading CNN, which solved both face detection and feature point positioning in the form of multitask and optimized the network structure through reasonable decomposition tasks. Its small CNN network cascading method not only had high precision but also had a fast detection speed due to its simple network structure. In addition, many face detection methods drew on the ideas of target detection. Jiang et al. [7] applied faster R-CNN, a representative of the object detection field to face detection tasks, resulting in satisfactory results. CMS-RCNN [8] applied human body information to faster R-CNN with contextual information fusion processing, further improving the performance of detection. The pictures collected by the surveillance video in the natural environment are greatly affected by the interference of the collection environment. There are many faces with low resolution, uneven light, changing shapes, and expressions. Especially in dark light, many facial textures are greatly disturbed. This has dramatically hindered the deployment of face detection in practical applications.

Unfortunately, there are many problems that need to be solved when applying to real-world scenarios: (1) surveillance cameras are used for security and anomaly detection in public places, companies, campuses, and other venues. Because the installation position of the surveillance camera is fixed, diverse sizes of faces are produced by changing distances. The multiscale problem is very prominent. Especially, the performance of face detection will be significantly reduced as the scale shrinks. Therefore, the application of face detection in video surveillance needs to concentrate on small-size faces. (2) Balance between accuracy and computation. In the video surveillance scene, a critical requirement is that the algorithm runs in real-time, which means its speed requirements are more stringent. The prior face detection algorithms with superior precision have the characteristics of slow detection speed because the improvement of precision is to calculate more information as a price.

Information from small objects can probably be weakened since the spatial resolution of the feature map in a large context reduces the information integration [9, 10]. Usually, the shallow layer has only lower semantics that may not be sufficient to identify the information instance of the object category. Actually, face detection has more reference to general object detection. Because the quality of face detection directly affects the technical trend and the application landing of face analysis, it has attracted wider attention in academia and industry. As a second classification task, owing to the nature of face shape and intricate background, we modify the method of object detection according to facial characteristics. For example, all of the faster R-CNN [7], R-CNN [8], and YOLO [11] have been extended and applied to face detection tasks. However, as described in the above paragraph, these object detection methods are often time-consuming. Moreover, these object detection methods are dependent on the proposal window generation method to locate the targets and cannot effectively locate small targets. Making mention of designing a convolutional neural network for the low-quality video scenario, two key points need to be considered. Firstly, the architecture is somewhat limited in its number of layers due to the low resolution of face images. Second, a robust extracting is required because the face descriptor for each image should be as compact as possible. Therefore, we choose EfficientNet [12, 13] as the backbone, and it has shown satisfactory performance in many works.

For this paper, the main contributions are as follows:

(i) Based upon EfficientNet, we augment the architecture using a context model and anchor strategy, which is more suitable for finding tiny faces

(ii) Stimulated by the success of visual tasks, we combine feature pyramid structure and attention mechanisms to design an efficient one-stage face detector, which enhances low-level semantics information

(iii) Based on Wider Face [14] and FDDB [15] datasets, we conducted a considerable amount of experiments to verify the efficiency and improvements of our model, confirming the applicability of our strategies

2. Related Work

2.1. Face Detection. Thanks to the significant accomplishments of general object detection, face detectors have an outstanding improvement in performance. The deep learning models trained on wide-ranging image data sets supply more discriminating features for face detectors than traditional hand-crafted features. In addition, end-to-end training methods promote better optimization. Depending

on whether following the proposal regions, the deep face detection approach can be segmented into two subcategories.

2.1.1. Sliding Windows-Based Methods. This kind of method outputs face detections at every location in a feature map at a given scale. These detections contain two parts: face detection score and bounding box. The SSH [16], based on an RPN, detects faces with various layers in a single forward network contemporaneously. In ref. [6], the author employs a deep cascaded multitask structure that integrates face detection with alignment tasks through unified CNNs. MTCNN [6] brings in candidate regions from the first CNN rapidly. Besides, another two large-scale CNNs screen out high-confidence results. Tian et al. [17] employed feature fusion and segmentation branch to expand the relationship between high-level and low-level. They notably improve the detecting accuracy. For real-time speed, Liu et al. [18] used a multibranch fully convolution network, which treats faces of diverse sizes through a single pass.

2.1.2. Region-Based Methods. After an expeditionary survey of finding small faces, Hu and Ramanan [19] utilized the characterizations of scale resolution to detect tiny objects. PyramidBox [20] designs a context-assisted single-shot face detector that makes full use of context information to overcome the difficulties of face detection. A low-level FPN and a context-sensitive module are added to the backbone. The context-sensitive prediction module acts as a branch network from each pyramid detection layer to get the final output. With an aim to get a trade-off between efficiency and accuracy, DCFPN [21] first shrinks the resolution of the input image, and they use a dense anchor strategy to maintain high accuracy. Object detection based on anchor has developed rapidly, and face detection has also gotten great ahead. However, the detection effect for small faces continues to be not very good. S3FD [22] mainly analyzes and improves the problem of the low detection rate of tiny faces. It enhances the recall rate of small-scale faces through the anchor matching strategy.

2.2. Attention Mechanism. Because of limited visual information, when reading a photo, the human optical system selectively concentrates on a special component of the photo while neglecting the remainder. As depicted in Figure 1, while the primary content of the figure is sky, people can first easily catch the airplane in the image. To simulate this procedure in artificial neural networks, an attention mechanism is presented. It has achieved excellent accomplishments in many fields such as image reconstruction [23], visual question answering [24], and face recognition.

Compared with convolution, the advantage of the attention mechanism is that it has a large number of elastic mappings. This is an effective method used to strongly connect any part of the input field. There are many works employing the attention mechanism to elevate the accuracy of the CNN classification model [25, 26]. In ref. [25], the author adopts a cascaded attention mechanism to guide the various layers of CNN and connect them in series to obtain discriminative representation as the input of the final linear classifier. A more recent work [26] performed universal object detection with domain-attention. However, it performed only moderately well without prior knowledge. Unlike the above methods, we attempt to put the pyramid structure and attention mechanism. The attention mechanism is applied to a multilevel pyramid network, and distinguished regions are used to classify and suppress noise information.

2.3. Low-Quality Surveillance Images. Due to the increasing demand for security, face recognition is more attractive than ever before. Different from gallery images, surveillance images are characterized by out-of-focus blur, low contrast, and low resolution. At present, most image-set technologies of face recognition are image-based face recognition. However, this result cannot be extended to real-life monitoring scenarios. One of the main challenges of surveillance images is the low image resolution, which may be because the monitored object is too far away from the camera to capture high-resolution images. Unfortunately, the methods developed for high-resolution images cannot be well extended to low-resolution images. It is very expensive and infeasible to build large-scale native face monitoring data as a benchmark for broader research. This is due to the restrictions on data access and very cumbersome data labels to a large extent and the high cost.

3. Proposed Method

Inspired by numerous works, we develop our detector for surveillance face detection. Firstly, we present the base network. Then, we modify the practical pyramid architecture and add four key modules to resolve the low-quality surveillance face. The overview architecture of our method is demonstrated in Figure 2.

3.1. Base Network: EfficientNet. We select EfficientNet as our base network for extracting features from face images and as a baseline for our subsequent ablation experiments. EfficientNet obtains compelling accuracy and efficiency performance by leveraging neural architecture search in object detection tasks. It expands width, depth, or input resolution of the backbone in a principled way instead of the traditional ways. Abundant experiments [12] prove that the precision of the model does increase with the scaling of a certain dimension of the model. When the scale of the model increases to a certain extent, the precision of the model will not continue to grow with the increase of the scale. Based on that, by tuning the width and height of the network structure, the generalization capability and representation power of the network have made a marked improvement. In previous works, expanding the width or height of the network structures tended to achieve better accuracy. However, the result rapidly saturates after gaining 80% accuracy. While it is feasible to transform width and height

Figure 1: The illustrations of attention mechanism. It cleverly highlights the areas of interest. The bright color parts mean more attention-focused. Large areas of blue indicate areas that paid less attention.

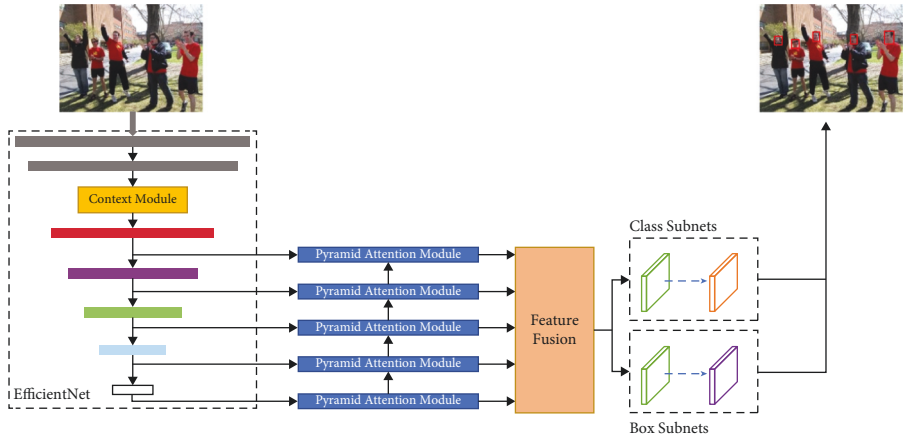

Figure 2: Pipeline for our method. We employ EfficientNet (the left panel) as the base network (the colorful boxes mean various CNN layers; we elaborate on this in Section 3.1), and we insert three dilated convolutions in the context module. The remaining section is built using a pyramid attention module, feature fusion, and two subnets.

discretionarily, arbitrary scaling still requires dreary manual adjusting. EfficientNet adopts a novel network scaling way in which dynamic adapts the sizes of depth, width, and resolution of networks. The neural network is represented by the following formula:

$$\mathcal{N} = \bigodot_{i=1\ldots s} \mathcal{F}_i^{L_i}\left(X_{H_i,W_i,C_i}\right), \tag{1}$$

where $\mathcal{F}_i^{L_i}$ means the layer $\mathcal{F}_i$ is repeated L_i times in stage i and H_i, W_i, C_i mean the matrix of tensor X of layer i. Formula (1) interprets every execution unit of the neural network as a functional expression, which sets appropriate optimization goals for the subsequent.

In order to obtain the more accurate model by limited resources, we formulate (1) as an optimization problem:

$$\begin{cases} \max\limits_{d,w,r} \text{Accuracy}\left(N\left(d,w,r\right)\right), \\ \text{s.t. } \mathcal{N}\left(d,w,r\right) = \%\bigodot_{i=1\ldots s} \widehat{\mathcal{F}}_i^{d\cdot\widehat{L}_i}\left(X_{r\cdot\widehat{H}_i,r\cdot\widehat{W}_i,w\cdot\widehat{C}_i}\right), \\ \text{Memory}\left(\mathcal{N}\right) \le \text{target}_{\text{memory}}, \\ \text{FLOPS}\left(N\right) \le \text{target}_{\text{flops}}, \end{cases} \tag{2}$$

where d, w, r are the coefficients for scaling network width, depth, and resolution and $\widehat{\mathcal{F}}_i, \widehat{L}_i, \widehat{H}_i, \widehat{W}_i, \widehat{C}_i$ denote predefined parameters in the baseline network. Apparently, the ultimate optimization goal of the model is to maximize the prediction accuracy by adjusting the scaling ratio of the depth, width, and resolution of the model.

Then, we define a simple yet effective compound coefficient ϕ to uniformly scale the dimensions of network:

$$\begin{cases} \text{depth: } d = \alpha^{\phi}, \\ \text{width: } w = \beta^{\phi}, \\ \text{resolution: } r = \gamma^{\phi}, \\ \text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \\ \alpha \ge 1,\ \beta \ge 1,\ \gamma \ge 1, \end{cases} \tag{3}$$

where α, β, γ are the values representing the proportion of resource distribution. Intuitively, ϕ is a user-specified constant that manages how many computing resources and storage memory are valid for model expansion. We set ϕ as 3 for an excellent performance. Table 1 illustrates the architecture of EfficientNet.

The EfficientNet acts as feature extractor for the face images. The initial layers of the EfficientNet abstract plain low-level features as edges and corners. The deeper layers extract more sophisticated high-level features as the semantics of the target. However, the receptive fields of the high-level features are much larger, making them unprecise

Table 1: The architecture of backbone.

Stage i	Operator	Resolution	Channels	Layers
1	Conv3 × 3	512 × 512	45	1
2	MBConv1	256 × 256	22	1
3	MBConv6	256 × 256	24	4
4	MBConv6	128 × 128	40	4
5	MBConv6	128 × 128	80	5
6	MBConv6	64 × 64	112	5
7	MBConv6	64 × 64	192	7
8	MBConv6	32 × 32	320	2
9	Conv1 × 1 and pooling and FC	32 × 32	1280	1

MBConv denotes mobile inverted convolutional bottleneck.

in localizing faces. Thus, we combine the high-level and low-level features together to deliver them as input to the attention pyramid networks for adaptively weighting allocating.

3.2. Single-Stage Detector. Furthermore, depending on the characteristics of tiny faces, we make two modifications to the EfficientNet architecture so that it can detect faces more efficiently. First, we abandon all the fully-connected layers and the last pooling layer of the EfficientNet to retain more details. The revised EfficientNet provided feature maps at five stages. Second, the sizes of anchors are particularly allocated according to the available receptive fields, accurately detecting faces with various resolutions in different scenes.

According to the articles [22, 23] on object detection based on anchor, the size of anchors has a huge influence on object detection. If the scale and ratio settings are not appropriate, the recall may not be high enough, or anchor may largely affect classification performance and speed. On the one hand, the vast majority of anchors is distributed in the background area if anchor is too dense. The loss of target box regression of the loss plays a minor role; on the other hand, preset anchor shapes cannot address the targets with extreme size and extreme aspect ratio. Therefore, when we handle the anchor design in the actual object detection, we are supposed to consider the distribution of scale, ratio, and anchor in the feature map. When the recall rate of a certain scale target is low, we should consider adding a small-scale anchor; Also, in the target of leakage, when the aspect ratio is more uniform, a ratio should be increased; when the false alarm is high in the object detection results, we should consider whether the number of the anchors is too high. We count the face resolution information in the Wider Face dataset [14], as demonstrated in Figure 3. The anchors setting of the network can be seen in Table 2. Considering the ratio of faces, we set the aspect ratio of anchors as 1 : 1 and 1: $\sqrt{2}$, because major faces are square-like shapes, and profile faces can be regarded as a rectangle. Small-size and large-size anchor boxes focus on shallow and deeper feature maps, receptively.

3.3. Context Module. Face detection in surveillance scenes is comparatively tricky due to the pedestrians being detected in a long-distance from the surveillance cameras. It is instituted that low-quality frames impair the performance of face detection. It turns out that contextual information is beneficial for detecting small faces [19]. The properties of covn3_3 have adequate spatial resolutions mapping from the original image, even though they have neither steady semantics nor context information. For receiving more receptive field, three dilated convolution layers are embedded at the beginning of the EfficientNet network. The structure of dilated convolution is shown in Figure 3. A standard 3 × 3 convolution kernel can only see the size of the corresponding region 3 × 3 (Figure 4), but dilated convolution makes it possible to see a more extensive range for the convolution kernel. Figures 4(b) and 4(c) can be understood as the convolution kernel size is still 3 × 3, but there is an interval between each convolution point. These three dilated convolutions can receive 3 × 3, 7 × 7, and 11 × 11 fields to extract multiscale context information, respectively. The weight of the rest points is 0. The receptive field expands exponentially while the number of parameters increases linearly. For the model, its running speed will not be influenced by this extra computation.

Unlike other multibranch networks sharing the standard input, we slice the input channels fairly enrolling into three dilated convolutions, making a smaller number of channels for each branch as shown in Figure 5. According to this method, we can have fully explored the faces which are not wholly detected.

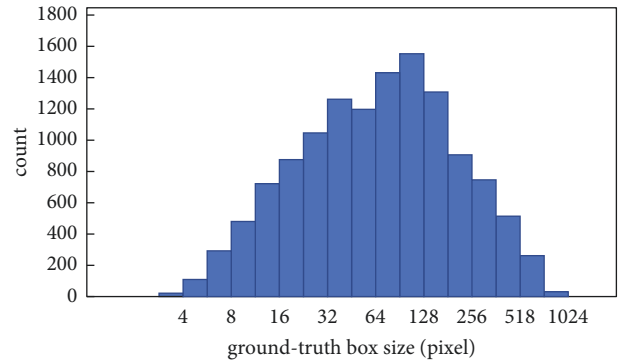

Figure 3: The resolution distribution of faces in the Wider Face dataset.

Table 2: Parameters of the five anchors for various sizes of surveillance face detection.

Layer	Stride	Anchor size (pixels)	Boxes
Conv3_3	4	16/ $\sqrt{2}$, 16	32768
Conv4_3	8	32/ $\sqrt{2}$, 32	8192
Conv5_3	16	64/ $\sqrt{2}$, 64	2048
Conv6_2	32	128/ $\sqrt{2}$, 128	512
Conv7_2	64	256/ $\sqrt{2}$, 256	128

3.4. Pyramid Feature Attention Module. Attention technique can be fairly understood as a method used for enhancing the response of the parts that have most information and

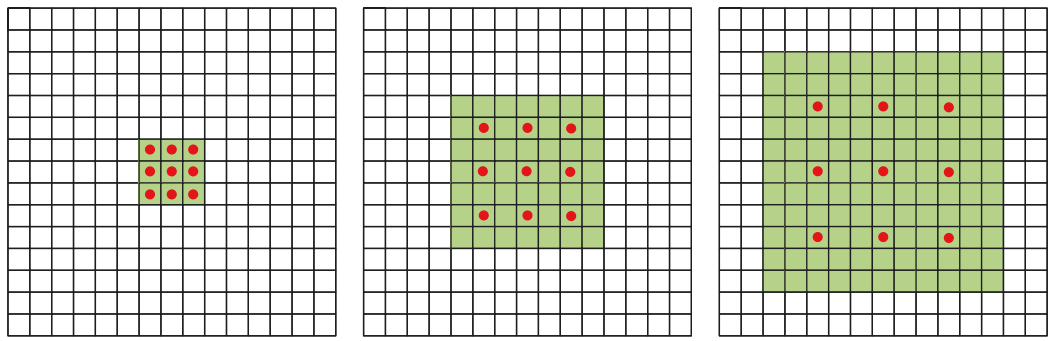

Figure 4: The receptive fields illustration of context module with three dilated rates. (a) Dilated rate = 1, (b) dilated rate = 2, and (c) dilated rate = 3.

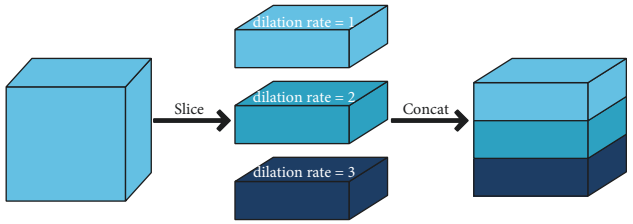

Figure 5: The slice-branch-concat operation of our context module.

suppressing the activation of others. It shows that background has a large influence, because the background has a high fraction in the images. But we should prevent them from being activated because this information commonly is useless to the classification. For the purpose of emphasizing the essential face features from the whole image, we propose a multiscale attention pyramid module, which calculates the corresponding attention map, to emphasize the scores that can be detected in small faces. The dimension of the output of the attention module operation is consistent with the inputs in order to facilitate access to the neural network as a common component. The structure of pyramid feature attention module is shown in Figure 6.

Specifically, the output feature of the attention module is calculated by

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \tag{4}$$

where x means the input feature. y denotes the output feature with the same size as x. $f(x_i, x_j)$ is the function of mapping the relationship between x_i and x_j. $g(x_j)$ computes a feature vector of the input signal at position j. For easier implementation, we set a 1×1 convolution filter as $g(x_j)$. $C(x)$ is a normalization function.

For better visual reasoning, we select concatenation function as $f(x_i, x_j)$:

$$f(x_i, x_j) = \text{Leaky ReLU}\left(w_f^T \left[\theta(x_i), \varphi(x_j)\right]\right). \tag{5}$$

w_f means a weighting coefficient that maps the concatenated vector to a scalar. The scalar is activated with Leaky ReLU for reducing information loss.

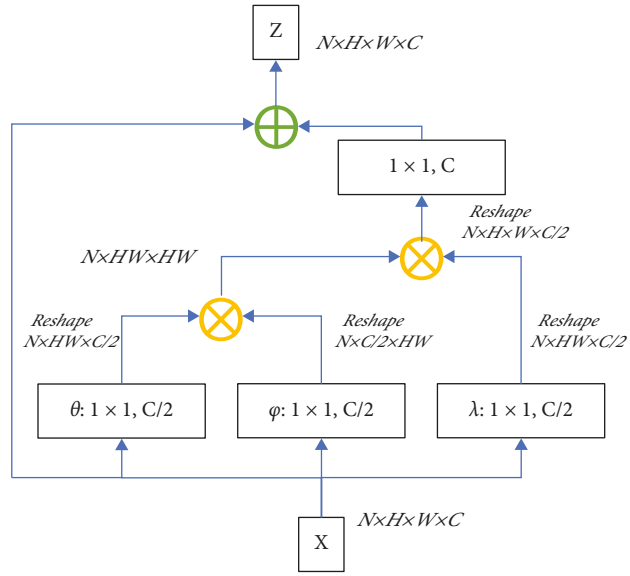

Figure 6: The structure of the pyramid feature attention module.

3.5. Feature Fusion. Feature fusion mechanism is utilized for enabling features to hold some low-level and high-level information. High-level features abstracted by deeper convolution layers include semantic information; however, they generally waste details such as positions and colors that are available in the detection. In contrast, low-level features contain more detailed information but introduce nonspecific noise. Feature fusion has been adopted by many object detection tasks to enhance performance by combining high-level and low-level features. Therefore, we choose the feature fusion technique to improve detection accuracy.

Referring to multiscale fusion, in the fusion of different input characteristics, most of the previous studies just merged the various feature eigenvectors into a sequence. A survey reveals that the contribution of varying input features is often unequal. To address this issue, we cascade a learnable weight to evaluate the worth of different input layers. Mathematically, the feature fusion function is formulated as

$$f^f = \%_{1 \le i \le 5} \odot f_i^a \cdot \omega(f_{i+1}^a; \rho) + f_i^a, \tag{6}$$

where f_i^a represents the out feature of the attention module. ω is the transposed convolution. ρ is the transposed parameter of ω. We combine high-level features with low-level features through element-wise multiplication. The combined vector involves both spatial and channel-wise information.

3.6. Hard Face Mining. In the training process, large amounts of training sets only contain high-quality faces, which does not help learn the robust detector for hard faces. Despite their success in most images, a significant performance gap persists, especially for hard training samples with low resolution, blur, and occlusion parts. More specifically, we utilize dynamically setting a difficulty rate to train images, which can judge whether an image is already well-detected or still useful for further training. This allows us to take full advantage of images that are not entirely detected to better facilitate the subsequent training process. We argue that this strategy can make our detector more robust for faces with challenging influence, meanwhile, without adding any computing costs. The hard face mining evolves the following four steps:

Step 1: for each dataset, we consider the whole samples in the training set as hard samples.

Step 2: when starting training, we use formula (7) to calculate the corresponding difficulty scores with all samples.

Step 3: if the score is below a threshold, the image sample will be marked as a hard face.

Step 4: after each epoch training, we collect all hard faces into a new subset. It will be trained in the next epoch.

$$\text{Score}(I;\theta) = \min_{a\in\mathscr{A}(I)^+} \frac{\exp\left(l(I;\varepsilon)_{a,1}\right)}{\exp\left(l(I;\varepsilon)_{a,1}\right) + \exp\left(l(I;\epsilon)_{a,0}\right)}, \quad (7)$$

where $\mathscr{A}(I)^+$ is the set of positive anchors for image I. l is the classification logit. $l(I;\varepsilon)_{a,1}$ and $l(I;\varepsilon)_{a,0}$ are the logits of anchor a for the image I to be foreground face and background.

4. Experiments

Firstly, some testing datasets and protocols will be introduced, which are aimed at verifying the performance of our methods. Next, the realization process is clearly described. We display the comparisons with the most advanced approaches and the capability of boosted EfficientNet. After, according to ablation studies, we research each method to get their performance, therefore, in order to explore the effectiveness of the boosted EfficientNet, a lot of experiments are achieved.

4.1. Benchmark Datasets and Metrics. Wider Face dataset [14] is a large-scale public face database containing 393,703 faces from 32,203 images. It is comprised of three parts: training (40%), validation (10%), and test (50%). In this dataset, the faces have occlusions, poses, race, and face bounding box annotations. According to the complexity of detection tasks, the provider of the database divides it into three parts: easy, medium, and hard subsets. The evaluation metric chosen by us used to evaluate the performance of the model is the average precision (AP) metric.

The FDDB dataset [15] aims to evaluate the strength of unconstrained face detection. FDDB provides 2,845 images involving a number of 5,171 faces collected in realistic conditions. The researchers manually marked the bounding box for the images. We use the receiver operating characteristic (ROC) curves for evaluating the performance.

Table 3: The ablation experiments of our strategies on the Wider Face validation test.

Method	Baseline					+DA
+ CM		√	√	√	√	√
+ FAP			√	√	√	√
+ FF				√	√	√
+ HFM					√	√
Accuracy (mAP[easy])	92.4	93.0	93.4	93.6	94.8	96.3
Accuracy (mAP [medium])	91.0	91.3	91.5	91.8	93.6	95.1
Accuracy (mAP[hard])	78.6	82.3	85.1	85.4	87.9	89.2

Table 4: Results comparison on Wider Face validation set.

Algorithms	Backbone	Easy	Medium	Hard
LDCF+ [27]	—	79.0	76.9	52.2
Multitask cascade-CNN [6]	—	84.8	82.5	59.8
ScaleFace [28]	ResNet50	86.8	86.7	77.2
CMS-RCNN [8]	VGG16	89.9	87.4	62.4
MSCNN [29]	VGG16	91.6	90.3	80.2
HR [19]	ResNet101	92.5	91.0	80.6
Zhu [30]	ResNet101	94.9	93.3	86.1
FAN [29]	ResNet50	94.3	94.2	88.8
Gao [31]	TinyYOLOv3	95.26	89.2	77.9
DBCFace [32]	ResNet50	95.84	94.96	90.34
Ours	Efficient-B7	96.3	95.1	89.2

4.2. Implementation Data Augmentation. During this testing, we resize the input images to 512×512 pixels. The threshold of nonmaximum suppression (NMS) for filtering out the redundant boxes is configured as 0.6. We train our model starting from a stochastic gradient descent optimizer with a momentum of 0.9, weight decay of $1e^{-4}$. For anchor settings, we set 12 anchor scales from the set {16, 32, 64, 128, 256} and anchor ratio as 1 : 1 and 1: $\sqrt{2}$. The experiment comes true by publicly available Pytorch framework on a machine for neural network training, which has 24 GB memory and four Tesla K80 GPUs.

Our primary task is to solve the problem of face detection in surveillance images, there is a lack of occluded and low-resolution faces in Wider Face [14] (around 20%). Data augmentation, as a common method in deep learning, is helpful to enhance the model in generalization ability and

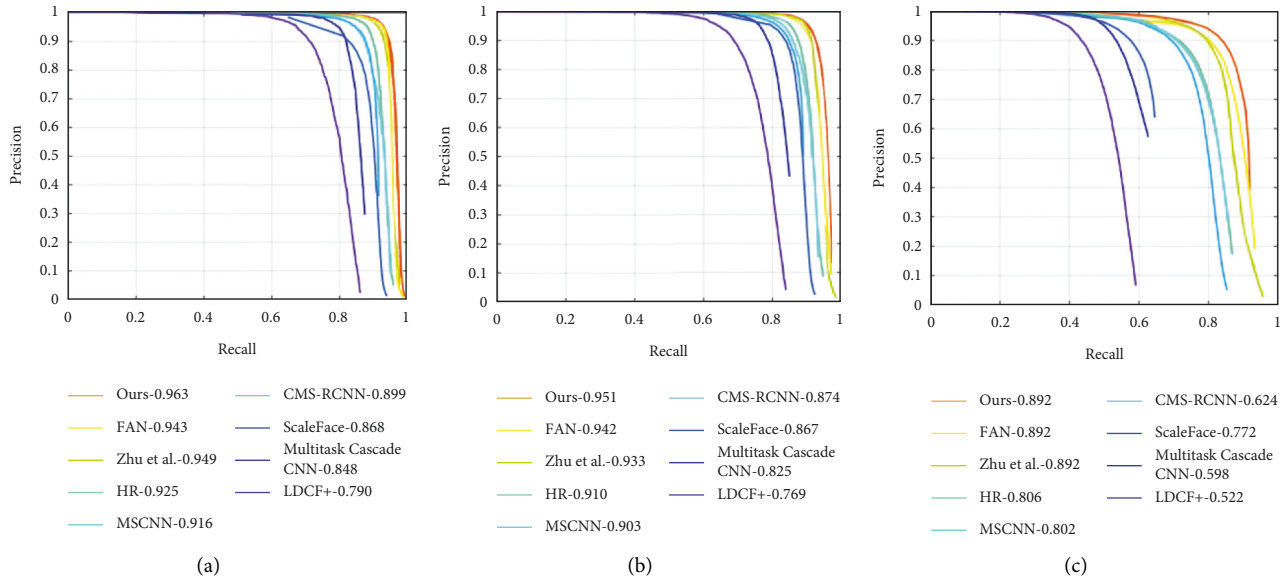

Figure 7: The evaluation results on Wider Face: easy subset (a), medium subset, (b) and hard subset (c).

Figure 8: Selected detected faces on Wider Face.

accuracy. Before the training, we disturbed and increased the data samples by image processing methods: (1) random rotation from −20° to +20°. (2) Random erasing 10% of the images. (3) Adding random value matrix sampled from the Gaussian distribution to the RGB pixels of the image. After these operations, we expanded the original training set by 2 times. In addition to being suitable for the serious occluded face, our augmentation is beneficial for small face detection, as more small faces are extended. Our data augmentation strategy is designed to scale up training data.

4.3. Ablation Studies. We do detailed ablation investigations to study each strategy respective roles on the face detector, including context module (CM), feature attention pyramid (FAP), feature fusion (FF), and hard face mining (HFM).

Table 3 shows the ablation analysis. It is easy to find that the baseline model also achieves good performance (92.4 mAP) due to the effectiveness of deep learning. The bottom three rows demonstrate that our four strategies practically

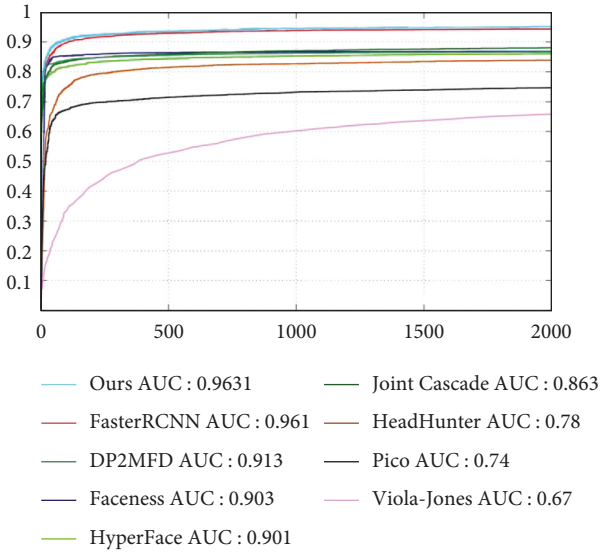

Figure 9: Performance evaluation on the FDDB dataset.

Figure 10: Selected detected faces on the FDDB dataset.

enhance the capability, particularly tiny faces. By context module, the AP is improved by 0.3% and 3.7% on medium and hard subsets, respectively, without increasing FLOPS. After we took in the feature attention pyramid module, the mAP was further expanded up to 85.1, and the mAP (hard) prominent raised by 2.3%. When inserting feature fusion and hard face mining into our model, the mAP increases by 85.4 and 87.9 on the hard set. We can find out that the mAP of the hard set has a steady increase from 78.6% to 89.2%.

With the help of data augmentation, our performance is improved over 1.4% on three subsets obviously. It shows that data augmentation is crucial for uneven data distribution.

The considerable improvement on hard subset validates our strategies indeed enhance the robustness of face detector, which means it can utilize more discriminative features from low-quality faces.

4.4. Results on Wider Face. Table 4 and Figure 7 illustrate the results of recently reported face detection models with our designed method. We compare our method against 8 recent deep learning approaches: LDCF+ [27], multitask cascade-CNN [6], ScaleFace [28], CMS-RCNN [8], MSCNN [29], HR [19], Zhu [30], and FAN [29]. We discover that our proposed method performs compelling performance with the SOTA methods on easy set (96.3%) and medium difficulty set (95.1%), respectively. After adopting our four strategies, our model outperforms all methods on the hard subset, reaching 0.892 mAP, and also has a better performance compared to most face detectors when applied to the Wider Face dataset.

In such case, the apparent enhancements on hard subclass verified that our feature attention pyramid module virtually reserves semantics from lower-level feature maps to high-level feature maps.

Figure 8 shows some detecting results on Wider Face dataset. It exhibits the robustness of our model in challenging cases.

4.5. Results on FDDB. From Figure 9, it can be seen that the score achieved by our proposed model is higher than any other methods on the continuous ROC curve. The proposed face detector has an accuracy of 96.31% when the number of false positives is equal to 2,000. It shows the better performance of our model in various scales, serious block, and greater blur regression in unconstrained scenarios.

Figure 10 reveals some qualitative detection on the FDDB dataset. According to the results, we can get the conclusion that the proposed method based on deep learning techniques is effective for face detection.

Table 5: The runtime and parameters compare our model with different approaches on Wider Face dataset.

Method	Backbone	mAP	Speed	Parameters	Runtime
FasterRCNN [33]	VGG19	71.2	<6 FPS	24M	180 ms
ScaleFace [28]	ResNet50	77.2	<8 FPS	37M	130 ms
HR [19]	ResNet101	80.6	3.1 FPS	58.16 M	360 ms
Gao [31]	TinyYOLOv3	91.5	<6 FPS	33M	182 ms
Ours	Efficient-B3	89.2	24 FPS	15M	80 ms

4.6. Runtime Analysis. For a better comprehension of the advantage about our method, we conduct a runtime comparison with the other three methods. As discussed in Section 3, different from early ways that arbitrarily scale these factors, our backbone can be scaled by an effective coefficient to balance the dimensions of depth, width, and resolution. As described in (3), we use the small grid search method to seek three constants. We set ϕ as 3 in (3) for controlling the available resources. With proper optimization of the new scaling strategy, our method becomes more outstanding in the advantage of runtime efficiency. As Table 5 shows, our method considerably outperforms all previous detectors on the Wider Face dataset. Compared with complicated CNN frameworks, our model is much smaller. The efficient model runs eight times faster than HR, while having a strong detection ability under practical runtime speed conditions.

5. Conclusions

In this work, we have designed a practical method to detect the face in low-quality surveillance images. Specifically, we

first develop a context model with multiple scales of view, which is beneficial for long-distance tiny faces detection. Then, we combine the pyramid attention module and feature fusion, which improves the accuracy both in the content and spatial locations. During training, hard face mining is used to handle the class imbalance problem in hard face images detection. Experimental results denote that the method we came up with promisingly promotes the performance of hard face detection. Moreover, the proposed model also enjoys efficient inference speed.

Data Availability

The Wider Face dataset and FDDB dataset used to support the findings of this study can be downloaded from http://shuoyang1213.me/WIDERFACE/ and http://vis-www.cs.umass.edu/fddb/index.html, respectively.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Foundation Strengthening Key Project of Science and Technology Commission (2019-JCJQ-ZD-113).

References

[1] R. Xia, Y. Chen, and B. Ren, "Improved Anti-occlusion Object Tracking Algorithm Using Unscented Rauch-Tung-Striebel Smoother and Kernel Correlation filter," *Journal of King Saud University-Computer and Information Sciences*, 2022.

[2] P. Viola and M. Jones, "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*, pp. 511–518, Kauai, HI, USA, December 2001.

[3] Z. Cai and N. Vasconcelos, "Cascade R-Cnn: Delving into High Quality Object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, Salt Lake City, UT, USA, June 2018.

[4] H. Qin, J. Yan, X. Li, and X. Hu, "Joint Training of Cascaded CNN for Face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3456–3465, Las Vegas, NV, USA, June 2016.

[5] S. Yang, P. Luo, C. C. Loy, and X Tang, "Faceness-net: face detection through deep facial Part Responses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1845–1859, 2018.

[6] K. Zhang, Z. Zhang, Z. Li, and Y Qiao, "Joint face detection and alignment using Multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[8] L. Shi, X. Xu, and I. A. Kakadiaris, "Detecting multi-scale faces using attention-based feature fusion and smoothed context enhancement," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 3, pp. 235–244, 2020.

[9] J. Zhang, W. Feng, T. Yuan, J. Wang, and A. K Sangaiah, "SCSTCF: spatial-channel selection and temporal regularized correlation filters for visual tracking," *Applied Soft Computing*, vol. 118, Article ID 108485, 2022.

[10] J. Zhang, J. Sun, J. Wang, Z. Li, and X. Chen, "An object tracking framework with recapture based on correlation filters and Siamese networks," *Computers & Electrical Engineering*, vol. 98, Article ID 107730, 2022.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, Long Beach, California, USA, May 2019.

[13] J. Wang, Q. Liu, H. Xie, Z. Yang, and H. Zhou, "Boosted EfficientNet: detection of lymph node metastases in breast cancer using convolutional neural networks," *Cancers*, vol. 13, no. 4, p. 661, 2021.

[14] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider Face: A Face Detection benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5525–5533, Las Vegas, NV, USA, June 2016.

[15] V. Jain and L.-M. Erik, "A Benchmark for Face Detection in Unconstrained settings," UMass Amherst technical report, Amherst MA, 2010.

[16] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "Ssh: Single Stage Headless Face detector," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4885–4894, Venice, Italy, October 2017.

[17] Z. Li, X. Tang, X. Wu, J. Liu, and R He, "Progressively refined face detection through semantics-enriched representation learning," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1394–1406, 2020.

[18] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: Single Shot Multibox detector," in *Proceedings of the European Conference on Computer Vision*, pp. 21–37, Amsterdam, The Netherlands, October 2016.

[19] P. Hu and D. Ramanan, "Finding Tiny faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1522–1530, Honolulu, HI, USA, June 2017.

[20] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A Context-Assisted Single Shot Face detector," in *Proceedings of the European Conference on Computer Vision*, pp. 812–828, Munich, Germany, March 2018.

[21] S. Zhang, X. Zhu, Z. Lei, X. Wang, H. Shi, and S. Z Li, "Detecting face with densely connected face proposal network," *Neurocomputing*, vol. 284, pp. 119–127, 2018.

[22] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single Shot Scale-Invariant Face detector," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 192–201, Honolulu, HI, USA, October 2017.

[23] Y. Chen, L. Liu, V. Phonevilay et al., "Image super-resolution reconstruction based on feature map attention mechanism," *Applied Intelligence*, vol. 51, no. 7, pp. 4367–4380, 2021.

[24] W. Li, Z. Yuan, X. Fang, and C. Wang, "Knowing where to look? Analysis on attention of visual question answering system," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 145–152, Munich, Germany, September 2018.

[25] J. Yang, P. Ren, D. Zhang et al., "Neural Aggregation Network for Video Face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5216–5225, Honolulu, HI, USA, July 2017.

[26] F. Wang, M. Jiang, C. Qian et al., "Residual Attention Network for Image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7289–7298, Long Beach,CA, USA, June 2019.

[27] A. Peña, A. Morales, I. Serna, J. Fierrez, and A. Lapedriza, "Facial expressions as a vulnerability in face recognition," in *Proceedings of the 2021 IEEE international conference on image processing (ICIP)*, pp. 2988–2992, IEEE, Anchorage, AK, USA, September 2021.

[28] S. Zhang, C. Chi, Z. Lei, and S. Z Li, "RefineFace: refinement neural network for high performance face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4008–4020, 2021.

[29] Y. Liu, P. Sun, N. Wergeles, and Y Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Systems with Applications*, vol. 172, Article ID 114602, 2021.

[30] L. Song, J. F. Yang, Q. Z. Shang, and M. A Li, "Dense face network: a dense face detector based on global context and visual attention mechanism," *Machine Intelligence Research*, vol. 19, no. 3, pp. 247–256, 2022.

[31] J. Gao and T. Yang, "Face detection algorithm based on improved TinyYOLOv3 and attention mechanism," *Computer Communications*, vol. 181, pp. 329–337, 2022.

[32] X. Li, S. Lai, and X. Qian, "DBCFace: towards pure convolutional neural network face detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1792–1804, 2022.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, pp. 91–98, 2015.

[34] S. Zhang, L. Wen, Z. Lei, and S. Z Li, "RefineDet++: single-shot refinement neural network for object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 674–687, 2021.