applied sciences

MDPI

Article

Real-Time Pre-Identification and Cascaded Detection for Tiny Faces

Ziyuan Yang [1,†], Jing Li [1,†], Weidong Min [2,3,*] and Qi Wang [1]

1 School of Information Engineering, Nanchang University, Nanchang 330031, China; yangziyuan@email.ncu.edu.cn (Z.Y.); jingli@ncu.edu.cn (J.L.); 351029018003@email.ncu.edu.cn (Q.W.)
2 School of Software, Nanchang University, Nanchang 330047, China
3 Jiangxi Key Laboratory of Smart City, Nanchang 330047, China
* Correspondence: minweidong@ncu.edu.cn; Tel.: +86-0791-8830-4080
† The first two authors contributed equally to this work.

Received: 6 September 2019; Accepted: 6 October 2019; Published: 15 October 2019

Abstract: Although the face detection problem has been studied for decades, searching tiny faces in the whole image is still a challenging task, especially in low-resolution images. Traditional face detection methods are based on hand-crafted features, but the features of tiny faces are different from those of normal-sized faces, and thus the detection robustness cannot be guaranteed. In order to alleviate the problem in existing methods, we propose a pre-identification mechanism and a cascaded detector (PMCD) for tiny-face detection. This pre-identification mechanism can greatly reduce background and other irrelevant information. The cascade detector is designed with two stages of deep convolutional neural network (CNN) to detect tiny faces in a coarse-to-fine manner, i.e., the face-area candidates are pre-identified as region of interest (RoI) based on a real-time pedestrian detector and the pre-identification mechanism, the set of RoI candidates is the input of the second sub-network instead of the whole image. Benefiting from the above mechanism, the second sub-network is designed as a shallow network which can keep high accuracy and real-time performance. The accuracy of PMCD is at least 4% higher than the other state-of-the-art methods on detecting tiny faces, while keeping real-time performance.

Keywords: face detection; tiny faces; pre-identification mechanism; cascaded detector; deep learning; convolutional neural network

1. Introduction

Face detection is one of the most hot topics in computer vision as it is a key step for many different applications, such as face recognition [1], facial expression analysis [2], eye-tracking [3], facial performance capture [4], facial expression transformation [5], etc. In fact, the applications are not limited to the traditional areas, there are still some exciting interdisciplinary applications [6–11] in the field of animation. However, many factors such as the illumination, occlusion, and the diversity of faces cause huge challenges in face detection.

Using universal face templates to detect faces is one of the main research fields of traditional methods. Determining whether there is a face is undertaken by calculating the correlation coefficient between the area which is detected and the template [12]. However, facial skin color, different expressions, and occlusion lead to the method being less robust and computationally complex.

In recent years, many researchers shift their attention from the traditional methods to convolutional neural networks (CNNs) [13–15] since they have achieved remarkable success in many important tasks of computer vision, such as classification, detection, and recognition. Lots of approaches have been proposed to solve the problem of tiny-face detection, which aims to search a tiny face in a whole image,

especially in a low-resolution image. However, these methods cannot achieve satisfactory performance because the features of tiny faces are different from those of big faces and tiny faces contain limited information available for face detection.

It is challenging to detect small objects in image detection. That is because these networks are designed to propose default boxes and the classification score is calculated from one single deep CNN. For example, faster Region-based Convolutional Neural Networks (R-CNN) [16] extracts features by visual geometry group (VGG)-16 [17], but when the face size is less than 16 × 16, the output in 'conv5' is less than one pixel. As the convolutional layer is deeper, each pixel in the feature map gathers more information outside of the original input area and lower information of the region of interest (RoI), which means these methods cannot keep the performance when targets are small. However, a single shallow CNN cannot get enough information for object detection, and thus a cascaded structure which is usually divided into a prediction part and a regression part becomes a popular and effective framework for face detection. A cascade framework can generate a large number of bounding box candidates based on a low threshold, and then extracts the regression scores of these candidates. That is, the information loss of shallow convolutional layers can be effectively relieved.

Moreover, the above-mentioned face detection algorithms cannot keep good performance in video surveillance, because they are designed for high-resolution images and big faces. Limited by the cost of large-scale surveillance and the scale of data, the face targets are often small and not clear. Additionally, the head movement tends to be more frequent than the body, making it difficult to detect tiny faces in multiple views. Previous methods are difficult to accurately detect tiny faces in this kind of scenario. If tiny faces can be detected automatically in real time, the telephoto lens would immediately be aimed at the suspicious persons' faces, which can help the police to conduct reconnaissance. This is very useful in public security, since violent incidents and terrorist attacks have occurred in many places in recent years.

In order to alleviate the above problems, this paper presents a cascaded framework named pre-identification mechanism and a cascaded detector (PMCD) to detect tiny faces based on two independent CNNs and a pre-identification mechanism. The two sub-networks can be trained separately, which greatly improves the flexibility in training networks. The first sub-network of PMCD and the pre-identification mechanism generate a set of RoI candidates, defined by the pedestrian area in the image. Then, the RoI candidates are resized to different scales to build an image pyramid as the input of the following network. As the input of the second sub-network has greatly reduced the irrelevant area, a shallow network can learn enough features and guarantee real-time performance.

We tested PMCD on a self-collected dataset and Caltech Pedestrian [18] dataset. Due to the scarcity of the tiny-face detection dataset, we collected 1370 images, 2450 faces in total and most faces were less than 20 × 20. In order to test in different situations, the size of 562 faces were larger than 20 × 20, while the rest were all smaller than the size in the self-collected dataset. As the results showed, compared with other state-of-the-art methods, PMCD achieves impressive performance in tiny-face detection and it can achieve real-time detection.

To this end, this paper aims to alleviate the issues discussed above which are tiny-face detection and keeping real-time features.

- We propose a new pre-identification mechanism to obtain a set of face candidates as the input for the face detector, which contains higher face proportion than that in the original image. The mechanism greatly reduces miss rate of tiny-face detection and leads to the robustness of PMCD.
- We propose a novel cascade neural network called PMCD for real-time tiny-face detection. The first sub-network is a deep pedestrian detector as a part of the pre-identification mechanism, and the second sub-network is designed based on a shallow multi-task CNN to detect faces in the RoI candidates.

The rest of this paper is as follows. Section 2 introduces related works on face detection. Section 3 details the framework including two sub-networks and the pre-identification mechanism. Section 4

mainly describes the performance of PMCD and compares it with other state-of-the-art methods. Section 5 concludes and looks forward to future works.

2. Related Works

Face detection is the key step in many different face-related applications and studies [19]. Most of the early work was designed for high-resolution images and large targets by using statistical learning methods to automatically extract features. Yaman et al. [20] proposed a framework to detect faces by utilizing histogram-based feature extraction with random subspace and voting ensemble learners. Luo et al. [21] proposed a face location algorithm which is developed to extract face regions with a high proportion of skin. This type of method increases the speed of the operation. Since the features are not completely selected by humans, the robustness is improved. However, the effect is still poor when the target is small. Mohanty et al. [22] proposed a new feature and combined it with the gray level feature and skin color feature, and this method improves detection speed in complex backgrounds and reduces the computational complexity. Additionally, Ma et al. [23] used the geometric relationship between facial organs to generate four Haar-like features for face detection based on the traditional Adaboost classifier, which greatly reduces the detection time. One of the most impressive traditional methods is Viola-Jones [24] which designed cascade classifiers based on the Haar feature and AdaBoost classifier, but all of these traditional methods focus on improving the performance with more effective hand-crafted features and more powerful classifiers [25–28]. These features or detection structures have certain subjective factors, which leads to the robustness of these frameworks being poor, and the operation being complicated and time consuming.

In recent years, deep learning models offer new ideas in solving many research problems such as classification, object detection, image segmentation, image restoration etc. Convolutional neural networks have achieved remarkable success in many different tasks of computer vision, especially for object detection. The family of Region-based Convolutional Neural Networks (R-CNN) [17,29,30], You Only Look Once (YOLO) [31], and Single Shot Multibox Detector (SSD) [32] are the most efficient and popular approaches for detecting objects these years. Inspired by these brilliant methods, researchers have proposed many effective and robust face detection structures. Jiang et al. [33] investigated applying faster R-CNN to face detection, compared with previous proposed models, and its accuracy has a significant improvement. Contextual multi-scale region-based CNN (CMS-RCNN) [34] combined multi-scale information which consists of the region proposal component and the RoI detection component to detect faces. Wan et al. [35] improved Faster R-CNN with the hard negative mining and significant boosts, and the hard negatives harvested from a large set of background examples. Sun et al. [36] applied some effective strategies including feature concatenation, hard negative mining, multi-scale training, model pre-training, and proper calibration to improve Faster R-CNN in face detection. Zhang et al. [37] light-designed R-CNN and improved the performance by integrating multi-scale training, multi-scale testing, some tricks for inference, and a vote-based ensemble method. Enlightened by SSD, Hsu et al. [38] proposed Multiple Dropout Framework to detect faces. Due to the anchor mechanism, the family of R-CNN can detect objects which occupy the majority of an image which are clear and huge, but the accuracy of these methods drop rapidly when the target is small.

Li et al. [39] proposed a novel cross-level parallel network (CLPNet) to extract low-level features and fuse them in the high-level stage, and CLPNet achieved remarkable performance on crowd counting. Triantafyllidou et al. [40] proposed a novel lightweight deep neural network and a new training method of progressive positive and hard negative sample mining to improve training speed and accuracy. A Fully Convolutional Network (FCN) [41] generated face proposals by the heat map of facial parts which are scored by a new facial parts responses method by their spatial structure and arrangement. As described in Section 1, cascaded structures are more advantageous than single networks in the task of face detection, and thus have been widely used in face detection, CascadedCNN [42] is a cascade framework built on CNNs and a CNN-based calibration stage was introduced to adjust the position of the detection window. Qin et al. [43] proposed joint training to achieve end-to-end

optimization for CNN cascade and showed back propagation used in training a single CNN can be naturally used in training the cascade CNN structure. Coarse-to-Fine Auto-encoder Network (CFAN) [44] cascades a few Stacked Auto-encoder Networks to accomplish different tasks including face detection. Multitask Cascaded Convolutional Networks (MTCNN) [45] is a multi-task cascade network to detect faces by three stages in a coarse-to-fine detection structure. Min et al. [46] proposed a multi-scale and multi-channel shallow convolutional network (MMSC) for real-time face detection after the pre-identified method detecting faces in the images based on a traditional pedestrian detection method. Hu et al. [47] proposed a multi-task detector with hybrid resolutions (HR), which detects different face scales from multiple layers of a single neural network.

The methods mentioned above have good performance in face detection when images are of high resolution and the faces are big, but all of these traditional methods and region-based single deep CNNs cannot keep a high accuracy in tiny-face detection, the reason is that the effective information of tiny faces is very limited, as explained in Section 1. Inspired by [46], we designed a novel structure to reduce the unnecessary input of the face detection order to improve the effectiveness and propose a new pre-identification mechanism.

3. Proposed Method

3.1. Overview of Our Method

In this paper, we used two convolutional neural networks (CNNs) for coarse-to-fine face detection. CNN is actually a multi-layer perception, where each layer is composed of multiple feature maps through different convolutional kernels. The most important advantage of this structure is that the parameters of CNN are self-learned through training data, so the structure avoids the generalization weakness caused by hand-crafted features. CNN usually consists of convolutional layers, pooling layers, and fully connected layers. A convolutional layer is used to extract features, a pooling layer aims to reduce the amount of data for calculation, and a fully connected layer is designed to combine the extracted local features into a powerful global feature. In practical operations, convolution layers and pooling layers are often designed as a whole, i.e., the structure would pool the feature maps after the convolutional operation. However, the output of convolutional layers is not immediately pooled, it is passed to the activation function first and then the results are passed to the pooling layers. Activation functions, which are always nonlinear functions, are used for feature mapping in order to help solve nonlinear problems. The parameters in convolutional layers and pooling layers are learned through the back-propagation algorithm, so the self-learned characteristic makes CNN far superior to the traditional algorithms in precision and robustness in many computer vision tasks such as classification, detection, tracking, etc.

Although face detection has received extensive attention in recent years, it still entails many challenges, such as occlusion, complex environments, small targets, etc. This paper aims to solve the problem of tiny-face detection. Due to the low resolution of face images, detecting faces is not an easy task when a pedestrian is far away from the camera. In this situation, the information of a body is much more than the face information. The pedestrians' faces can be roughly calibrated based on body context information. Considering this, we defined the region at the top of the pedestrians' bounding boxes as RoI candidates, obtained the candidate regions by a pedestrian detector and a novel pre-identification algorithm to reduce the search range as pre-processing.

Methods of pedestrian detection include the method based on statistical learning methods, background modeling, and neural networks. The most influential pedestrian detection structure is what Dalal et al. [48] proposed, which uses Support Vector Machine (SVM) as the classifier to detect faces in the images based on the histograms of oriented gradient (HOG) features. Lots of features have been proposed with better robustness and result in better classification accuracy for pedestrian detection [49–52]. However, these features are often specifically designed to a particular situation, hence the robustness of these traditional methods is not good. In order to alleviate the problems

mentioned above, more and more researchers have adopted convolutional neural networks (CNNs) to detect objects these years [17,29–32,53]. There are many images from different scenes in the training set, so the robustness of features learned by neural networks is better than hand-crafted features. Therefore, these methods based on CNNs have a huge improvement in accuracy and robustness.

Our method consists of two parts, pre-identification detection and face detection. In the pre-identification detection part, we detected the bounding box of the pedestrian by a deep CNN first, then the RoI candidates would be selected by the self-adaptively algorithm, after that we built an image pyramid, which was used to adapt different sizes of faces, as the input for the face detection network. In the face detection part, image pyramid would be passed to our proposed multitask neural network to detect faces. As the previous steps significantly reduce background interference, the multi-task face detector requires only a shallow structure to perform efficiently and in real-time. What we proposed is a cascaded framework, and we can train the two sub-networks separately. The whole framework is shown in Figure 1.

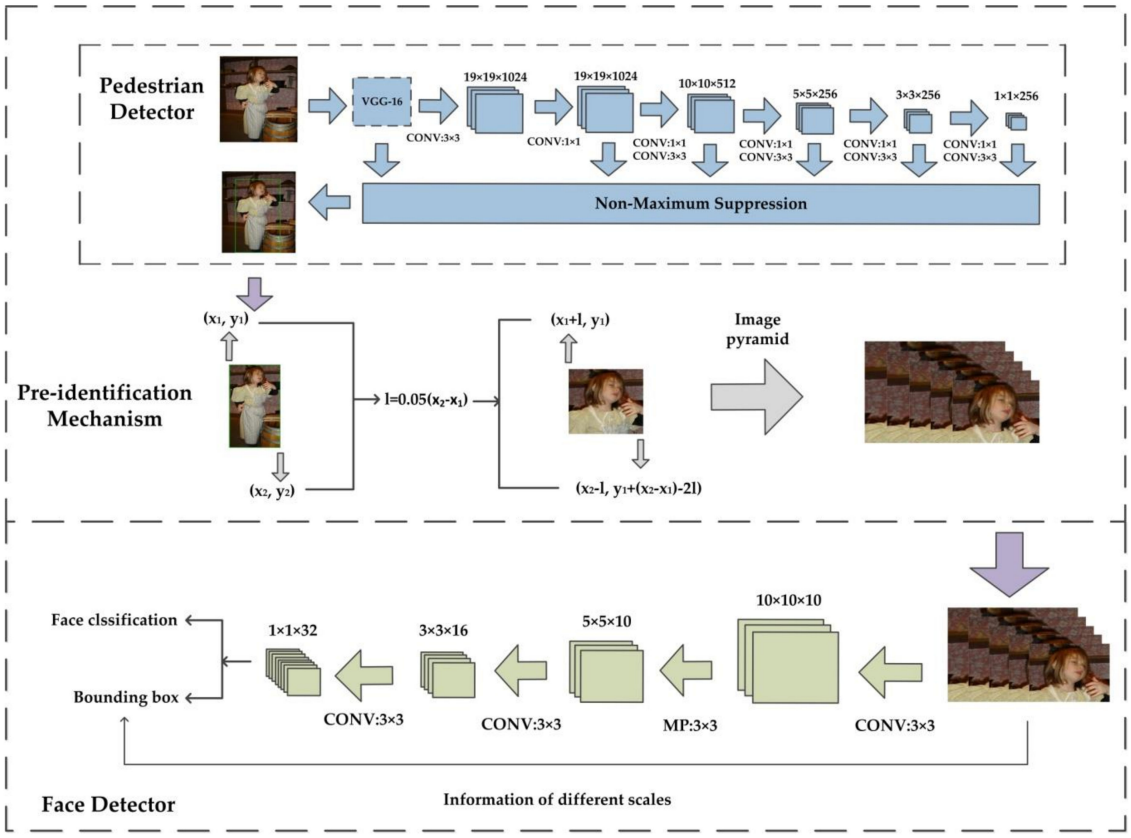

Figure 1. The whole framework.

3.2. Proposed Pre-Identification Mechanism

3.2.1. Pedestrian Detector

The bounding boxes of the pedestrians are detected by a deep CNN [32] in order to narrow the detection range of the face detector, as seen in Figure 1. Firstly, the image size is resized to 300 × 300. Afterwards, the features are extracted by VGG-16 and then fed to six additional convolutional layers for detecting the targets. All default boxes are predicted by combining many feature maps with different scales and ratios, multiple default boxes are set to cover various sizes and shapes of targets. Finally, non-maximum suppression (NMS) is used to obtain the final bounding box which has the highest score in the set of bounding boxes of the target.

Convolutional layers, except VGG-16 in the pedestrian detector, predict bounding boxes and offsets. There are different receptive fields in different levels of feature maps. The scale of the default boxes is computed as:

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 1}(k - 1),\ k\ \epsilon\ [1,\ m], \tag{1}$$

where s_{min} is 0.2 and s_{max} is 0.9, the length l_k and the width w_k of the bounding box are calculated based on s_k; m is the number of default boxes, W is the width of the input and L is the length of the input, r is the ratio of the length and the width in each feature map. The detailed calculation formula is shown as:

$$\begin{cases} w_k = W \times s_k \times \sqrt{r} \\ l_k = L \times s_k / \sqrt{r} \end{cases}. \tag{2}$$

3.2.2. Pre-Identification Mechanism

In order to reduce the misclassification rate in the face detector and reduce the interference of other information, we should reduce the input of the face detector. Face areas are adaptively estimated according to the proportion of the pedestrian area. The proportional relationship between the face-region and the pedestrian-region is estimated by our newly proposed mechanism. The face-region selection operator is used to generate a set of face candidate regions which is described in Algorithm 1, and the whole process of face pre-identification is shown in Figure 1. Herein, N_i and D_i are two sets of coordinates, N_i contains the upper-left corner coordinates of the bounding box N_i^1 and the lower-right corner coordinates N_i^2 through the pedestrian detector, N_i^1 and N_i^2 are composed by x and y; D_i contains the upper-left corner coordinates of the bounding box D_i^1 and the lower-right corner coordinates D_i^2 through the pre-identification algorithm, D_i^1 and D_i^2 are composed by x and y, and i means the i-th person in the image.

Algorithm 1. Pre-identification.

Input: Coordinates of the bounding boxes N_i through the pedestrian detector
Output: Coordinates of the RoID_i

1. $T \leftarrow 0.9$
2. $i \leftarrow 0$
3. $\theta \leftarrow (1 - T)/2$
4. **while** $N_i \neq \varnothing$
5. $l \leftarrow \theta \times \left(N_i^2.x - N_i^1.x\right)$
6. $D_i^1.x \leftarrow N_i^1.x + l$
7. $D_i^1.y \leftarrow N_i^1.y$
8. $D_i^2.x \leftarrow N_i^2.x - l$
9. $D_i^2.y \leftarrow N_i^1.y + \left(N_i^2.x - N_i^1.x\right) - 2 * l$
10. $i \leftarrow i + 1$
11. **end while**

3.2.3. Image Pyramid

The input image size is fixed for the face detector, but the size of the target is not fixed. Therefore, image pyramids can be used to detect different sizes of objects with the fixed input size. We build image pyramids upon the set of RoI candidates. RoIs are resized to different scales which are adaptively computed by the image pyramid method to build an image pyramid, which is the input of the second multi-task face detector. The image pyramid algorithm is described in Algorithm 2. Herein, D_i is the coordinate set of RoI, which is the same as in Section 3.2.2, i means the i-th person in the image, factor is the scaling factor, P_i is an image pyramid of the i-th person, containing many resized images. The threshold of minlin Algorithm 2 is set to 12, since the size of input of the face detector is 12 and the information is pretty limited if the size of face is smaller than 12.

Algorithm 2. Image pyramid.

Input: Coordinates of the RoI D_i
Output: Image pyramid P_i
$w \leftarrow D_i^2.x - D_i^1.x$
$l \leftarrow D_i^2.y - D_i^1.y$
$count \leftarrow 0$
$factor \leftarrow 0.7$
if $w < l$
$minl \leftarrow w$
else
$minl \leftarrow l$
end if
if $minl < 12$
$minl \leftarrow 12$
end if
$m \leftarrow 12/minl$
$minl \leftarrow \text{minl} * \text{m}$
while $\text{minl} \geq 12$
$scales \leftarrow scales + m * factor^{count}$
$minl \leftarrow minl * factor$
$P_i^{count} \leftarrow Resize(D_i, scales)$
$count \leftarrow count + 1$
end while

3.3. Multitask Face Detector

Before detecting the face, we pre-processed the image to reduce the size of the input. The detector is a shallow neural network, but the depth is enough to learn useful features of tiny faces, because the input of the detector is preprocessed, and background information and other interferences is greatly reduced. Another benefit of shallow neural networks is that the parameters are much smaller than deeper detectors, resulting in fast operation and real-time performance.

Each image pyramid is passed to the face detector to detect the coordinates of the face. The real coordinates are calculated based on the detected coordinates of bounding boxes and the image pyramid scales. Finally, the NMS method is used to eliminate the redundant bounding boxes by getting multiple overlapped bounding boxes and reducing them to only one. The structure of our face detector is shown in Figure 1. The stride of the convolutional layer is 1, and the step size of the pooling layer is 2.

Different convolutional kernel sizes may result in different feature extraction effects. Considering that the detected target is a tiny face with the input size is 12×12, we apply a 3×3 filter for all the convolutional layers. There are two tasks in the face detector, face classification and bounding box regression. Compared to other complex multiclass objection classification and detection methods, the input of our proposed network is the image pyramid generated from the set of RoI candidates and the number of classes is only 2, face and non-face, so this network does not need a deep network structure and a lot of convolutional kernels in each layer. ReLU is applied as nonlinearity activation function in this detector, which is:

$$\mathrm{f(x)} = \begin{cases} 0, & x \leq 0 \\ \mathrm{x}, & x < 0 \end{cases}. \tag{3}$$

Face classification is a binary classification problem, so we use the cross-entropy logistic regression function as the loss function. The function is designed as follows:

$$\mathrm{L_i} = -\big(\mathrm{y_i} \log \mathrm{p_i} + \big(1 - \mathrm{y_i}\big)\big(1 - \log \mathrm{p_i}\big)\big), \tag{4}$$

where $\mathrm{p_i}$ is the possibility calculated based on the input $\mathrm{x_i}$; $\mathrm{y_i}$ is the predicted class, $\mathrm{y_i} \in \{1,\ 0\}$.

For the task of bounding box regression, we employ the Euclidean loss of the offset between the predicted bounding box and the nearest ground truth. Predicted by the network are the upper-left coordinate, width, and length. The loss function is given by:

$$G_i = \| \hat{b_i} - b_i \|^2, \tag{5}$$

where $\hat{b_i}$ is the class that the network predicts and b_i is the ground-truth coordinate. We use four values to represent a bounding box, i.e., the upper-left coordinate, width, and length. So $\hat{b_i}$ and b_i are four-dimensional vectors.

The total loss of this network is the weighted sum of the loss values of the above two functions. Faces are contained in RoI candidates after we preprocess the input of the multitask detector, therefore we give the face detector a high tolerance for features of tiny faces. This trick makes the tiny faces pass to the classification task of the multitask detector with a low threshold. Therefore, we give a big weight to the loss function of the bounding box. In addition, the total loss function is as follows:

$$Loss_i = t_1 \times L_i + t_2 \times G_i, \tag{6}$$

where L_i and G_i are the losses of face classification and bounding box regression respectively; t_1 and t_2 are the weights of each task. Here, t_1 is set to 0.3 and t_2 is set to 0.7.

4. Experiments

The experimental environment used in this paper is as follows: Intel Xeon E-2136 CPU @3.30 GHz, 16 GB internal storage, Windows 10 64 bit operating system, Microsoft, US.

We tested the performance of our framework on different datasets composed of a self-collected dataset and Caltech Pedestrian dataset 18. The self-collected dataset was introduced in Section 1. The Caltech dataset was collected by the California Institute of Technology on 2012, and it is often used in the design and testing of pedestrian detection algorithms. It contains a video of a city environment with a duration of about 10 h, and the image resolution is 640×480. The pedestrian targets are divided into different levels according to the size and the occlusion.

Figure 2 shows PMCD could detect the face accurately when the pedestrians are in different environments containing occlusion, incomplete body, and poor light.

Our mechanism consists of SSD [32] and the pre-identification mechanism, which is mentioned in Section 3.2.2. The accuracy of face detection in our framework is directly related to the performance of the mechanism, so we tested our mechanism on the Caltech Pedestrian dataset. In order to prove that our mechanism is superior to other state-of-the-art methods, we drew the curve of false positive per image (FPPI)-miss rate by the evaluation method [54], as shown by:

$$FPPI = m/N, \tag{7}$$

where m is the number of false positive, N is the number of images.

$$\text{miss rate} = fa/(fa + tr), \tag{8}$$

where fa is the number of false negatives, tr is the number of true positives.

We compared our pre-identification mechanism with Viola-Jones (VJ) [24], histogram of oriented gradient (HOG) [48], Scale Aware (SA)-Fast RCNN [55] and region proposal network(RPN)+ boosted forests (BF) [56]. As shown in Figure 3, ours performs better than the other methods, and the detection results of the corresponding method are better when the miss rate is lower.

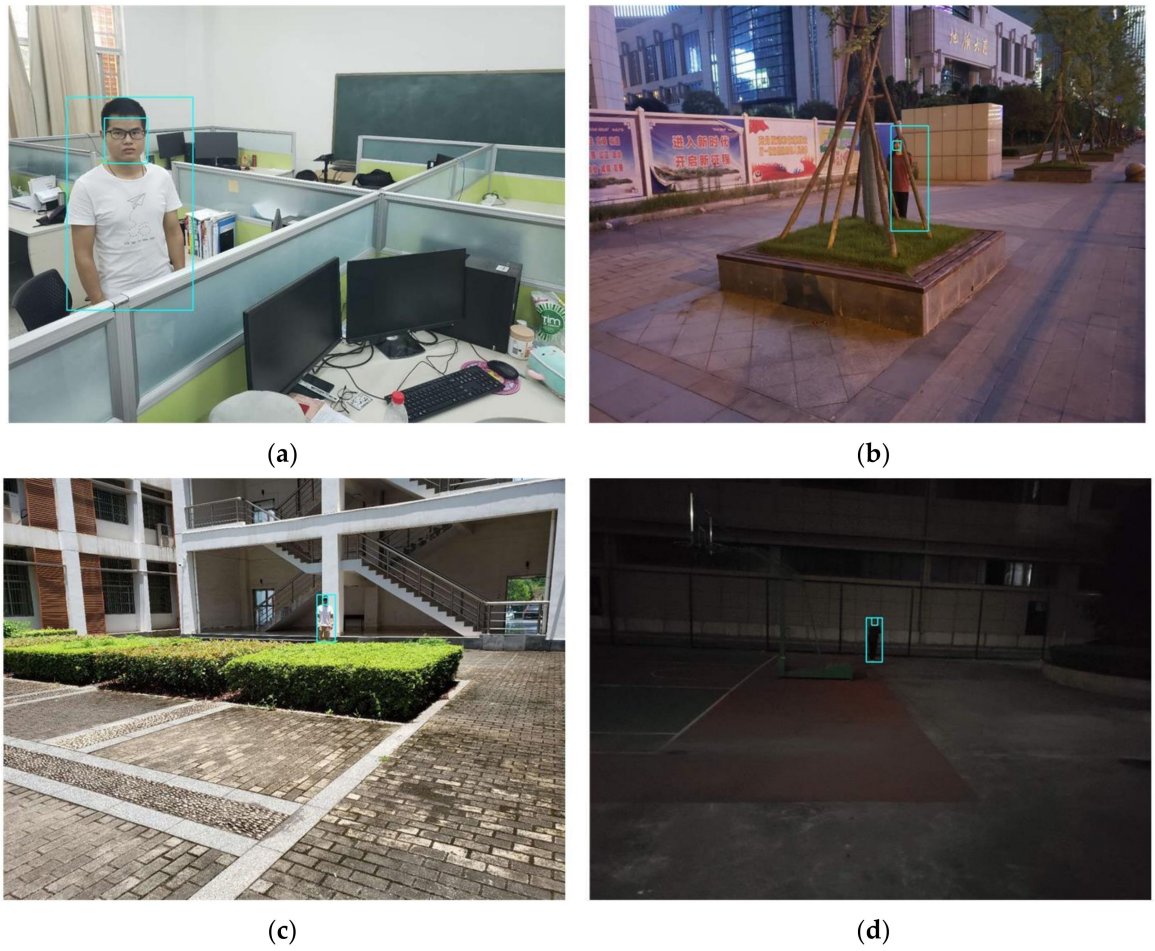

Figure 2. Face detection in different situations based on pre-identification mechanism and a cascaded detector (PMCD): (**a**) the face detected with occlusion on some part of the body; (**b**) the face detected with occlusion on some part of the body in complex backgrounds; (**c**) the tiny face detected with good light; (**d**) the tiny face detected with poor light.

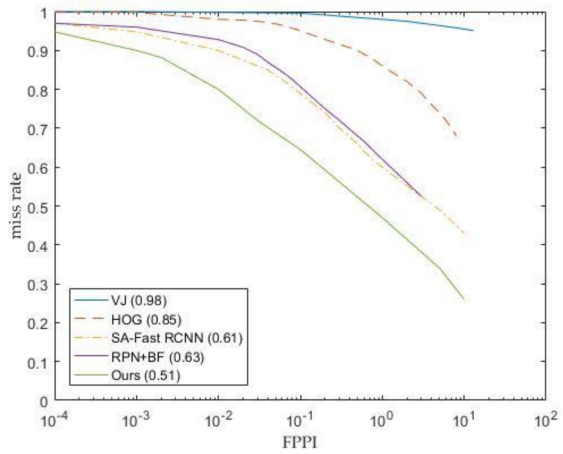

Figure 3. The curve of false positive per image (FPPI)-miss rate.

As shown in Figure 3, SSD achieves the best performance in these methods. Meanwhile, the miss rates of VJ and HOG are high, the main reason is that traditional methods usually train the classifiers by hand-crafted features, resulting in poorer robustness than that of convolutional neural networks.

We compared the whole framework with five of the most important face detection algorithms on the self-collected dataset, which are Viola-Jones(VJ) [24], single shot multibox detector(SSD) [32], multi-task cascaded neural networks (MTCNN) [45], multi-channel shallow convolutional network(MMSC) [46], and hybrid resolutions(HR) [47]. Figure 4 shows the results of the six methods in different situations. VJ and MTCNN cannot detect faces correctly when the face is fairly small. MMSC uses traditional methods to extract pre-processing regions, and thus it cannot keep good performance in detecting faces on incomplete bodies or in relatively complex background. SSD can detect faces roughly, but it cannot correctly obtain the bounding boxes when the faces are small. On one hand, HR can achieve remarkable performance in detecting tiny faces when the environment is relatively simple, but when the light is dim or faces are in a complex environment, the performance of HR would be very bad. On the other hand, PMCD achieves the best performance among the six methods. In the case of poor lighting conditions, only PMCD can achieve remarkable performance.

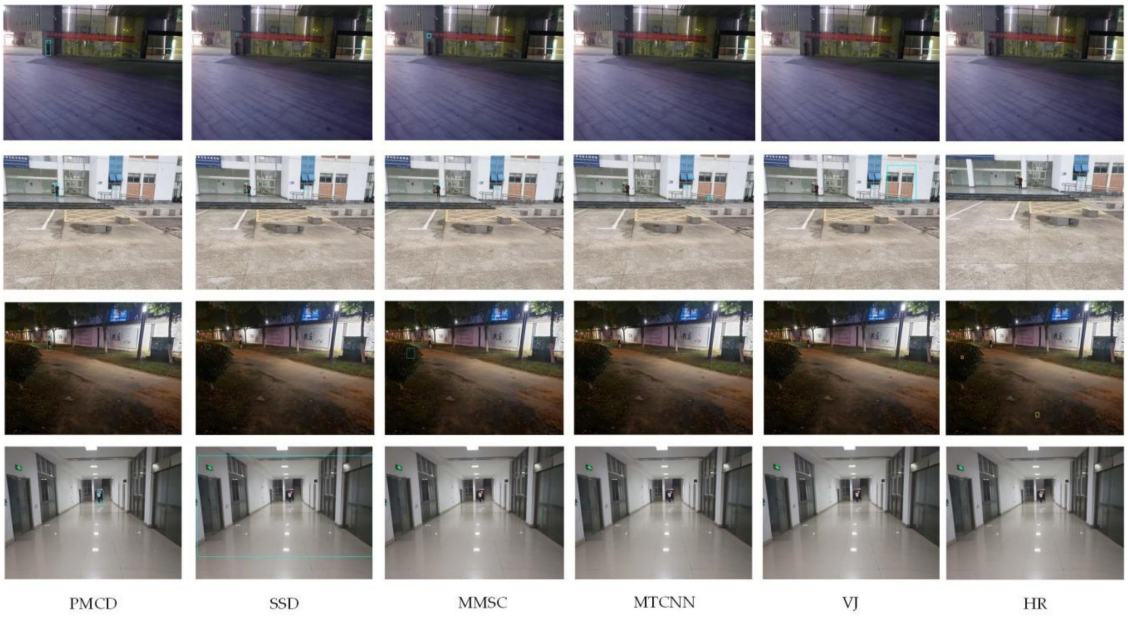

Figure 4. Face detection results.

Intersection over union (IoU) is the ratio between the intersection and the union of the candidate bounding box predicted by the model and the ground truth bounding box. The formula is:

$$\mathrm{IoU} = (\mathrm{area}(C) \cap \mathrm{area}(G))/(\mathrm{area}(C) \cup \mathrm{area}(G)) \tag{9}$$

where $\mathrm{area}(C)$ is the area of the candidate bounding box; $\mathrm{area}(G)$ is the area of the ground truth bounding box.

We collected the true positives' IoUs which are predicted by the above five methods and calculated the mean IoU (MIoU) to judge the accuracy of the predicted bounding boxes. The formula is shown as follows:

$$\mathrm{MIoU} = \sum_{i=1}^{n} \mathrm{IoU}_i / n, \tag{10}$$

where n is the total number of true positives; IoU_i is the IoU score of the i-th true positive.

The results are shown in Table 1. PMCD achieves the best performance in these six methods. The second best is HR, no matter how small the face is, HR would be detected very precisely when the light is good. However, when the face is in a dark environment, the face's bounding box will be detected with some deviation. While the performance of SSD is the lowest because SSD cannot retain its ability when the target is small even though it can detect tiny faces, because its default anchors are set to

different shapes, but the box shape with the highest score is not always the most suitable one. The other three methods can get great scores only if the targets are big.

Table 1. Mean intersection over union (MIoU) on the self-collected dataset.

	PMCD	HR	MMSC	MTCNN	SSD	VJ
MIoU	84.8%	83.4%	76.9%	79.2%	71.2%	73.6%

In order to verify the effectiveness of PMCD, the F1 measures, where precision and recall are equal, are shown in Table 2. The table of our model is higher than benchmark results of the other three important face detection neural networks. The MTCNN, VJ, and MMSC cannot keep high performance when targets are less than 20×20, which could result in many misclassifications. The framework proposed in this paper can solve this problem very well because we use two neural networks for grading detection, which can avoid the subjectiveness of hand-crafted features. In addition, the pre-identification mechanism can avoid the information loss of small targets, which often happens in a deep neural network.

Table 2. The F1 measures of different methods.

	VJ	MTCNN	MMSC	SSD	HR	PMCD
F1 measures	0.32	0.554	0.569	0.602	0.673	0.712

In addition, we tested the detection speed of these methods based on frames per second (FPS), and the results are shown in Table 3. Although PMCD is not the fastest method, it still exceeds 24 FPS, maintaining real-time performance. PMCD is designed as a cascade framework, in which some time is wasted in passing data, that is, the end-to-end detectors can outperform cascade frameworks in detection speed. Nevertheless, for the tiny-face detection task, the advantages of the cascaded detector are obvious.

Table 3. Detection speed of different methods.

	VJ	MTCNN	MMSC	SSD	HR	PMCD
FPS	1.2	32.8	3.9	43.2	2.6	34.6

5. Conclusions and Future Works

A real-time pre-identification mechanism and a cascaded detector were proposed for tiny-face detection in this paper. We combined pedestrian detection with face detection to reduce the search region. After obtaining the bounding box of the pedestrian, we estimated the face area as RoI based on the proportion of the size of the bounding box. The mechanism not only can improve efficiency but also reduce the number of false positives. After that, we built an image pyramid as the input of the face detector. The algorithm can ensure the performance of the face detector on the targets of different sizes. Based on the experimental results, the proposed cascaded neural network is more efficient in tiny-face detection than other state-of-the-art methods.

There are many open studies in the above research. Our method is constrained, but when the body is not completely occluded, PMCD can perform better in tiny-face detection under extreme conditions such as poor light and occlusion, etc. Traditional face detection methods and region-based CNNs cannot keep a high accuracy in tiny-face detection by themselves, but it is possible to combine these unconstrained methods with our proposed framework in the future. The highest priority should be given to this framework to avoid unnecessary false positives and reduce the miss rate of tiny-face detection.

As mentioned in Section 1, cascade frameworks have significant advantages in detecting tiny faces, and the experimental results also support this conclusion in Section 4. This is because the characteristics of tiny faces are gradually lost with convolutional layers going deeper. Therefore, this problem can be solved by finding a larger upper-body to complete the coarse-to-fine detection. PMCD has achieved remarkable results on our self-collected dataset when the targets are small or in extreme environments such as darkness, complex background, occlusion, etc. When the targets are of normal sizes, the end-to-end detection frameworks can obtain a faster speed. Therefore, PMCD still has potential for further improvement. In the future, we will design an end-to-end detection network based on PMCD, which can internally extract hierarchical features for the face detection task. This network can ensure satisfactory performance on tiny-face detection while keeping a faster detection speed than PMCD.

Author Contributions: All authors of the paper made significant contributions to this work. Z.Y. and J.L. contributed equally to this paper, conceived the idea of work, implemented algorithms, analyzed the experiment data, and wrote the manuscript. W.M. led the project, directed and revised the paper writing. Q.W. helped to code and analysis the experiment data.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 61762061, 61963027, and 61703198), the Natural Science Foundation of Jiangxi Province, China (Grant No. 20161ACB20004), Natural Science Foundation for Distinguished Young Scholars of Jiangxi Province (Grant No. 2018ACB21014), and Jiangxi Key Laboratory of Smart City (Grant No. 20192BCD40002).

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

References

1. Dang, L.M.; Hassan, S.I.; Im, S.; Lee, J.; Lee, S.; Moon, H. Deep Learning Based Computer Generated Face Identification Using Convolutional Neural Network. *Appl. Sci.* **2018**, *8*, 2610. [CrossRef]
2. Bai, W.; Quan, C.; Luo, Z. Uncertainty Flow Facilitates Zero-Shot Multi-Label Learning in Affective Facial Analysis. *Appl. Sci.* **2018**, *8*, 300. [CrossRef]
3. Kang, S.J. Multi-user identification-based eye-tracking algorithm using position estimation. *Sensors* **2016**, *17*, 41. [CrossRef] [PubMed]
4. Ma, L.; Deng, Z.G. Real-time hierarchical facial performance capture. In Proceedings of the Symposium on Interactive 3D Graphics and Games (ACM), New York, NY, USA, 21–23 May 2019. [CrossRef]
5. Weise, T.; Li, H.; Gool, L.V.; Pauly, M. Face/off: Live facial puppetry. In Proceedings of the SIGGRAPH/Eurographics ACM Symposium on Computer animation, New Orleans, LA, USA, 1–2 August 2009. [CrossRef]
6. Bouaziz, S.; Li, H.; Pauly, M. Realtime performance-based facial animation. *ACM Trans. Graph.* **2011**, *30*, 77.
7. Li, H.; Weise, T.; Mark, P.X. Example-based facial rigging. *ACM Trans. Graph.* **2010**, *29*, 32. [CrossRef]
8. Li, W.; Deng, Z.G. A practical model for live speech driven lip-sync. *IEEE Comput. Graph. Appl.* **2014**, *35*, 70–78.
9. Li, H.; Yu, J.H.; Ye, Y.T.; Bregler, C. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph* **2013**, *32*, 42. [CrossRef]
10. Ouzounis, C.; Kilias, A.; Mousas, C. Kernel projection of latent structures regression for facial animation retargeting. *arXiv* **2017**, arXiv:1707.09629.
11. Ma, L.; Deng, Z. Real-time Facial Expression Transormation for Monocular RGB Video. *Comput. Graph. Forum Wiley Online Libr.* **2019**, *38*, 470–481. [CrossRef]
12. Kaewmart, P.; Markus, B. The shape of the face template: Geometric distortions of faces and their detection in natural scenes. *Vis. Res.* **2015**, *109*, 99–106.
13. Ranjan, R.; Patel, V.M.; Chellappa, R. HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 121–135. [CrossRef] [PubMed]
14. Liao, Y.Q.; Xiong, P.W.; Min, W.D.; Min, W.Q.; Lu, J.H. Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. *IEEE Access* **2019**, *7*, 38044–38054. [CrossRef]
15. Min, W.D.; Fan, M.D.; Guo, X.G.; Han, Q. A new approach to track multiple vehicles with the combination of robust detection and two classifiers. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 174–186. [CrossRef]

16. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Wojek, C.; Dollar, P.; Schiele, B.; Perona, P. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761. Available online: http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/ (accessed on 18 March 2009).
19. Zou, F.Y.; Li, J.; Min, W.D. Distributed Face Recognition Based on Load Balancing and Dynamic Prediction. *Appl. Sci.* **2019**, *9*, 94. [CrossRef]
20. Yaman, M.A.; Subasi, A.; Rattay, F. Comparison of Random Subspace and Voting Ensemble Machine Learning Methods for Face Recognition. *Symmetry* **2018**, *10*, 651. [CrossRef]
21. Luo, Y.; Guan, Y.P. Adaptive skin detection using face location and facial structure estimation. *IET Comput. Vis.* **2017**, *11*, 550–559. [CrossRef]
22. Mohanty, R.; Raghunadh, M.V. A new approach to face detection based on YCgCr color model and improved AdaBoost algorithm. In Proceedings of the International Conference on Communication and Signal Processing, Melmaruvathur, India, 6–8 April 2016; pp. 1392–1396.
23. Ma, S.; Bai, L. A face detection algorithm based on Adaboost and new Haar-Like feature. In Proceedings of the IEEE International Conference on Software Engineering and Service Science, Beijing, China, 26–28 August 2016; pp. 651–654.
24. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
25. Liao, S.C.; Jain, A.K.; Li, S.Z. A fast and accurate unconstrained face detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 211–223. [CrossRef] [PubMed]
26. Yang, B.; Yan, J.J.; Lei, Z.; Li, S.Z. Aggregate channel features for multi-view face detection. In Proceedings of the IEEE International Joint Conference on Biometrics, Clearwater, FL, USA, 29 September–2 October 2014; pp. 1–8.
27. Bilal, M. Algorithmic optimisation of histogram intersection kernel support vector machine-based pedestrian detection using low complexity features. *IET Comput. Vis.* **2017**, *11*, 350–357. [CrossRef]
28. Baek, J.; Kim, J.; Kim, E. Fast and efficient pedestrian detection via the cascade implementation of an additive kernel support vector machine. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 902–916. [CrossRef]
29. Girshick, R.; Donahue, J.; Darrell, T. Region-based Convolutional Networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [CrossRef] [PubMed]
30. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Science, Wuhan, China, 20–22 November 2015; pp. 1440–1448. [CrossRef]
31. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
33. Jiang, H.Z.; Learned-Miller, E. Face detection with the Faster R-CNN. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, 30 May–3 June 2017; pp. 650–657.
34. Zhu, C.; Zheng, Y.; Luu, K.; Savvides, M. CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection. *arXiv* **2016**, arXiv:1606.054413.
35. Wan, S.; Chen, Z.; Zhang, T.; Zhang, B.; Wong, K.K. Bootstrapping face detection with hard negative examples. *arXiv* **2016**, arXiv:1608.02236.
36. Sun, X.; Wu, P.; Hoi, S.C.H. Face detection using deep learning: An improved faster RCNN approach. *arXiv* **2017**, arXiv:1701.08289. [CrossRef]
37. Zhang, C.; Xu, X.; Tu, D. Face detection using improved Faster RCNN. *Neurocomputing* **2018**, *299*, 42–50.
38. Hsu, G.S.; Hsieh, C.H. Cross-pose landmark localization using multi-dropout framework. In Proceedings of the IEEE International Joint Conference on Biometrics, Denver, CO, USA, 1–4 October 2017; pp. 390–396.
39. Li, J.; Xue, Y.; Wang, W.; Ouyang, G. Cross-level Parallel Network for Crowd Counting. *IEEE Trans. Ind. Inform.* **2019**. [CrossRef]

40. Triantafyllidou, D.; Nousi, P.; Tefas, A. Fast deep convolutional face detection in the wild exploiting hard sample mining. *Big Data Res.* **2018**, *11*, 65–76. [CrossRef]
41. Yang, S.; Luo, P.; Loy, C.C.; Tang, X. From facial parts responses to face detection: A deep learning approach. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 3676–3684.
42. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.
43. Qin, H.; Yan, J.; Li, X. Joint training of cascaded CNN for face detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3456–3465.
44. Zhang, J.; Shan, S.; Kan, M.; Chen, X. Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 1–16.
45. 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]
46. Min, W.D.; Fan, M.D.; Li, J.; Han, Q. Real-time face recognition based on face pre-identification detection and multi-scale classification. *IET Comput. Vis.* **2018**, *13*, 165–171. [CrossRef]
47. Hu, P.; Ramanan, D. Finding tiny faces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1522–1530.
48. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 886–893.
49. Xie, Y.; Yang, L.; Sun, X.; Wu, D.; Chen, Q.; Zeng, Y.; Liu, G. An auto-adaptive background subtraction method for Raman spectra. *Spectrochim. Part A Mol. Biomol. Spectrosc.* **2016**, *161*, 58–63. [CrossRef] [PubMed]
50. Han, J.; Quan, R.; Zhang, D.; Nie, F. Robust object co-segmentation using background prior. *IEEE Trans. Image Process.* **2018**, *27*, 1639–1651. [CrossRef] [PubMed]
51. Kim, G.; Yang, S.; Sim, J.Y. Saliency-based initialization of Gaussian mixture models for fully-automatic object segmentation. *Electron. Lett.* **2017**, *53*, 1648–1649. [CrossRef]
52. Chan, K. Segmentation of moving objects in image sequence based on perceptual similarity of local texture and photometric features. *EURASIP J. Image Video Process.* **2018**, 62. [CrossRef]
53. Park, K.; Kim, S.; Sohn, K. Unified multi-spectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognit.* **2018**, *80*, 143–155. [CrossRef]
54. Determe, J.F.; Louveaux, J.; Jacques, L.; Horlin, F. Improving the Correlation Lower Bound for Simultaneous Orthogonal Matching Pursuit. *IEEE Signal Proc. Lett.* **2016**, *23*, 1642–1646. [CrossRef]
55. Li, J.N.; Liang, X.D.; Shen, S.M.; Xu, T.F.; Feng, J.S.; Yan, S.C. Scale-Aware Fast R-CNN for Pedestrian Detection. *IEEE Trans. Multimed.* **2018**, *20*, 985–996. [CrossRef]
56. Zhang, L.; Lin, L.; Liang, X.; He, K. Is faster R-CNN Doing Well for Pedestrian Detection? *arXiv* **2016**, arXiv:1607.07032v2.