information

MDPI

Article

A Lightweight Face Detector via Bi-Stream Convolutional Neural Network and Vision Transformer

Zekun Zhang, Qingqing Chao, Shijie Wang and Teng Yu *

College of Electronic Information, Qingdao University, Qingdao 260000, China; 2021020655@qdu.edu.cn (Z.Z.); 2021023773@qdu.edu.cn (Q.C.); 2021023782@qdu.edu.cn (S.W.)
* Correspondence: yuteng@qdu.edu.cn

Abstract: Lightweight convolutional neural networks are widely used for face detection due to their ability to learn local representations through spatial induction bias and translational invariance. However, convolutional face detectors have limitations in detecting faces under challenging conditions like occlusion, blurring, or changes in facial poses, primarily attributed to fixed-size receptive fields and a lack of global modeling. Transformer-based models have advantages on learning global representations but are insensitive to capture local patterns. To address these limitations, we propose an efficient face detector that combines convolutional neural network and transformer architectures. We introduce a bi-stream structure that integrates convolutional neural network and transformer blocks within the backbone network, enabling the preservation of local pattern features and the extraction of global context. To further preserve the local details captured by convolutional neural networks, we propose a feature enhancement convolution block in a hierarchical backbone structure. Additionally, we devise a multiscale feature aggregation module to enhance obscured and blurred facial features. Experimental results demonstrate that our method has achieved improved lightweight face detection accuracy with an average precision of 95.30%, 94.20%, and 87.56% across the easy, medium, and hard subdatasets of WIDER FACE, respectively. Therefore, we believe our method will be a useful supplement to the collection of current artificial intelligence models and benefit the engineering applications of face detection.

Keywords: artificial intelligence; face detection; transformer

check for updates

Citation: Zhang, Z.; Chao, Q.; Wang, S.; Yu, T. A Lightweight Face Detector via Bi-Stream Convolutional Neural Network and Vision Transformer. *Information* **2024**, *15*, 290. https://doi.org/10.3390/info15050290

Academic Editors: Jae-Mo Kang and Vikas Palakonda

Received: 30 April 2024
Revised: 15 May 2024
Accepted: 17 May 2024
Published: 20 May 2024

1. Introduction

The process of facial detection involves the meticulous identification and localization of human facial features within images. Face detection is a crucial task in the field of computer vision and has garnered significant attention due to its pivotal role in downstream applications like face recognition and reconstruction [1–4]. In recent years, face detection methods [5,6] have witnessed significant advancements in detection accuracy and speed owing to the emergence and refinement of convolutional neural networks (CNNs). Sophisticated face detection models, such as MogFace [7], RetinaFace [8], and AInnoFace [9], have demonstrated impressive performance in face detection tasks. However, complex models usually come with a huge number of parameters, which inevitably slow down the speed of detection.

The transformer model's ability to capture long-range dependencies has led to its impressive success in various natural language processing (NLP) tasks [10,11]. Recognizing the model's vast potential, scholars have begun to adapt the transformer structure for a wide range of tasks beyond NLP. For example, an innovative work by Wang et al., who introduced a unique variant of the transformer, termed 3Mformer [12]. This advanced model is designed to fuse multi-stage feature representations, significantly enhancing its performance in skeletal action recognition tasks. Similarly, Li et al. [13] leveraged the transformer to construct a strong semantic scene completion framework, demonstrating excellent experimental results. The vision transformer (ViT) was introduced into computer vision by [14], leading to exceptional outcomes for image classification tasks. ViT and its variants [15,16] demonstrate a remarkable capacity to substitute for CNN in computer vision tasks. Mehta et al. have developed a novel variant of ViT, known as MobileViT [17], which integrates a transformer block with CNNs to achieve remarkable performance in various mobile vision tasks. The lightweight design of MobileViT enables efficient processing on mobile devices, and the synergistic combination of the transformers and CNNs results in state-of-the-art(SOTA) performance. However, the aggressive downsampling system in the vision transformer model often leads to inadequate low-level feature extraction [18], leading to the loss of image representation and poor generalization of the model. Consequently, the existing models fail to detect human faces in complex environments accurately.

Generally, the conventional face detector usually consists of three parts: a CNN-based backbone, a feature pyramid network (FPN) [19], and a detector head. The backbone network is responsible for extracting the features from the input image, while FPN is employed to merge the deep semantic and shallow information from these features. Finally, the obtained features are fed into the detector head to complete the detection task, as illustrated at the top region in Figure 1. However, these existing detection solutions tend to reach their performance bottleneck, since the fixed-size receptive fields of CNN lead to the backbone being unable to effectively extract semantic information with limited model parameters.

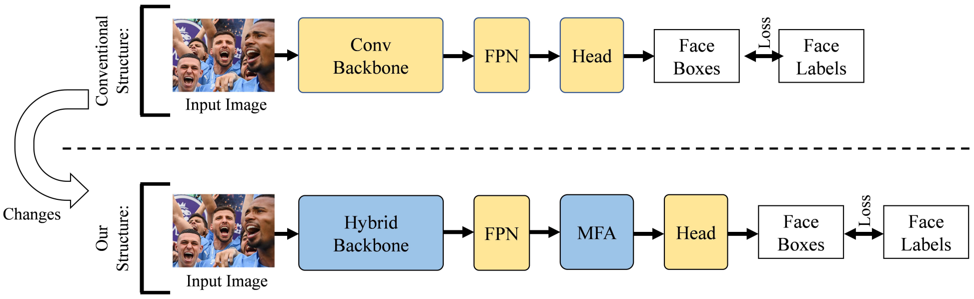


Figure 1. Pipeline overview. The main highlight of our model is the novel hybrid backbone to replace the traditional CNN-based backbone, and proposing a multiscale feature aggregation module in the neck part to improve face detection performance, with model details provided in the subsequent chapters.

To address the aforementioned challenges, we present a novel lightweight model, entitled E-CT Face, with a bi-stream architecture that aims to improve the face detection performance in a trade-off between detection accuracy and efficiency. In our method, we introduce a hybrid backbone architecture that combines CNNs and transformer blocks in a bi-stream manner. This structure enables the model to capture both global and local features while retaining rich facial texture details [20–22]. To effectively integrate feature maps from transformers and CNNs, mitigate the loss of face details during down-sampling operations, and enhance the feature extraction capability of the hybrid backbone, we present a novel convolutional block called the feature enhancement convolution (FEC) block. The proposed FEC architecture comprises a detail preservation (DP) layer and standard convolutional layers. The standard convolutional layers are employed to capture and encode local patterns within the input feature maps and merge features from the corresponding transformer block. The DP layer reconstructs the spatial dimension of the input feature maps, retaining fine-grained details while reducing the width and height of the feature map to half of its original size. This layer proves significantly beneficial for detecting blurred and small faces, ultimately improving the overall detection accuracy. Previous detectors primarily focus on spatial features of the feature map, neglecting the rich texture contained within the inter-channel features [23]. To deal with this limitation, we introduce the multiscale feature aggregation (MFA) module, positioned between the FPN module and the head, as illustrated at the bottom region in Figure 1. This innovative approach enables the aggregation of valuable interchannel features to enhance the accurate of face detection. The MFA module consists of three standard convolutional layers with different kernel sizes and a branch channel attention part [24]. By processing the input feature maps in spatial and channel dimensions, the MFA module enhances the detector's performance in detecting faces under challenging conditions such as occlusion, insufficient lighting, atypical poses, and small scales. Through standard convolution layers with different kernel sizes and branch channel attention, tje MFA module effectively distinguishes the face from background, achieving better detection performance. Our multibranch head follows the design of the RetinaFace [8]. In the following sections, we will provide detailed explanations of these components and present experimental results to demonstrate the superiority of our method.

In summary, the key contributions of this work are as follows:

1. We have introduced a novel backbone architecture for efficient face detection, which has leveraged the advantages of both CNNs and transformers and outputs multiscale features through a hybrid backbone to detect faces with scale variations.
2. The proposed FEC block employs a spatial dimension reconstruction operation and standard convolutional stacks to optimize the preservation of detailed facial textures while facilitating feature fusion between the transformer and CNN blocks.
3. By combining the standard convolution layers and a branch channel attention architecture, our proposed MFA module is able to enhances the ability of the detector to differentiate between facial features and background elements along both the spatial and channel dimensions.

2. Related Works

2.1. CNN-Based Face Detectors

In the past few years, significant advancements in CNN-based face detection methods have been noteworthy, owing to the exceptional feature extraction ability and visual induction bias. The existing CNN-based face detection methods can be categorized into two distinct groups: two-stage detectors and one-stage detectors. Two-stage detectors involve the generation of candidate regions in the initial step, followed by the classification and regression of the candidate regions in the subsequent steps. Notable examples of two-stage detectors include Face R-CNN [4], ScaleFace [25], and FDNet [26]. These two-stage detectors of face detection achieve superior performance but at the cost of slower detection speed. One-stage detectors directly classify and regress images based on the anchor without needing a separate region generation step. Prominent One-Stage methods include SSD [27], RetinaFace [8], and Faster R-CNN [28]. One-Stage detectors have emerged as the prevailing research direction for face detection algorithms due to their ability to balance performance and speed. The facial detection models, designed by [7,9,29], have demonstrated remarkable performance on public datasets. However, the high number of parameters associated with these models poses a significant challenge in deploying the detectors on edge devices. Implementing face detection tasks on devices with limited computational resources is a prevalent concern. The majority of face detection tasks are executed through devices with limited computational resources. To implement detector to edge devices, lots of great lightweight detectors have sprung up. By regarding the receptive fields as natural "anchors" which can cover continuous face scales, LFFD [30] designed a light and fast face detector. Qi et al. redesigned the detection head and loss function of the Yolov5 object detector and obtained the Yolov5 Face [31], which achieves state-of-the-art performance in the WIDER FACE dataset. EfficientFace [32] integrated three key features upon the EfficientNet to obtain a high-performance lightweight detector. The Extremely Tiny Face Detector (EXTD) designed by [33], which generates the feature maps by iteratively reusing a shared lightweight and shallow backbone network, and this approach significantly reduces the number of parameters. However, these lightweight face detectors with a small number of parameters are limited due to the fixed-size receptive fields of CNNs, limiting their effectiveness in extracting features.

2.2. Transformer-Based Vision Tasks

The transformer architecture introduced by [10] represents a seminal contribution to the field of natural language processing. By implementing the multihead self-attention and fully connected feed forward networks, this structure enables the network to extract global context in a highly effective manner, leading to significant advances in the field. The advent of ViT [14] marks a significant milestone in computer vision, as it represents the first instance of transformer architecture being applied to this field. ViT reduces model complexity and the number of parameters by partitioning images into uniform patches, which are subsequently encoded for processing via the transformer. The detection transformer (DETR) [34] designs an end-to-end target detector based on the transformer, leveraging its global representation capabilities to improve the accuracy and efficiency of the model. ViT is inherently less efficient in detection tasks due to the large model parameter size, motivating the need for lightweight variants such as Mobile ViT [17], which employs a global context extraction module designed to optimize the model's efficiency. Mehta et al.'s subsequent research [35] introduced novel techniques for computing self-attention that reduced the model complexity while improving the performance. However, the aggressive downsampling system of the vision transformer loses the image representation, leading to the model exhibiting suboptimal performance in detecting faces with complex backgrounds and small targets. To address this limitation, we have developed a novel hybrid lightweight face detector that merges the merits of transformers and CNNs in a bi-stream manner and introduces a detail-preserving layer to mitigate the loss of representative features caused by the aggressive downsampling system employed by the transformer.

3. Proposed Method

3.1. Method Overview

In this section, we have presented the overall framework of the E-CT Face detector. Subsequently, we have explicated the principal components of the model in the subsequent subsections. As shown in Figure 2, the framework of E-CT face is familiar to conventional object detectors, which consists of three parts: the hybrid backbone, neck, and head.

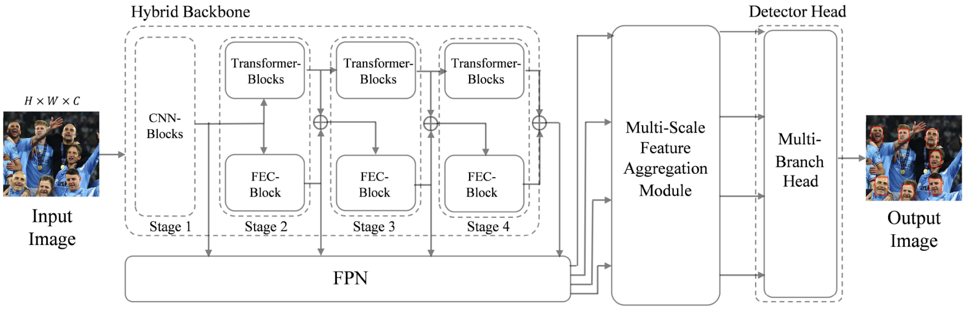


Figure 2. Overall network architecture of our proposed method. The input image is first fed into the hybrid backbone to extract transformer and CNN features; the semantic and spatial information from each stage is integrated by employing the FPN. After that, the feature maps are systematically fed into the multiscale feature aggregation module to achieve feature fusion. Finally, the aggregated feature maps go through the multibranch head to output the face detection result.

We have designed the backbone network of the detector with a bi-stream configuration by integrating the CNNs and transformer blocks in a hybrid manner, in which the FEC block is proposed to effectively integrate the feature maps derived from transformer and CNN blocks. For the neck part of the detector, a conventional FPN structure is adopted to merge the deep semantic and shallow information from the input features. The subsequent MFA module employs the standard convolutional layers with attention mechanism and multiple kernel sizes to strengthen the capability of the model to detect faces in complex environments. Finally, the feature map is sent to the head part for classification and localization.

3.2. Hybrid Backbone

In this section, we have presented a detailed exposition of the backbone component of our hybrid architecture and elucidate the benefits of our innovative backbone, which combines the merits of the transformer block and the CNN block for face-detection tasks. Transformer-based models have advantages on learning global representations but are weakly to capture local patterns. The lightweight CNN models are limited due to the fixed-size receptive field and lack of global modeling, but they have good ability to learn local representations. To make our backbone be aware of both global and local features and good model generalization capability, we have proposed the bi-stream backbone architecture, one steam is the transformer pipeline composed of transformer blocks, in which the MobileVitV2 block is adopted to extract the global context of the input image. The other one is the CNN pipeline composed of FEC blocks, in which consist of standard convolutional layers and a DP layer. A comparison of our hybrid backbone with the conventional backbone is illustrated in Figure 3.

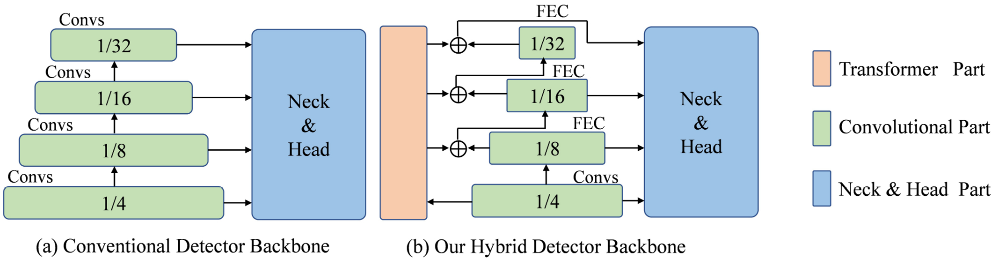


Figure 3. The comparison between CNN-based backbone and our proposed hybrid backbone. (**a**) depicts the backbone of conventional detector, primarily consisting of a hierarchical CNN structure. (**b**) depicts our purposed hybrid backbone. Compared to conventional detector backbone, our hybrid detector backbone incorporates a transformer part for global modeling, while maintaining the hierarchical structure to retain multiscale features.

At first, the input image is fed into stage 1 of the backbone by using the traditional CNN block to reduce the spatial resolution of the feature map to 1/4 and increases the dimension of channels to 96. Then, we feed the output feature map of stage 1 into the subsequent bi-stream pipelines. The transformer block extracts global context and implicitly learn the semantic information within the face, body, and background [36]; the FEC block extracts the local face features while adding up the feature maps form the transformer block and the FEC block before feeding them to the subsequent layers to fusion features. As a result, we can obtain multilevel feature maps through four stages These output feature maps are delivered to the subsequent neck part, and the resolution 4 stages are 1/4, 1/8, 1/16 and 1/32 of the input image in our method, respectively. The embedding dimensions of each stage are 96, 192, 288, and 384, and the dimensions of the transformer block and FEC block in Hybrid Stages are 96, 144, and 192, respectively. The hierarchical structure ensures that the backbone extracts face features from coarse to fine and sparse to dense. Hence, our proposed method can precisely capture the semantic information of face images and flexibly achieve face detection tasks in various scenarios.

3.3. Feature Enhancement Convolution Block

The transformer block captures complex relational interactions between different spatial patches more easily, whereas the low-level features containing images details are inadequate in the vision transformer's aggressive down-sampling system. To overcome this limitation, we have designed the FEC block as a compensation for the transformer, which consists of standard convolutional layers and a DP layer, as shown in Figure 4. The DP layer retains the low-level features containing images details by downsampling the input feature maps through reconstructing the spatial dimensions of the input feature maps. The FEC block in stage 2 takes the output feature maps from stage 1 as input. Similarly, the input of the FEC block in stage $P + 1$ $(P = 2, 3)$ is formed by adding up the output feature maps of the FEC block and the Transformer block from stage P. Finally, the output feature maps from the FEC block in stage 4 are fed into the neck part of the entire network.

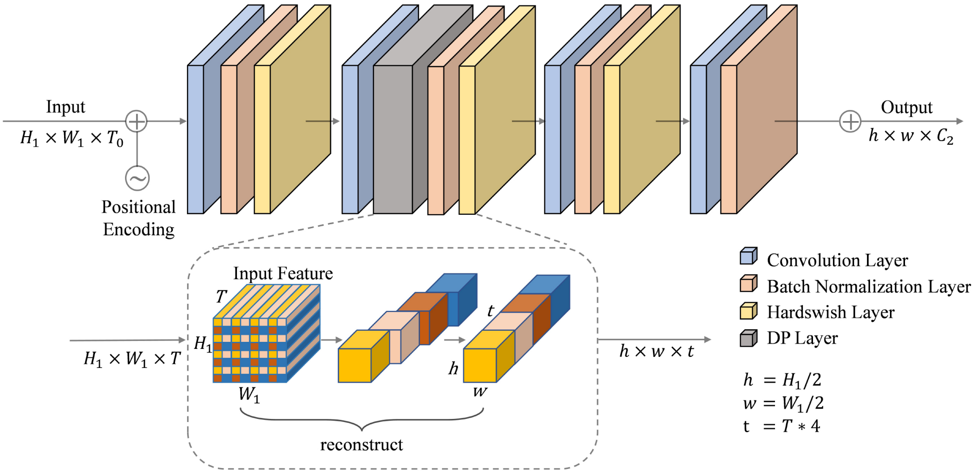


Figure 4. The architecture of the proposed feature enhancement convolution (FEC) block. The FEC block comprises several key components: a positional encoding layer, convolutional layers, batch normalization layers, activation function layers, and a DP layer. We have listed the pipeline of FEC block at the top, and we have demonstrated the DP layer structure at the bottom.

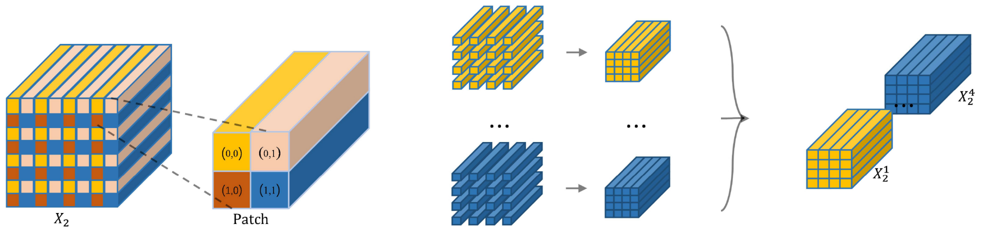
At the stage P, the FEC block takes an input tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times C_1}$ and applies pointwise convolution layer to generate $\mathbf{X}_1 \in \mathbb{R}^{H \times W \times T}$, followed by a standard convolution layer with kernel size 3×3 and stride of 1 to generate $\mathbf{X}_2 \in \mathbb{R}^{H \times W \times T}$. The pointwise convolution layer projects the tensor into a T-dimensional space by learning a linear combination of the input channels, T is the dimension of the transformer block from the corresponding stage P. The convolutional layer with kernel size of 3×3 encodes the local spatial features of the face. To make the model preserve the image details when learning local features of tiny faces, we feed $\mathbf{X}_2$ into the DP layer in order to reconstruct it into four tensors of the same size, as shown in Figure 5. More specifically, the output of the DP layer is computed as

$$\mathbf{X}_3 = Concat(\mathbf{X}_2^1, \mathbf{X}_2^2, \mathbf{X}_2^3, \mathbf{X}_2^4) \tag{1}$$

$$\mathbf{X}_2^i = R(\mathbf{P}_1(n,m), \cdots, \mathbf{P}_k(n,m), \cdots, \mathbf{P}_K(n,m),) \tag{2}$$

where $Concat(\bullet)$ represents the tensor using the concatenation operation in the channel dimension. At first, the DP layer splits $\mathbf{X}_2$ into $K = H/2 * W/2$ patches. The dimension of each patch is $\mathbb{R}^{2 \times 2 \times T}$; $\mathbf{P}_k$ denotes the kth patch In Equation (2). Here, $R(\bullet)$ represents the reorganization of the tensor at (n, m) of each patch into a new tensor $\mathbf{X}_2^i \in \mathbb{R}^{h \times w \times T}$, $i \in \{1, 2, 3, 4\}$, where the correspondence among i , m, and n can be described as

$$n \cdot 2^0 + m \cdot 2^1 = i - 1 \qquad n, m \in \{0, 1\} \tag{3}$$



(1) Split into patches (2) Reorganize into a new tensor (3) Concat

Figure 5. The main operations of detail preservation (DP) layer include the following: (**1**) Split into patches: We split the input tensor $\mathbf{X}_2$ into $H/2 * W/2$ patches of uniform size. Within each of these patches, there are four tensors with dimensions of $\mathbb{R}^{1\times1\times T}$, which are designated with spatial indices of (0,0), (0,1), (1,0), and (1,1), respectively. (**2**) Reorganization into a new tensor: We reorganized the tensors located at corresponding index positions within each patch to construct a new tensor. (**3**) Concat: We concatenated the four reorganized tensors together along the channel dimension.

The proposed DP layer effectively reduces the loss of low-level features by transferring the information of adjacent pixels from the spatial dimension to the channel dimension. After that, we apply a standard convolutional layer with kernel size 3 × 3 to merge these concatenated features in $\mathbf{X}_3 \in \mathbb{R}^{h\times w\times t}$, in which $t = T * 4$. Then, we use a pointwise convolution layer to project the tensor into the low-dimensional space to obtain $\mathbf{X}_4 \in \mathbb{R}^{h\times w\times C_2}$. In mathematics, the FEC block can be described as

$$\mathbf{X}_3 = DP(Conv_{3\times3}(Conv_{1\times1}(\mathbf{X} + \mathbb{P}))) \tag{4}$$

$$\mathbf{X}_4 = Conv_{1\times1}(Conv_{3\times3}(\mathbf{X}_3)) \tag{5}$$

Among them, the stride of $Conv_{3\times3}$ and $Conv_{1\times1}$ are both 1, and $\mathbb{P}$ denotes the position encoding (PE). For position-sensitive tasks such as face detection, the preservation of spatial structure is crucial. The PE layer can effectively capture the position information of the input feature and ensure the model's sensitivity to the spatial position of the input feature maps.

3.4. Multiscale Feature Aggregation Module

One of the major challenges for face detection task is to detect faces in the crowd. Since high-resolution feature maps can improve the sensibility of the detector for small face detection, we have developed a top-down connection method for four multilevel feature maps of different scales in the neck part of the E-CT face detector to obtain additional semantic information with high-resolution feature maps to achieve the purpose of detecting faces at different scales. In the real application scenarios, problems such as occlusion, illumination, face pose variation, and face scale variation frequently occur. To overcome this problem, the MFA module is proposed in this study, the structure of which is shown in Figure 6. The MFA module processes the input feature maps separately to obtain four feature maps with different scales and then delivers the four feature maps hierarchically to the head part for classification and localization. For a single feature map of the input, the specific structure of MFA can be defined as

$$D_S(F_M) = Conv_{3\times3}(F_M), Conv_{5\times5}(F_M), Conv_{7\times7}(F_M) \tag{6}$$

$$\mathbf{Res}_1, \mathbf{Res}_2, \mathbf{Res}_3 = D_S(F_M) \tag{7}$$

$$\mathbf{Res} = Concat(\mathbf{Res}_1, \mathbf{Res}_2, \mathbf{Res}_3) \tag{8}$$

$$\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3 = Swish\Big(Conv_{1\times1}\Big(Pooling(\mathbf{Res}_1, \mathbf{Res}_2, \mathbf{Res}_3)\Big)\Big) \tag{9}$$

$$O_M = \mathbf{Res} + \mathbf{Res} * \Big(Sigmoid\Big(Conv_{1\times1}\Big(Concat(\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3)\Big)\Big)\Big) \tag{10}$$

where O_M and F_M represent the single input and output of MFA. The $Pooling(\bullet)$ operation contains two methods, the Maxpool method and the Avgpool method. We used both average pooling and maximum pooling to obtain the model with optimum performance. We applied maximum pooling to the feature maps after convolutional layers with kernel size 7 × 7 to preserve the contour information of the big face and applied average pooling to the feature maps after convolutional layers with kernel size 5 × 5 and convolutional layers with kernel size 3 × 3 to preserve the overall features of the small face. In Section 4.3, we demonstrate the ablation experiments on pooling methods.

The input feature maps of MFA are four feature maps at different scales, which is the output of the FPN network. Each feature map is processed separately by standard convolutions with three different kernel sizes for multiscale feature extraction, allowing the model to better distinguish faces from other backgrounds on the spatial dimension. The feature maps from each of our three standard convolutional outputs are fed into the branch channel attention part to enable our method to focus on channels that are more relevant for face details [37].

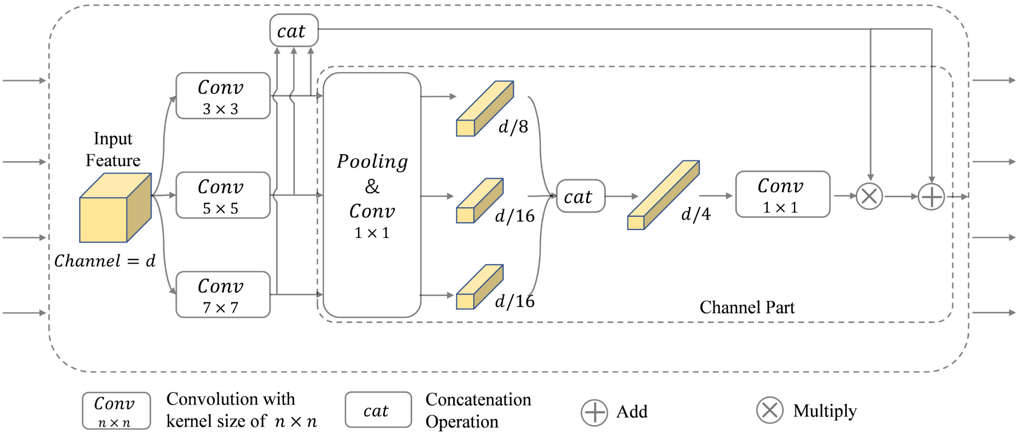


Figure 6. The structure of the multiscale feature aggregation (MFA) module. MFA module comprises convolutional layers with varying kernel sizes and a branched channel attention component.

4. Experimental Results

4.1. Datasets and Evaluation Metrics

We have evaluated our approach on two publicly available datasets: WIDER FACE [38], FDDB [39]. The WIDER FACE dataset contains 32,203 images and 393,703 annotated faces for face detection and recognition. Since the WIDER FACE dataset contains various scenes, lighting conditions, and variations of face pose, it is widely applied in face detection and recognition. The Face Detection Dataset and Benchmark (FDDB) contains 5171 faces with various challenges and is one of the most significant datasets in face detection. FDDB have two different evaluation methods: discrete score and continuous score. We chose discrete score as our evaluation method to avoid the labeling style of the training dataset affecting the test results since we only use FDDB as a test dataset.

For a fair comparison with the SOTA methods, we also divide the WIDER FACE dataset into three subsets: train (40%), validation (10%), and test (50%). The validation subset and test subset are divided into three subsets according to the difficulty level of the detection, which are easy, medium and hard subsets. The training subset of WIDER FACE was used for training, the validation subset and test subset of WIDER FACE were used for performance evaluation, and we also evaluated our method on the FDDB dataset. All of the training images were resized to 640 × 640, and the testing images were kept at their original size, which means that no rescaling was used. Our evaluation metric on the easy, medium, and hard subset was the average precision (AP) of $IoU = 0.5$.

4.2. Implementation Details

Our experiments were performed on a workstation with two NVIDIA GeForce RTX 2080Ti GPUs (with 24 G memory) and 64 G RAM. Our network was implemented in Python 3.7 with Pytorch 1.8.1. We adopt the AdamW optimizer [40] for parameter optimization. We set the batch size to 8, and we trained our model for 70 epochs. The learning rate was set to 0.001 at the first 2500 iterations as the warm-up stage and then it annealed to 0.000001. In addition, our hybrid backbone was pretrained using ImageNet-1k datasets.

4.3. Component Evaluation

In this section, to confirm the effectiveness of the major components in our method. We have conducted the ablation experiments to quantitatively verify the performance of the hybrid backbone, FEC, and MFA of our method.

4.3.1. Ablation Study on the Hybrid Backbone

Our hybrid backbone can merge the advantages of the transformer block and FEC block to extract multiscale global contextual information of human faces effectively. We have further conducted several experiments to demonstrate the superiority of our hybrid backbone. We replace the hybrid backbone with two leading lightweight backbone networks, MobileNet V3 and EfficientNet, which have been widely used in lightweight face detection models. Therefore, we compare the performance of using our hybrid backbone, MobileNet V3, and EfficientNet-B0 as the backbone individually on the WIDER FACE validation subset. As shown in Table 1, our face detector benefits from the strong information extraction ability based on the hybrid backbone and achieves the best results on the easy, medium, and hard subsets.

Table 1. Quantitative comparison results of different backbones on the WIDER FACE validation subset.

Backbones	**Easy**	**Medium**	**Hard**	**Params (M)**
MobileNet V3 [3]	93.75%	91.48%	81.29%	3.67
EfficientNet-B0 [41]	94.27%	92.59%	83.82%	4.77
Ours	95.30%	94.20%	87.56%	3.80

4.3.2. Ablation Study on the FFC

The FEC block has two key points: one is that it extracts image details through reconstructing operations in the spatial dimension to improve the performance of detecting small faces, and the other one is encoding the local patterns and merging the features from the corresponding transformer block. To validate the superiority of our FEC block, we have conducted several experiments, with the quantitative results presented in Table 2. Similarly, we have compared the performance of three models on the WIDER FACE validation dataset. The "Baseline" in the first row of Table 2, denotes the results of the E-CT face without FEC block, the "Baseline + FEC (w Conv)" denotes the results of the FEC block with the standard convolution with stride 2, and the "Baseline + FEC (w DP)" denotes the results of the complete FEC block.

Table 2. Quantitative comparison results of the FEC component on the WIDER FACE validation subset.

Model	**Easy**	**Medium**	**Hard**
Baseline	93.78%	92.70%	86.87%
Baseline + FEC (w Conv)	94.13%	93.09%	87.03%
Baseline + FEC (w DP)	95.30%	94.20%	87.56%

The results in Table 2 show that our FEC block has a significant impact in improving the overall performance of the model, achieving 1.52%, 1.398%, and 0.69% improvement in the easy, medium, and hard subsets, respectively, compared to the baseline model. The decreased performance of the model without the DP layer also proves the effectiveness in spatially reconstructing features.

4.3.3. Ablation Study on the MFA

We have also analyzed the performance of the model with/without MFA to verify the effectiveness of MFA module. The quantitative results of our experiments are presented in Table 3. As for the visual comparison shown in Figure 7, it can be seen that the images obtained from the model without MFA have more missing faces compared with those from model with MFA in detecting faces with occlusion, different poses, scales, and illumination. The results demonstrated that MFA effectively improves the detector's performance in these critical aspects.

Table 3. Quantitative results with/without the MFA on the WIDER FACE validation subset.

Dataset	w/o MFA	w MFA
Easy	93.70%	95.30%
Medium	92.69%	94.20%
Hard	84.82%	87.56%

Simultaneously, to validate the effectiveness of the pooling method in the MFA module, we have conducted ablation experiments including "E-CT face + Avgpool", "E-CT face + Maxpool", and "E-CT face + Avgpool & Maxpool". The quantitative results are presented in Table 4, in which "E-CT face + Avgpool" and "E-CT face + Max pool" denote the results of the E-CT face with only the average pooling method and maximum pooling method, respectively, and "E-CT face + Avgpool & Maxpool" denotes the results of the E-CT face with both average and maximum pooling methods. According to the results in Table 4, the detection performance of the model with only the average pooling method on the hard subset is superior to the model with only the maximum pooling method, and the detection performance of the model with only the maximum pooling method on the easy and medium subset is superior to the model with only average pooling method. In contrast, the model with both maximum and average pooling methods has the best performance in all subsets. The experimental results also show the importance of the appropriate combination of pooling methods in MFA module for the improvement of model performance.

Table 4. Quantitative comparison results with different pooling methods of MFA on the WIDER FACE validation subset.

Model	Easy	Medium	Hard
E-CT Face + Avgpool	94.77%	93.87%	87.28%
E-CT Face + Maxpool	95.14%	93.96%	86.65%
E-CT Face + Avgpool & Maxpool	95.30%	94.20%	87.56%

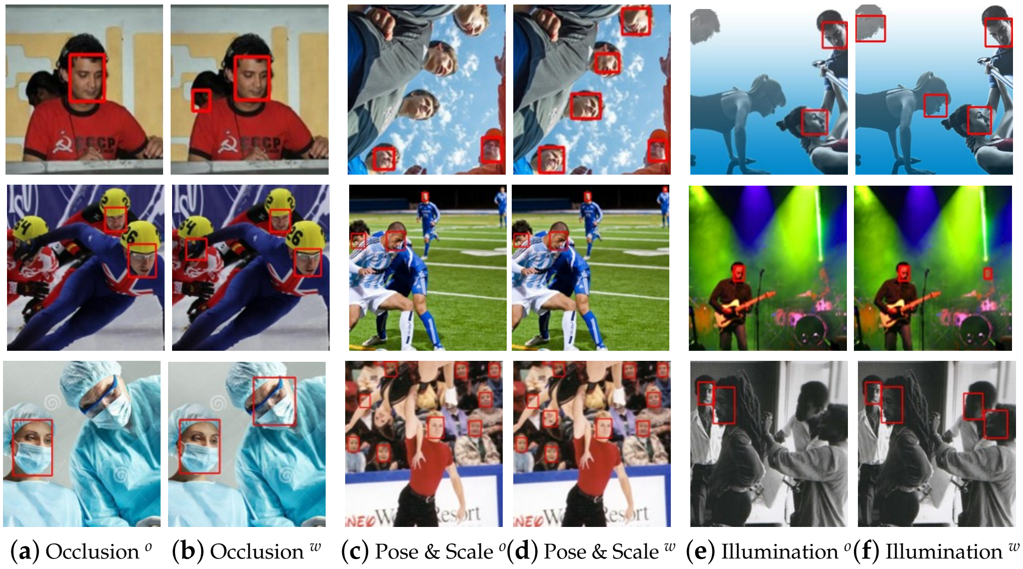

(**a**) Occlusion ^{o}
(**b**) Occlusion ^{w}
(**c**) Pose & Scale ^{o}
(**d**) Pose & Scale ^{w}
(**e**) Illumination ^{o}
(**f**) Illumination ^{w}

Figure 7. Visual results of with/without the MFA. $(\bullet)^{o}$ represents without MFA, $(\bullet)^{w}$ represents with MFA. Red box represents the detection result.

4.4. Comparison with the SOTA Methods

We have compared our proposed method with state-of-the-art (STOA) lightweight face detection methods and heavyweight detection methods on the WIDER FACE dataset to visualize the superiority of our proposed method. We have strictly followed the standard evaluation protocols for the WIDER FACE dataset, by training the model only on the training subset and testing it on the validation and test subsets.

The results of the quantitative comparison with lightweight face detectors on the WIDE FACE validation subset are present in Table 5, and the results of the quantitative comparison with heavyweight face detectors are present in Table 6.

Table 5. Quantitative comparison results between our face detector and other lightweight face detectors on the WIDER FACE validation subset.

Light Detector	**Easy**	**Medium**	**Hard**	**Params (M)**
YoloV5 Face-n [31]	93.6%	91.5%	80.5%	1.72
YoloV5 Face-s [31]	94.3%	92.6%	83.1%	7.06
EXTD [33]	92.1%	91.1%	85.6%	0.16
LFFD [30]	91.0%	88.1%	78.0%	2.15
OS-LFFD [42]	91.6%	88.4%	77.1%	1.44
Efficient Face-B0 [32]	91.0%	89.1%	83.6%	3.94
Efficient Face-B1 [32]	91.9%	90.2%	85.1%	6.64
Efficient Face-B2 [32]	92.5%	91.0%	86.3%	7.98
SCRFD-10GF [43]	95.1%	93.8%	83.0%	3.86
IRNet [44]	91.8%	89.3%	76.6%	1.68
Ours	**95.30**%	94.20%	87.56%	3.80

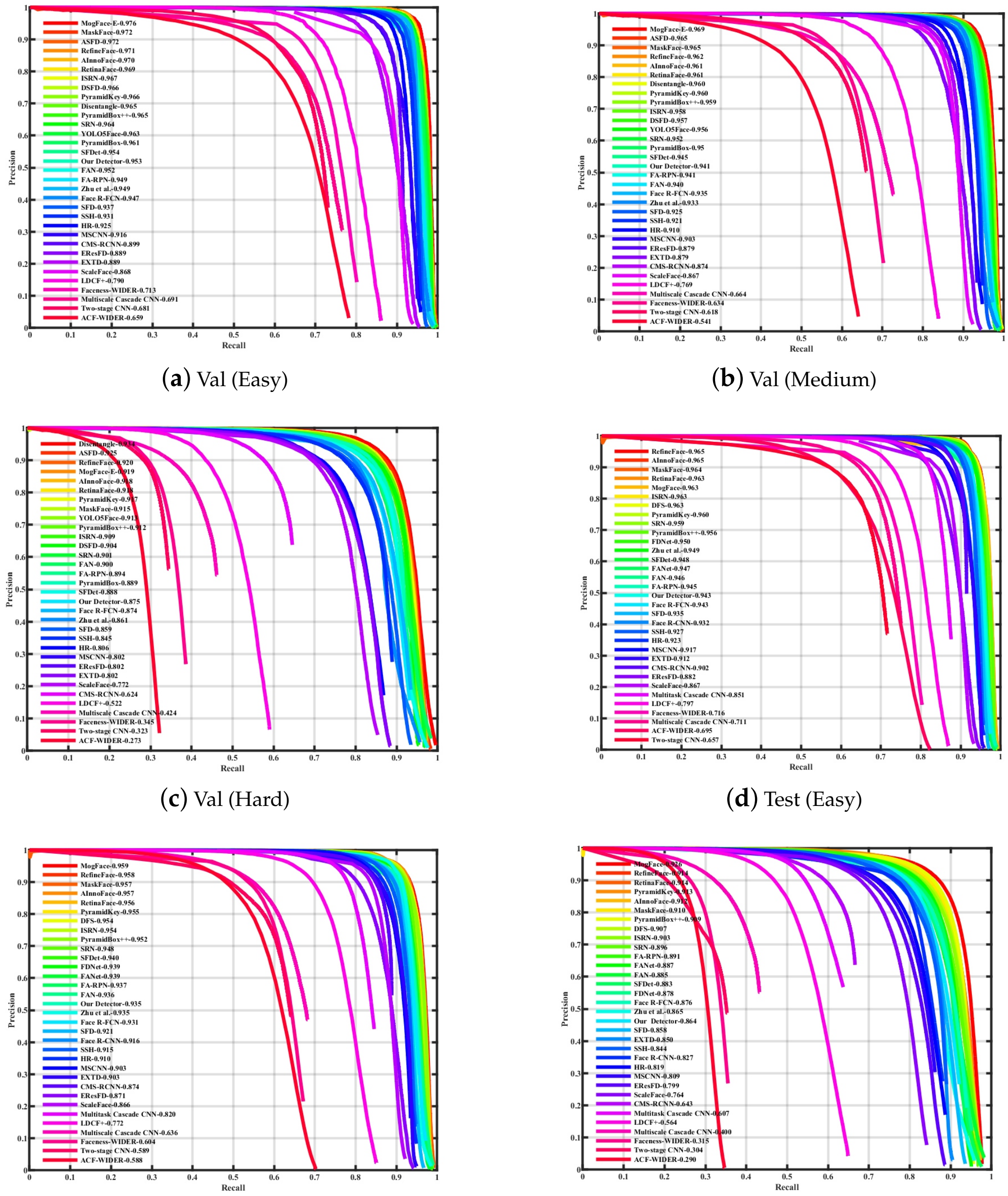
Table 6. Quantitative comparison results between our face detector and other heavyweight face detectors on the WIDER FACE validation subset.

Heavy Detector	Easy	Medium	Hard	Params (M)
AInnoFace [9]	97.0%	96.1%	91.8%	88.01
MogFaceAli-AMS [7]	94.6%	93.6%	87.3%	36.07
MogFace [7]	97.0%	94.3%	93.0%	85.26
TinaFace [45]	95.6%	94.2%	81.4%	172.95
YoloV5 Face-X6 [31]	96.67%	95.08%	86.55%	88.665
Ours	95.30%	94.20%	87.56%	3.8

The current STOA lightweight face detection methods have demonstrated impressive performance on the WIDER FACE validation set, achieving average precision (AP) ranging from 91.0% to 95.16% on the easy subset, 88.1% to 93.87% on the medium subset, and 76.6% to 86.3% on the challenging hard subset. Among these methods, excluding DS-Face, SCRFD-10GF stands out, utilizing 3.86 M parameters to achieve the highest AP of 95.1% and 93.8% on the easy and medium subsets, respectively. Meanwhile, EfficienetFace-B2, with 7.98 M parameters, leads the pack on the hard subset with an AP of 86.3%. Our proposed DS-Face detector achieves remarkable performance on the WIDER FACE validation set, attaining optimal AP scores of 95.30%, 94.20%, and 87.56% on the easy, medium, and hard subsets, respectively, with a mere 3.8 M parameters. Specifically, DS-Face exhibits the best performance among lightweight face detection methods with similar parameter counts. Notably, on the challenging hard subset of the WIDER FACE validation set, DS-Face outperforms EfficienetFace-B2 by achieving a 1.26% improvement in performance while using less than half the number of parameters. This significant enhancement is primarily attributed to the dual-stream architecture employed by DS-Face, which enables the detector to perform exceptionally well in more complex detection environments.

The existing STOA heavyweight face detection methods achieve AP ranging from 94.6% to 97.0% on the Easy subset, 93.6% to 96.1% on the Medium subset, and 81.4% to 91.8% on the Hard subset of the WIDER FACE validation set. Among these heavyweight face detection methods, MogFace utilizes 22 times more parameters (85.26 M) than DS-Face, yet it only achieves a marginal improvement of 1.7% and 5.44% in AP on the easy and hard subsets, respectively. Similarly, AInnoFace, with 23 times more parameters (88.01 M) than DS-Face, attains just a 1.902% increase in AP on the medium subset. Remarkably, DS-Face uses only 1/50th of the parameters compared to TinaFace, yet it matches TinaFace's performance closely, with only a 0.3% and 0.002% decrease in AP on the easy and medium subsets, respectively. Moreover, DS-Face outperforms TinaFace by 6.16% on the challenging hard subset. Overall, DS-Face demonstrates competitiveness even when compared to heavyweight face detection methods, highlighting the efficiency of the DS-Face detector and the effectiveness of the design proposed in this paper.

Figure 8 provides the precision–recall curve of our model and the existing model in the WIDER FACE validation subset and test subset. As depicted in the figure, the proposed DS-Face face detection method achieves an average precision (AP) of 94.30%, 93.5%, and 86.4% on the easy, medium, and hard subsets of the WIDER FACE test set, respectively. Compared to the results obtained on the validation set, these values are slightly lower by 1%, 0.698%, and 1.16%, respectively. This minor decrease in performance indicates that the DS-Face face detection method possesses good generalization capabilities, suggesting its robustness and adaptability to different images. We have used the FDDB dataset for testing without any modification to better demonstrate the competitiveness of our model, and the results of the receiver operating characteristic curve (ROC) are shown in Figure 9.



(**a**) Val (Easy) (**b**) Val (Medium)

(**c**) Val (Hard) (**d**) Test (Easy)

(**e**) Test (Medium) (**f**) Test (Hard)

Figure 8. Precision–recall curves of different methods on the WIDE FACE validation and test subset.

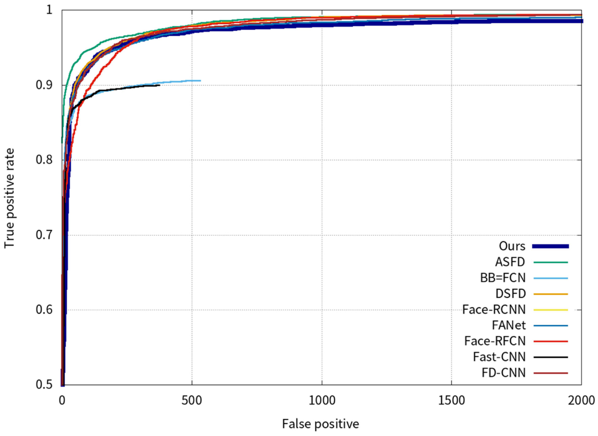


Figure 9. ROC curves of our model's detection results on the FDDB dataset.

4.5. Running Efficiency

We have analyzed the inference speed of our method under different input sizes and on different platforms. Our inference speed is measured in "ms". We measured the inference time of our method on each of the three platforms, NVIDIA GeForce 2080 Ti and Intel i9-10940X). Also, we set the size of the input images as Small $(1 \times 3 \times 320 \times 320)$, Mid $(1 \times 3 \times 640 \times 640)$, and Big $(1 \times 3 \times 960 \times 960)$. The final results are presented in Table 7. The results show that our model is able to detect faces in realtime on PCs, and also it can detect faces at 35ms latency in small images on CPU.

Table 7. Quantitative comparison results of the inference efficiency among different devices on different input sizes $(320 \times 320, 640 \times 640, 960 \times 960)$. The CPU is Intel-10940X.

Platforms	320 × 320	640 × 640	960 × 960
NVIDIA GeForce 2080Ti	9 ms (111 FPS)	13 ms (76 FPS)	28 ms (35 FPS)
CPU	35 ms (28 FPS)	114 ms (8 FPS)	---- ----

5. Limitation and Future Work

Despite achieving a competitive performance on the WIDER FACE dataset with a model utilizing only 3.8 M parameters, our approach has not yet attained optimal performance when compared to heavyweight face detectors. In our future work, we intend to investigate heavyweight face detection models with a larger parameter count. This exploration aims to unlock the full potential of our network architecture and enhance its adaptability to diverse detection scenarios.

6. Conclusions

This paper has proposed a novel efficient face detector, termed E-CT Face. We have followed the widely used face detection structure as our baseline network. One of its main limitations is its inability to extract features from images efficiently, which may impact its detection performance in a complex environment. To handle this limitation, we have adopted the hybrid backbone that leverages the advantages of both transformer and CNN architectures, enabling our detector to achieve high performance with little parameter size in face object detection tasks. Moreover, to enhance the effectiveness of our model in detecting diminutive faces and optimizing its detection capabilities in various complex scenarios, we have devised the feature enhancement convolution block and multiscale feature aggregation module. Our method demonstrates superior performance compared to the SOTA lightweight face detection methods on the face detection benchmarks. We anticipate that our works will offer valuable insights for future research on the application of transformers in the field of face detection.

Author Contributions: Conceptualization, Z.Z. and S.W.; methodology, Z.Z. and Q.C.; software, Z.Z.; validation, Q.C., S.W. and Z.Z.; formal analysis, Z.Z. and Q.C.; investigation, S.W.; resources, T.Y.; data curation, S.W.; writing—original draft preparation, Z.Z.; writing—review and editing, T.Y.; visualization, T.Y.; supervision, T.Y.; project administration, T.Y.; funding acquisition, T.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

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

Notations

In this manuscript, we employ specific mathematical notation to ensure clarity and consistency. The following conventions are used throughout: Notably, $\mathbb{R}^{H\times W\times C}$ signifies the dimensionality of tensors. Tensors are denoted by uppercase blodface, e.g., $\mathbf{X}$. The superscripts and subscripts of $\mathbf{X}$ denote its indices. Therefore, the dimensionality of the tensor $\mathbf{X}_2^1$ can be expressed as $\mathbf{X}_2^1 \in \mathbb{R}^{H\times W\times C}$.

References

1. Zhang, S.; Zhu, R.; Wang, X.; Shi, H.; Fu, T.; Wang, S.; Mei, T.; Li, S. Improved selective refinement network for face detection. *arXiv* **2019**, arXiv:1901.06651.
2. Kuzdeuov, A.; Koishigarina, D.; Varol, H.A. Anyface: A data-centric approach for input-agnostic face detection. In Proceedings of the 2023 IEEE International Conference on Big Data and Smart Computing(BigComp), Jeju, Republic of Korea, 13–16 February 2023; pp. 211–218.
3. Howard, A.G.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
4. Wang, H.; Li, Z.; Ji, X.; Wang, Y. Face r-cnn. *arXiv* **2017**, arXiv:1706.01061.
5. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask r-cnn. *arXiv* **2017**, arXiv:1703.06870.
6. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
7. Liu, Y.; Wang, F.; Sun, B.; Li, H. Mogface: Towards a deeper appreciation on face detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4083–4092.
8. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. Retinaface: Single-shot multi-level face localisation in the wild. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5202–5211.
9. Zhang, F.; Fan, X.; Ai, G.; Song, J.; Qin, Y.; Wu, J. Accurate face detection for high performance. *arXiv* **2019**, arXiv:1905.01585.
10. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017.
11. Wang, S.; Li, B.Z.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-attention with linear complexity. *arXiv* **2020**, arXiv:2006.04768.
12. Wang, L.; Koniusz, P. 3mformer: Multi-order multi-mode transformer for skeletal action recognition. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 5620–5631.
13. Li, Y.; Yu, Z.; Choy, C.B.; Xiao, C.; Álvarez, J.M.; Fidler, S.; Feng, C.; Anandkumar, A. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 9087–9098.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Shen, C. Conditional positional encodings for vision transformers. In Proceedings of the International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
16. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision(ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
17. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
18. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *arXiv* **2021**, arXiv:2106.04803.
19. Lin, T.-Y.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
20. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 805–815.
21. Zhang, H.; Hu, W.; Wang, X. Parc-net: Position aware circular convolution with merits from convnets and transformer. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022.
22. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. In Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Online, 7–10 December 2021.
23. Liu, J.; Li, H.; Kong, W. Multi-level learning counting via pyramid vision transformer and cnn. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106184. [CrossRef]
24. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.-S. Cbam: Convolutional block attention module. *arXiv* **2018**, arXiv:1807.06521.

25. Yang, S.; Xiong, Y.; Loy, C.C.; Tang, X. Face detection through scale-friendly deep convolutional networks. *arXiv* **2017**, arXiv:1706.02863.
26. Zhang, C.; Xu, X.; Tu, D. Face detection using improved faster rcnn. *arXiv* **2018**, arXiv:1802.02142.
27. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
28. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef] [PubMed]
29. Li, J.; Wang, Y.; Wang, C.; Tai, Y.; Qian, J.; Yang, J.; Wang, C.; Li, J.; Huang, F. Dsfd: Dual shot face detector. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5055–5064.
30. He, Y.; Xu, D.; Wu, L.; Jian, M.; Xiang, S.; Pan, C. Lffd: A light and fast face detector for edge devices. *arXiv* **2019**, arXiv:1904.10633.
31. Qi, D.; Tan, W.; Yao, Q.; Liu, J. Yolo5face: Why reinventing a face detector. In *Computer Vision–ECCV 2022 Workshops. ECCV 2022*; Springer: Cham, Switzerland, 2021.
32. Wang, G.Q.; Li, J.Y.; Wu, Z.; Xu, J.; Shen, J.; Yang, W. Efficientface: An efficient deep network with feature enhancement for accurate face detection. *arXiv* **2023**, arXiv:2302.11816.
33. Yoo, Y.J.; Han, D.; Yun, S. Extd: Extremely tiny face detector via iterative filter reuse. *arXiv* **2019**, arXiv:1906.06579.
34. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. *arXiv* **2020**, arXiv:2005.12872.
35. Mehta, S.; Rastegari, M. Separable self-attention for mobile vision transformers. *arXiv* **2022**, arXiv:2206.02680.
36. Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. Relation networks for object detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3588–3597.
37. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Zhang, Z.-L.; Lin, H.; Sun, Y.; He, T.; Mueller, J.W.; Manmatha, R.; et al. Resnest: Split-attention networks. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 2735–2745.
38. Yang, S.; Luo, P.; Loy, C.C.; Tang, X. Wider face: A face detection benchmark. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5525–5533.
39. Jain, V.; Learned-Miller, E.G. *Fddb: A Benchmark for Face Detection in Unconstrained Settings*; UMass Amherst: Amherst, MA, USA, 2010.
40. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
41. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
42. Xu, D.; Wu, L.; He, Y.; Zhao, Q.; Jian, M.; Yan, J.; Zhao, L. Os-lffd: A light and fast face detector with ommateum structure. *Multimed. Tools Appl.* **2020**, *80*, 34153–34172. [CrossRef]
43. Guo, J.; Deng, J.; Lattas, A.; Zafeiriou, S. Sample and computation redistribution for efficient face detection. *arXiv* **2021**, arXiv:2105.04714.
44. Jiang, C.; Ma, H.; Li, L. Irnet: An improved retinanet model for face detection. In Proceedings of the 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China, 26–28 July 2022; pp. 129–134.
45. Zhu, Y.; Cai, H.; Zhang, S.; Wang, C.; Xiong, Y. Tinaface: Strong but simple baseline for face detection. *arXiv* **2020**, arXiv:2011.13183.

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