applied sciences

MDPI

*Article*

# Face Keypoint Detection Method Based on Blaze_ghost Network

**Ning Yu [1,†], Yongping Tian [2,*,†], Xiaochuan Zhang [2] and Xiaofeng Yin [2]**

[1] School of Vehicle Engineering, Chongqing University of Technology, Chongqing 400054, China
[2] School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China
* Correspondence: tianyp@stu.cqut.edu.cn
† These authors contributed equally to this work.

**Abstract:** The accuracy and speed of facial keypoint detection are crucial factors for effectively extracting fatigue features, such as eye blinking and yawning. This paper focuses on the improvement and optimization of facial keypoint detection algorithms, presenting a facial keypoint detection method based on the Blaze_ghost network and providing more reliable support for facial fatigue analysis. Firstly, the Blaze_ghost network is designed as the backbone network with a deeper structure and more parameters to better capture facial detail features, improving the accuracy of keypoint localization. Secondly, HuberWingloss is designed as the loss function to further reduce the training difficulty of the model and enhance its generalization ability. Compared to traditional loss functions, HuberWingloss can reduce the interference of outliers (such as noise and occlusion) in model training, improve the model's robustness to complex situations, and further enhance the accuracy of keypoint detection. Experimental results show that the proposed method achieves significant improvements in both the NME (Normal Mean Error) and FR (Failure Rate) evaluation metrics. Compared to traditional methods, the proposed model demonstrates a considerable improvement in keypoint localization accuracy while still maintaining high detection efficiency.

**Keywords:** face keypoint detection; PFLD; Blaze_ghost; HuberWingloss



**Citation:** Yu, N.; Tian, Y.; Zhang, X.; Yin, X. Face Keypoint Detection Method Based on Blaze_ghost Network. *Appl. Sci.* **2023**, *13*, 10385. https://doi.org/10.3390/app131810385

Academic Editors: Xin Ning, Weijun Li and Sahraoui Dhelim

Received: 13 July 2023
Revised: 7 September 2023
Accepted: 11 September 2023
Published: 17 September 2023



## 1. Introduction

Since AlexNet kickstarted the era of deep learning in 2012, using convolutional neural networks to detect facial features has become straightforward. Simultaneously, the mainstream approach to identifying a driver's fatigue state involves combining facial keypoint detection with facial fatigue evaluation indicators. However, recognizing facial fatigue indicators imposes strict requirements on the accuracy and precision of facial keypoint detection. Therefore, it is necessary to investigate existing facial keypoint detection algorithms and devise a high-precision and real-time keypoint detection model that accurately localizes keypoints to address the challenges in fatigue driving identification. Existing methods for facial keypoint detection can be classified into three categories: those based on ASM and AAM models, those based on cascaded shape regression models, and those based on deep learning.

In 1995, Cootes et al. [1] proposed the Active Shape Model (ASM), which is a method for extracting keypoint coordinates on the distribution model of feature points (PDM). The overall process of detection involves the "manual calibration of training sets–aligned shape model construction–search matching". After improving the ASM model structure, the Active Appearance Model (AAM) [2] was proposed, which combines shape and texture information to locate keypoints based on the grayscale value of all pixels in the face. The ASM and AAM algorithms are pioneering algorithms in the field of facial keypoint detection, and many researchers have developed their own methods based on them [3,4]. In 2010, Dollar [5] proposed the Cascaded Pose Regression (CPR) algorithm, which first sets the initial predicted value and then uses cascade regressors to further narrow the range of the initial prediction while gradually determining the shape of the object and

finally combines all regressors to obtain the final detection result. Afterward, many facial keypoint detection algorithms using feature detection combined with cascade regression appeared [6,7]. In 2013, Professor Tang Xiao'ou [8] of the Chinese University of Hong Kong and his team proposed the first application of the deep convolutional neural network (DCNN) in detecting facial keypoints, which consists of three cascaded networks of different levels to detect the five keypoints of the left eye, right eye, nose, and mouth on both sides. The obtained keypoint coordinates are more accurate compared to the first two methods. Zhou et al. [9] used a 68-point dataset for training the cascade network and proposed an improved DCNN model. Since then, cascade regression convolutional neural networks have become the mainstream method for researching facial keypoint detection, including many algorithms, such as the TCDCN [10] network and MTCNN [11] network proposed by Zhang et al. and the PFLD algorithm proposed by Guo Xiaojie and others [12].

In summary, the current facial keypoint algorithms have poor generalization in complex environments, and the balance between accuracy and speed needs improvement. Therefore, this study aims to address these issues and enhance the robustness and performance of facial keypoint detection algorithms. Specifically, based on the PFLD facial keypoint algorithm, this paper integrates BlazeNet [13] and Ghost Module [14] models to design a new backbone network called Blaze_ghost. ASPP (Atrous Spatial Pyramid Pooling) is also incorporated for more accurate keypoint extraction. Additionally, a new loss function called HuberWingloss is introduced, and the Adadeleta [15] optimizer is used. These improvements enable the enhanced facial keypoint model to extract keypoints more accurately in a shorter time, meeting the requirements for fatigue feature extraction in driving environments.

## 2. Baseline Model

In this paper, the PFLD (Practical Facial Landmark Detector) model is chosen as the baseline model. The facial keypoint detection algorithm called the PFLD was jointly proposed by Tianjin University, Wuhan University, Tencent AI Lab, and others in February 2019. This algorithm exhibits significant advantages in terms of accuracy, efficiency, and model compression. The model structure consists of a backbone network and an auxiliary network. The backbone network is responsible for predicting facial keypoints, while the auxiliary network is used to predict facial poses.

### *2.1. PFLD Backbone Network*

As shown in Table 1, the backbone network of the PFLD adopts a structurally optimized MobileNet-V2 [16] lightweight network. This network is used to locate the position coordinates of facial keypoints and greatly reduces the model's parameter and computational complexity due to its unique network structure, thereby improving the model's execution speed.

**Table 1.** PFLD backbone network.

| Input | Operator |
|---|---|
| $112^2 \times 3$ | Conv3 × 3 |
| $56^2 \times 64$ | Depthwise Conv3 × 3 |
| $56^2 \times 64$ | Bottleneck |
| $28^2 \times 64$ | Bottleneck |
| $14^2 \times 128$ | Bottleneck |
| $14^2 \times 128$ | Bottleneck |
| (S1) $14^2 \times 16$ | Conv3 × 3 |
| (S2) $7^2 \times 32$ | Conv7 × 7 |
| (S2) $7^2 \times 32$ | - |
| S1, S2, S3 | Full Connection |

### 2.2. PFLD Auxiliary Network

The auxiliary network of the PFLD, as shown in Table 2, is a branch of the backbone network used for head pose prediction during training to improve the localization accuracy of keypoints. By default, the auxiliary network is not used during testing. The purpose of this is to adjust the loss parameters based on the head pose angle obtained during training, making the model pay more attention to rare samples and samples with large pose angles and predict keypoint position coordinates more stably and robustly.

**Table 2.** PFLD auxiliary network.

| Input | Operator |
|---|---|
| $28^2 \times 64$ | Conv3 × 3 |
| $14^2 \times 128$ | Conv3 × 3 |
| $14^2 \times 128$ | Conv3 × 3 |
| $7^2 \times 32$ | Conv7 × 7 |
| $1^2 \times 128$ | Full Connection |
| $1^2 \times 32$ | Full Connection |

### 2.3. The Loss Function of PFLD

The initial loss function of the PFLD algorithm is expressed as the formula below, where $M$ represents the number of samples, $N$ represents the number of keypoints, $\gamma_n$ represents different weights, and $\| * \|$ is the distance metric for feature points.

$$L = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_n \| d_n^m \| \tag{1}$$

The loss function of the PFLD algorithm takes into account the possible significant differences in the number of samples from different categories in the training set. The head pose angles obtained from the auxiliary branch are applied to the loss penalties, and rare samples are assigned higher weights to further refine $\gamma_n$. The optimized loss function is expressed as the formula below.

$$L = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} \left( \sum_{c}^{C} \omega_n^c \sum_{k=1}^{K} \left( 1 - cos\theta_n^k \right) \right) \| d_n^m \|_2^2 \tag{2}$$

## 3. Proposed Method

Based on the PFLD, this paper retains the output of keypoint coordinates and Euler angles from the auxiliary network for pose correction and designs a new backbone network called Blaze_ghost and a new loss function named HuberWingloss. This results in a facial keypoint model with a higher accuracy, better performance, and generalization ability for complex environments.

### 3.1. Related Theory

- The idea of BlazeNet is to achieve efficient feature extraction and accurate face detection through the use of depthwise separable convolution, a lightweight feature pyramid, and efficient model design strategies. Compared with traditional deep convolutional networks, BlazeNet has the advantages of low computational complexity, a small storage size, fewer parameters, and a fast speed, making it suitable for edge devices such as mobile phones. As shown in Table 3, BlazeNet's structure consists of a series of BlazeBlocks and DoubleBlazeBlocks, where the BlazeBlock is a basic block consisting of multiple depthwise separable convolution layers, batch normalization layers, and ReLU activation functions, as well as shortcut connections to help information transfer and gradient back propagation. The DoubleBlazeBlock, in contrast, is a block consisting of two BlazeBlocks and a shortcut connection, with a more

complex structure that can further improve the model's detection ability. In addition to the basic block, BlazeNet also employs the idea of a lightweight feature pyramid, where each DoubleBlazeBlock downsamples its output feature map and stacks it with the feature maps obtained from different scales, forming a set of feature maps with different scales; the low-resolution feature map can capture a larger range of facial information, while the high-resolution feature map is more suitable for detecting small faces. Stacking feature maps of different levels improves the model's detection ability for faces of various scales.

- The Ghost Module is a lightweight module based on depthwise separable convolution, designed to reduce the model's parameter and computational complexity. The Ghost Module consists of two parts: primary convolution and a cheap operation. The primary convolution is a regular convolutional operation, used to extract the main features from the input feature map. The cheap operation, in contrast, is a depthwise separable convolution, used for more detailed processing of the input feature map. Finally, the feature maps obtained from the primary convolution and cheap operation are concatenated and returned as a feature map with the specified output channel number.
- ASPP (Atrous Spatial Pyramid Pooling) is a neural network module used for image semantic segmentation, which can enhance the model's receptive field and improve the accuracy of segmentation. ASPP captures multi-scale information by introducing atrous convolution kernels of different scales, achieving the effect of spatial pyramid pooling. Specifically, ASPP applies different atrous convolution kernels at a given spatial scale to obtain feature information at different scales. Then, it uses a global average pooling layer to compress the convolutional result into a feature vector and then maps the feature vector to the same size as the input feature map through $1 \times 1$ convolution, which serves as the output ASPP feature map. In ASPP, the expansion size of the atrous convolution kernel used is called the dilation rate, and the larger the dilation rate, the wider the range of the convolution kernel's receptive field. ASPP is widely used in visual tasks that require the capture of multi-scale information, such as object detection and keypoint detection.

**Table 3.** BlazeNet network structure.

| Input | Layer |
|---|---|
| $128^2 \times 3$ | Convolution |
| $64^2 \times 4$ | Single BlazeBlock |
| $64^2 \times 4$ | Single BlazeBlock |
| $64^2 \times 4$ | Single BlazeBlock |
| $32^2 \times 48$ | Single BlazeBlock |
| $32^2 \times 48$ | Single BlazeBlock |
| $32^2 \times 48$ | Double BlazeBlock |
| $16^2 \times 96$ | Double BlazeBlock |
| $16^2 \times 96$ | Double BlazeBlock |
| $16^2 \times 96$ | Double BlazeBlock |
| $8^2 \times 96$ | Double BlazeBlock |
| $8^2 \times 96$ | Double BlazeBlock |

### *3.2. Backbone Network Design*

The Blaze_ghost network proposed in this article first utilizes the Ghost Module to replace some of the convolutional layers in BlazeNet. The Ghost Module is a lightweight convolutional layer module that can significantly reduce the computational complexity and model size while maintaining accuracy. It achieves this by dividing the input feature map into smaller subsets and applying different linear transformations to each subset, which reduces the number of parameters that need to be learned in the model and improves its generalization ability.

Secondly, the model uses an ASPP module, which is a multi-scale spatial pyramid pooling module that can capture global contextual information from different scales and enhance the model's ability to distinguish specific objects. Specifically, the ASPP module performs pyramid pooling on the feature maps at different scales and uses convolutional layers and interpolation operations to upsample the feature maps to the original size and then concatenates them at these different scales to obtain more global contextual information.

In summary, this fusion model has been optimized for computational efficiency and the model size, while using the ASPP module to improve the model's recognition performance. The optimized backbone network model structure is shown in Table 4.

**Table 4.** Blaze_ghost network structure.

| Layer | Input | Output |
|---|---|---|
| firstconv | (B,3,H,W) | (B,24,H/2,W/2) |
| BlazeBlock | (B,24,H/2,W/2) | (B,48,H/4,W/4) |
| Double BlazeBlock1 | (B,48,H/4,W/4) | (B,96,H/8,W/8) |
| Double BlazeBlock2 | (B,96,H/8,W/8) | (B,96,H/8,W/8) |
| ASPP | (B,96,H/8,W/8) | (B,96,H/8,W/8) |
| FC | (B,96 × 7 × 7) | (B,numclass) |

In Table 4, B represents batch size and H and W represent the height and width of the input image. The network structure of Blaze_ghost mainly consists of the Ghost Module, BlazeBlock, Double BlazeBlock, ASPP module, and fully connected layers, which are used to implement the task of facial landmark detection. Here are detailed explanations of each module:

- First convolutional layer: After passing through this layer, the input image's channel number changes from 3 to 24, which can effectively extract high-level features of the image.
- The subscript for the permeability of vacuum $\mu_0$, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- BlazeBlock feature extractor: Consisting of several Ghost Modules, it outputs features maps with higher dimensions for further processing in the next step.
- Double BlazeBlock feature extractor: Similar to BlazeBlock, it consists of multiple Ghost Modules and is used to extract more complex features. Two Double BlazeBlock modules are used in this model to give the model stronger representation capabilities.
- ASPP module: It adopts the method of multi-scale parallel convolution for feature fusion, which is used to capture information of different scales and extract more contextual features. In this model, the ASPP module receives the output of the last Double BlazeBlock module as input and the output feature is used to predict facial landmarks.
- Fully connected layer: It maps the previously generated feature maps to a 136-dimensional vector, which is used to predict the position of facial landmarks.

Overall, Blaze_ghost has a certain level of complexity while being lightweight, which can achieve high accuracy and computational efficiency. It is well suited for tasks such as facial landmark detection.

### *3.3. The Design of the Loss Function*

In this article, a new loss function called HuberWingloss is designed by linearly combining huber_loss [17] and wing_loss [18], as shown in the formula below, to supervise the training of the model. Huber_loss is typically used to handle outliers in regression problems, as it is more robust and less sensitive to outliers than L2 loss; wing_loss is a

loss function that balances smoothness and robustness, which can effectively improve the accuracy of the model. Combining the two can further enhance the robustness of the model.

$$abs_{error} = |y - \hat{y}| \tag{3}$$

$$huberloss = \begin{cases} \frac{1}{2}(abs_{error})^2, & abs_{error} \leq \delta \\ \delta\left(abs_{error} - \frac{1}{2}\delta\right), & abs_{error} > \delta \end{cases} \tag{4}$$

$$wingloss = \begin{cases} w * \ln\left(1 + \frac{abs_{error}}{\epsilon}\right), & abs_{error} \leq w \\ abs_{error} - c, & abs_{error} > w \end{cases} \tag{5}$$

$$c = w - w * \ln\left(1 + \frac{w}{\epsilon}\right) \tag{6}$$

$$L(y, \hat{y}) = huber_{loss} + wingloss \tag{7}$$

$$L_{weighted} = euler\_angle\_weights * L(y, \hat{y}) \tag{8}$$

$$HuberWingloss = \frac{1}{N}\sum_{i=1}^{N} L_{weighted} \tag{9}$$

In the equation, $y$ represents the true label, $\hat{y}$ represents the predicted value of the model, *abs_error* represents the absolute difference between the two, *huberloss* represents the result obtained from the calculation of the huber loss function, $\delta$ represents the critical point where the huber loss changes from square loss to linear loss, *wingloss* represents the result obtained from the calculation of the wing loss function, $w$ and $c$ are parameters of the wing loss function, and $c$ is a constant. $L(y, \hat{y})$ represents the calculation method of the total loss function, using different calculation methods in different *abs_error* intervals. *euler_angle_weights* represents the weight factors of the euler angle error in each sample. When calculating the final loss value, if the *euler_angle_weights* parameter is not empty, then each loss function value will be further multiplied, which helps adjust the importance of different euler angle dimensions in the loss function. This can make the model pay more attention to samples with larger euler angle errors. $L_{weighted}$ represents the weighted loss value obtained by applying the weight factor to the loss value of each sample, and, finally, *HuberWingloss* represents the sum of the weighted loss values of all samples divided by the number of samples, which is the final loss value of the model.

## 4. Experiment and Result Analysis

### *4.1. Dataset*

The WFLW [19] dataset is primarily a facial keypoint localization dataset used for training facial alignment algorithms. As shown in Figure 1, it contains rich attribute annotations, such as occlusion, poses, makeup, lighting, blur, and expressions, enabling comprehensive analysis of existing algorithms. Compared to previous datasets, the WFLW dataset exhibits significant variations in facial expressions, poses, and occlusion, allowing for the evaluation of robustness in these aspects. The WFLW dataset consists of 10,000 images, each accompanied by an annotation file. These images cover a wide range of features, expressions, and poses from faces of different ages, races, and genders. The annotation information in the dataset includes the bounding box of the face and the coordinates of 98 keypoints, which mark important facial locations, such as the eyes, nose, and mouth.


**Figure 1.** Partial WFLW dataset.

The common facial landmark localization methods include 5-point, 68-point, and 98-point localization. The 5-point localization can only be used for facial contour localization; the 98-point localization contains too much facial information, which can cause a large computational load and low recognition efficiency. The 68-point localization can accurately describe the facial contour and local features, such as the eyes and mouth, while reducing the computation time and improving the recognition efficiency. Therefore, in this paper, we processed the WFLW dataset for the training and testing of 68-keypoint localization. As shown in Figure 2, in 68-point localization there are 6 keypoints in the left eye (37–42), 6 keypoints in the right eye (43–48), and 20 keypoints in the mouth (49–68), while the rest are facial contour keypoints.

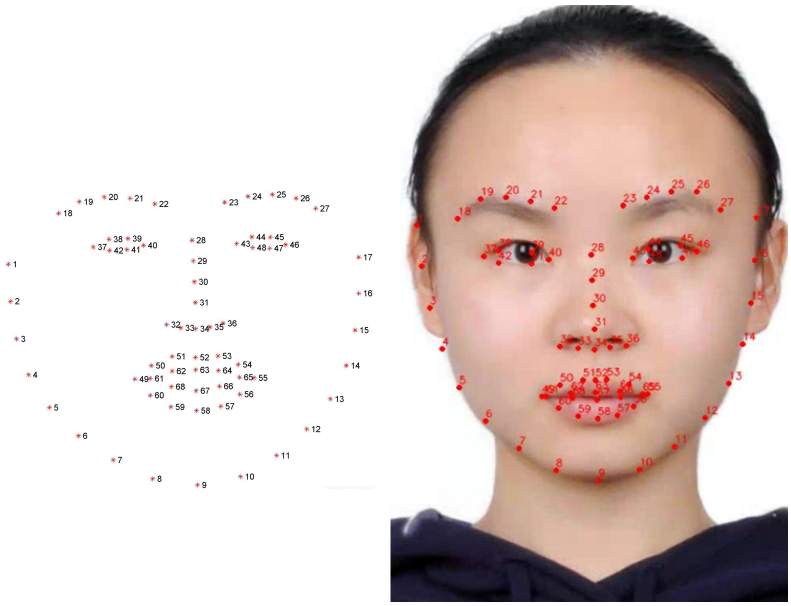

**Figure 2.** Sixty-eight points of face keypoints.

### *4.2. Experimental Environment and Parameter Settings*

The experimental environment for this experiment is the Ubuntu 20.04.5 LTS operating system, CUDA10.0, cudnn7.6.5, an IntelR CoreTM i5-9500 CPU @ 3.00GHz × 6 processor, an NVIDIA Corporation TU104 (GeForce RTX2080 SUPER) graphics card, 16 G of memory, the Pycharm community compilation platform, python3.7, and pytorch1.7.1.

Before training the model, the network training parameters were set, and, based on factors such as the actual graphics memory size, a batch_size parameter value of 16 was set in this paper. Adadelta was selected as the optimizer, and the decay rate rho was set to the default value of 0.95, which controls the exponential weighted average of historical gradient

squares and determines the adaptive range of the learning rate. The increment stability coefficient epsilon was set to the default value of $1 \times 10^{-6}$, which is a small constant used to avoid division by zero, and the initial value of the smooth average is an epsilon. The Adadelta algorithm is a powerful and adaptive gradient descent algorithm, which further simplifies the parameter adjustment process compared with Adam and SGD algorithms since it does not require the manual setting of global learning rates and momentum, and it is more adaptive and robust. In the entire training process, a total of 100 epochs were iterated; that is, all samples were trained 100 times.

For data preprocessing, in order to improve the robustness and generalization ability of the model, this paper adopted a series of data augmentation methods, including color transformation, scaling, rotation, and Gaussian blur, to increase the sample diversity of the training set and improve the robustness of the model.

### *4.3. Result and Discussion*

#### 4.3.1. Comparative Experiment

This paper uses the NME (Normal Mean Error), FR (Failure Rate), Inference Time, and model size as evaluation indicators. Among them, the NME is the normalized average error, which is the value obtained by averaging the normalized estimated errors of all keypoints on a face. The FR is another standard for evaluating the accuracy of facial keypoint positioning algorithms. The Inference Time is the time it takes to locate the keypoints. This paper uses 2500 testing data from WFLW for testing and compares the experimental data of the proposed model with the data of the original network model, as shown in Table 5.

**Table 5.** Comparative experiment.

| Backbone Network | NME | FR | Model Size | Inference Time |
|---|---|---|---|---|
| Mobilev2 (PFLD) | 0.062 | 0.125 | 1.1 M | 0.121 (s) |
| Resnet50 | 0.053 | 0.073 | 122.27 M | 0.304 (s) |
| Blaze_ghost | 0.056 | 0.073 | 5.66 M | 0.141 (s) |

From the experimental data comparison in Table 5, it can be seen that the improved backbone network proposed in this paper, Blaze_ghost, has reduced the NME and FR to varying degrees in the evaluation indicators. At the same time, although the Inference Time and model size have not decreased, the model still maintains its lightweight structure. These data indicate that the overall performance of the improved facial keypoint detection algorithm has been significantly improved.

#### 4.3.2. Ablation Experiment

In order to verify the effectiveness of the Blaze_ghost backbone network, the Huber-Wingloss loss function designed in this paper, and the use of Adadelta optimizer, a series of ablation experiments were conducted, and the experimental results are shown in Table 6.

**Table 6.** Ablation experiment.

| Backbone Network | Loss Function | Optimizer | NME | FR | Model Size |
|---|---|---|---|---|---|
| Blazelandmark | Wingloss | SGD | 0.081 | 0.197 | 7.52 M |
| Blaze_ghost | Wingloss | SGD | 0.077 | 0.178 | 5.66 M |
| Blaze_ghost | Wingloss | Adadelta | 0.057 | 0.082 | 5.66 M |
| Blaze_ghost | HuberWingloss | Adadelta | 0.056 | 0.073 | 5.66 M |

From the experimental comparison in Table 6, it can be seen that Blaze_ghost is an effective optimization of the BlazeLandMark network based on the Ghost Module, making the model more lightweight and improving the model's inference speed. At the same time, it can also be seen from the table that the introduction of the Adadelta optimizer and the HuberWingloss loss function proposed in this paper have significantly reduced the two indicators of the NME and FR. This indicates that the Adadelta optimizer and the HuberWingloss loss function can effectively apply to the improved keypoint detection algorithm proposed in this paper, thereby improving the performance of keypoint detection.

#### 4.3.3. Detection Performance Validation

In order to visually check the effect of keypoint localization, the detection results of some facial images were output and displayed. The detection results under normal conditions are shown in Figure 3, which can intuitively show that the labeled keypoints can accurately depict the contour of the key facial area, demonstrating good performance.

**Figure 3.** Effect of detection under normal conditions.

Due to the complexity and variability of the real driving environment, drivers may wear face coverings, such as glasses or masks, and there may be an uneven facial lighting distribution due to lighting environment effects. In addition, image blurring may occur due to issues with the acquisition equipment such as infrared cameras. These factors can all affect the effect of keypoint localization, as shown in Figures 4–6, which show the localization results of the improved model under different influencing factors. It can be seen intuitively from the detection result figures that the above factors did not cause too much interference in the detection results, indicating that the improved algorithm has good robustness.

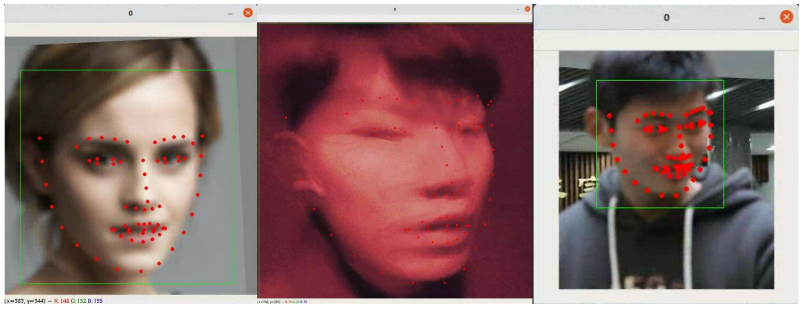

**Figure 4.** Detection effect when the acquisition image is blurred.

To verify the performance of our keypoint detection model in extracting fatigue features, we conducted experiments in a simple system that recognizes blinking and yawning behavior through mouth aspect ratio (MAR) and eye aspect ratio (EAR) calculation using face keypoints. We first used the retinaface [20] face detection network to locate and refine the target area of the portrait in the video image frames from coarse to fine to improve the localization efficiency and reduce the size of subsequent input images. Then, we used our proposed keypoint detection algorithm to extract keypoints and finally extract the fatigue features. Different faces were tested in the experiment, and the results

are shown in Figure 7. It can be seen that under oblique facial conditions the model can fit the eye contour well and accurately reflect the opening and closing status of the eyes. When the driver yawns, the mouth keypoints can also fit well and accurately reflect the driver's mouth state. In summary, based on the results of different facial fatigue features, our proposed face keypoint detection algorithm can accurately detect and identify the face in the image, extract 68 keypoints of the face, and accurately extract fatigue features, such as blinking and yawning, indicating that our algorithm has high effectiveness and stability and can meet the requirements of fatigue feature extraction.

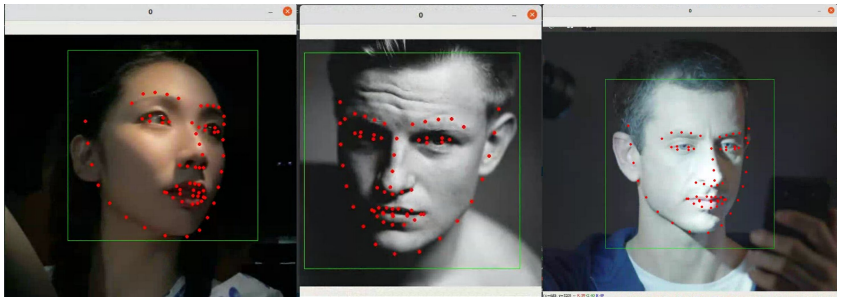

**Figure 5.** Detection effect when light distribution is not uniform.

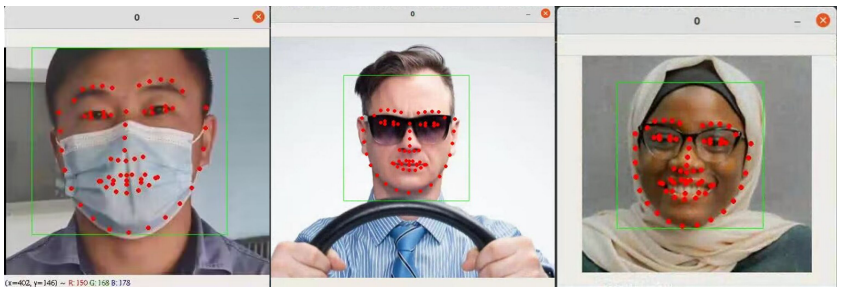

**Figure 6.** Detection effect in case of partial occlusion.

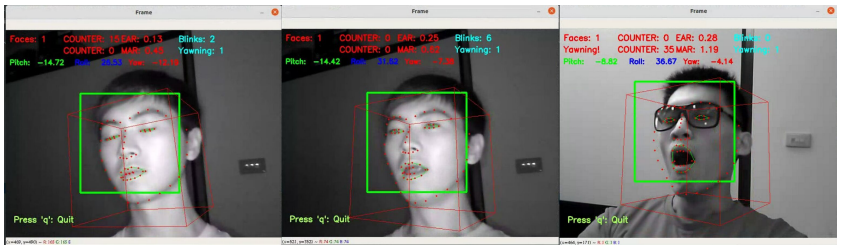


**Figure 7.** Detection effect of fatigue features.

## 5. Conclusions and Future Work

This paper proposes a facial keypoint detection method based on the Blaze_ghost network. It designs the Blaze_ghost network as the backbone network and utilizes Huber-Wingloss as the loss function. Adadelta is used as the optimizer. The proposed algorithm is compared and analyzed with the PFLD algorithm in the WFLW dataset for 68-point keypoint detection. The advantages of the proposed facial keypoint detection method are validated through metrics such as the NME, FR, Inference Time, and model size. Additionally, the real effects of the proposed method on facial keypoint detection and simple fatigue feature extraction tasks like eye blinking and yawning are visually demonstrated using images. In future research, we will further expand and improve the facial keypoint localization method. Firstly, we plan to consider training and evaluating with larger and more diverse datasets to enhance the algorithm's robustness and generalization ability. Secondly, we will explore more advanced deep learning models and algorithms or introduce self-supervised learning methods to reduce the reliance on labeled data.

**Author Contributions:** Conceptualization, Y.T. and N.Y.; methodology, Y.T.; software, Y.T.; validation, Y.T.; investigation, Y.T. and X.Y.; resources, N.Y. and X.Z.; data curation, Y.T.; writing—original draft preparation, Y.T.; writing—review and editing, Y.T.; visualization, Y.T.; supervision, N.Y. and X.Z. 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:** The source code and datasets of the experiments can be obtained upon request.

**Acknowledgments:** The authors would like to thank Ming Gao for his valuable suggestions. The authors would like to thank the National Engineering Research Centre for High Mobility Riot Vehicle Technology 2023 Open Fund Project for supporting the paper.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| PFLD | A face keypoint detection model |
| Blaze_ghost | The proposed face keypoint detection model |
| NME | Normal Mean Error |
| FR | Failure Rate |
| ASPP | Atrous Spatial Pyramid Pooling |

## References

1. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active Shape Models-Their Training and Application. *Comput. Vision Image Underst.* **1995**, *61*, 38–59. [CrossRef]
2. Cootes, T.F.; Wheeler, G.V.; Walker, K.N.; Taylor, C.J. View-Based Active Appearance Models. *Image Vis. Comput.* **2002**, *20*, 657–664. [CrossRef]
3. Cootes, T.F.; Taylor, C.J. *Combining Elastic and Statistical Models of Appearance Variation*; Springer: Berlin/Heidelberg, Germany, 2000.
4. Matthews, I.; Baker, S. Active Appearance Models Revisited. *Int. J. Comput. Vis.* **2004**, *60*, 135–164. [CrossRef]
5. Dollár, P.; Welinder, P.; Perona, P. Cascaded pose regression. In Proceedings of the OAI, San Francisco, CA, USA, 13–18 June 2010.
6. Fu, J.; Shucheng, H. Research on cascade regression face alignment method for multi-feature fusion. *J. Jiangsu Univ. Sci. Technol. (Nat. Sci. Ed.)* **2020**, *34*, 6.
7. Zhao, D.Y.G.; Yuan, L. Residual Neural Network Improvement of Landmark with Cascaded Framework. *Softw. Guide* **2018**, *17*, 5.
8. Sun, Y.; Wang, X.; Tang, X. Deep Convolutional Network Cascade for Facial Point Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.
9. Zhou, E.; Fan, H.; Cao, Z.; Jiang, Y.; Yin, Q. Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013.
10. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Facial Landmark Detection by Deep Multi-task Learning. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
11. 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]
12. Guo, X.; Li, S.; Zhang, J.; Ma, J.; Ling, H. PFLD: A Practical Facial Landmark Detector. *arXiv* **2019**, arXiv:1902.10859.
13. Bazarevsky, V.; Kartynnik, Y.; Vakunov, A.; Raveendran, K.; Grundmann, M. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. *arXiv* **2019**, arXiv:1907.05047.
14. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C. GhostNet: More Features From Cheap Operations. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
15. Yang, Z.; Jun, X.U. Face Recognition System Based on MobileNetV2 and Raspberry Pi. *Comput. Syst. Appl.* **2021**, *30*, 6.
16. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
17. Gokcesu, K.; Gokcesu, H. Generalized Huber Loss for Robust Learning and its Efficient Minimization for a Robust Statistics. *arXiv* **2021**, arXiv:2108.12627.
18. Feng, Z.H.; Kittler, J.; Awais, M.; Huber, P.; Wu, X.J. Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.

19. Wu, W.; Qian, C.; Yang, S.; Wang, Q.; Cai, Y.; Zhou, Q. Look at Boundary: A Boundary-Aware Face Alignment Algorithm. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
20. Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-stage Dense Face Localisation in the Wild. *arXiv* **2019**, arXiv:1905.00641.