electronics

MDPI

Article

Micro-Expression Spotting Based on VoVNet, Driven by Multi-Scale Features

Jun Yang, Zilu Wu and Renbiao Wu *

Tianjin Key Laboratory for Advanced Signal Processing, Civil Aviation University of China, Tianjin 300300, China; junyang@cauc.edu.cn (J.Y.); 2021021098@cauc.edu.cn (Z.W.)
* Correspondence: rbwu@cauc.edu.cn; Tel.: +86-22-24092003

Abstract: Micro-expressions are a type of real emotional expression, which are unconscious and difficult to hide. Identifying these expressions has great potential applications in areas such as civil aviation security, criminal interrogation, and clinical medicine. However, because of their characteristics such as short duration, low intensity, and sparse action units, this makes micro-expression spotting difficult. To address this problem and inspired by object detection methods, we propose a VoVNet-based micro-expression spotting model, driven by multi-scale features. Firstly, VoVNet is used to achieve the extraction and reuse of different scale perceptual field features to improve the feature extraction capability. Secondly, multi-scale features are extracted and fused using the Feature Pyramid Network module, incorporating optical flow features, and by realizing the interactive fusion of fine-grained feature information and semantic feature information. Finally, the model is trained and optimized on CAS(ME)^2 and SAMM Long Video. The experimental results show that the F1 score of the proposed model is improved by 0.1963 and 0.2441 on the two datasets compared with the baseline method, which outperforms the most popular spotting methods.

Keywords: micro-expression spotting; multi-scale; optical flow

Citation: Yang, J.; Wu, Z.; Wu, R. Micro-Expression Spotting Based on VoVNet, Driven by Multi-Scale Features. *Electronics* **2023**, *12*, 4459. https://doi.org/10.3390/electronics12214459

Academic Editor: Andrea Asperti

Received: 10 October 2023
Revised: 19 October 2023
Accepted: 26 October 2023
Published: 30 October 2023

1. Introduction

Facial expressions are mainly divided into micro-expressions and macro-expressions [1]. Micro-expressions are facial expressions that are unconsciously revealed by humans. When micro-expressions occur, the changes to the facial features are insignificant, mainly characterized by short duration, low intensity, and sparse facial action units. However, compared with macro-expressions, they can realistically reveal people's emotions. Therefore, micro-expression research has a large application value in the fields of civil aviation security screening [2], criminal interrogation [3], and clinical medicine [4].

Research related to micro-expressions is divided into two categories: micro-expression spotting and micro-expression recognition. Micro-expression spotting refers to locating the clips of micro-expressions in a video. Micro-expression recognition refers to the classification of a detected micro-expression slice, and then the classification results are applied to different clips. Micro-expression recognition research is more mature, while micro-expression spotting is still in the preliminary research stage. In this paper, inspired by object detection methods, a multi-scale feature fusion method is applied to micro-expression spotting to improve the accuracy of the micro-expression spotting model.

The object detection method consists of three parts: backbone, neck, and head [5]. The backbone part serves to perform feature extraction, where high quality features retain more information in the image and make subsequent detection more accurate. The neck part serves to perform feature fusion, which aims to fuse different features and enrich the feature connotation. The head part is used to predict the results, such as the location and the classification of the result. Similar to object detection, in micro-expression spotting, facial features are extracted in the backbone part, the extracted features are fused in the

neck part, and finally the location and category of the micro-expressions are output in the head part.

In feature extraction, a simple backbone network will ignore the detailed information of micro-expressions, while a complex backbone network will slow down the model's speed. Therefore, we use VoVNet [6] for feature extraction and to concatenate the feature maps of each layer in the end, which not only achieves feature reuse and improves feature extraction capability, but also reduces the model parameters and improves the model's training speed. Different levels of feature maps have different focuses. The shallow feature maps mainly reflect the content, such as the light and dark of an image; the deep feature maps express the overall structural information. Therefore, in the neck part, the Feature Pyramid Network (FPN) [7] is applied to fuse the deep-level features with the shallow-level features, making the whole feature richer.

The optical flow method is widely used in the computer field, which can provide key information for many vision tasks and help to improve performance. As an important method in computer vision and image processing, optical flow can provide dynamic features about facial motion, combining temporal and spatial information to improve the accuracy of micro-expression spotting. At the same time, the optical flow method is robust to common factors such as facial occlusion, illumination change, and noise. It enables the network to be more able to deal with various interference factors in the actual scene. Optical flow is incorporated into the features, and the motion information is extracted by analyzing the pixel changes between consecutive frames, which can better capture the change in micro-expressions.

The organizational structure of this paper is as follows: the first section introduces the background and significance of this paper; the second section introduces the current status of micro-expression spotting; the third section, the micro-expression spotting based on VoVNet, driven by multi-scale features is introduced in detail; the fourth section shows the analysis of the experimental results; and the fifth section summarizes the work of this paper.

2. Related Work

In the early stage of micro-expression spotting, the algorithms were mainly focused on traditional methods. Shreve [8] calculated the optical flow from the onset frame to each frame of the video sequence and determined the micro-expression interval according to the magnitude of the optical flow change. Moilanen [9] used a local binary pattern (LBP) to analyze the feature difference between consecutive frames for micro-expression spotting. Patel [10] computed optical flow over local spatial regions and used a heuristic algorithm to filter out non-micro-expressions. This could detect the onset frame, the vertex frame, and the offset frame. Li [11] proposed using a local temporal pattern (LTP) and a local binary pattern (LBP) for micro-expression spotting and used them as the benchmark for the Micro-Expression Spotting Challenge, 2019. Later, He Y [12] proposed the MDMD method using the maximum difference of optical flow features to detect micro-expressions.

In recent years, deep learning methods have been widely used in various fields, and more and more experts and scholars are exploring the use of deep learning methods for micro-expression spotting and recognition. Xia [13] applied machine learning to micro-expression spotting and considered the relationship between frames and used adaboost to predict the probability of a certain frame as a micro-expression. Hong [14] used a sliding window to detect micro-expressions in samples with a fixed number of frames and treated micro-expression spotting as a binary classification task. Nag [15] proposed a joint architecture of temporal and spatial information to detect the onset frame and offset frame of micro-expressions. Verburg M [16] applied the computed HOOF features into a recurrent neural network (RNN) for micro-expression localization, which combined deep learning and traditional methods and applied them to micro-expression spotting. Pan et al. [17] proposed putting each frame of a video into the local bilinear convolutional neural network (LBCNN) to judge whether each frame belonged to a micro-expression, a macro-expression, or a nat-

ural expression. Yap et al. [18] proposed a 3D-CNN model that compared each frame with a reference frame, which is a pure deep learning scheme. Liong et al. [19] proposed a shallow optical flow three-stream CNN (SOFTNet), which used different optical flow components in three channels to capture different motion information. Fang Y [20] used the phase calculated by the Riesz Pyramid to represent motion and used CNN to calculate the probability that each frame is a micro-expression. Many of these micro-expression spotting methods draw on the idea of micro-expression recognition to judge whether a clip or a certain frame in a video is a micro-expression. This is essentially a classification problem and does not locate the clips in the video where the micro-expression occurs. Li J et al. [21] first introduced the self-supervised learning method into the construction of the micro-expression spotting model. By using auxiliary tasks in a large number of unsupervised videos, a model with temporal and spatial features of micro-expressions was constructed. Cao [22] designed a micro-expression spotting framework based on outlier spotting. Song [23] proposed a BERT network-based micro-expression spotting algorithm composed of candidate fragment generation, a spatio-temporal feature extraction module, and a grouping module.

Object detection methods based on deep learning are widely used in areas such as facial detection [24], pedestrian detection [25], and license plate detection [26]. Inspired by object detection, some scholars began to apply the methods of object detection in the spatial dimension to micro-expression spotting in the temporal dimension. For example, Yu et al. proposed using the detection method for micro-expression spotting and achieved good results in the Facial Micro-Expression (FME) Challenge. In this paper, we draw on the methods and ideas of object detection to carry out micro-expression spotting research and propose a VoVNet-based micro-expression spotting method driven by multi-scale features.

3. Proposed Method

3.1. Micro-Expression Spotting Method Based on VoVNet

Although micro-expressions are short in duration, there is still a process of facial change. Here, we define the starting point where the micro-expression occurs as the onset frame, the frame where the micro-expression changes most significantly as the apex frame, and the offset frame of the micro-expression as the offset frame. The main task of micro-expression spotting is to locate the apex frame and offset frame of the micro-expression. Figure 1 shows the structure of the micro-expression spotting model. Firstly, the micro-expression samples and the corresponding optical flow are concatenated and input into the VoVNet for feature extraction. Secondly, the fusion of the extracted features is performed by the FPN module. Finally, the micro-expression spotting results are output. In the whole process, feature extraction and fusion play a key role in ensuring the accuracy of the micro-expression spotting. Figure 1 shows the network structure of the micro-expression spotting.

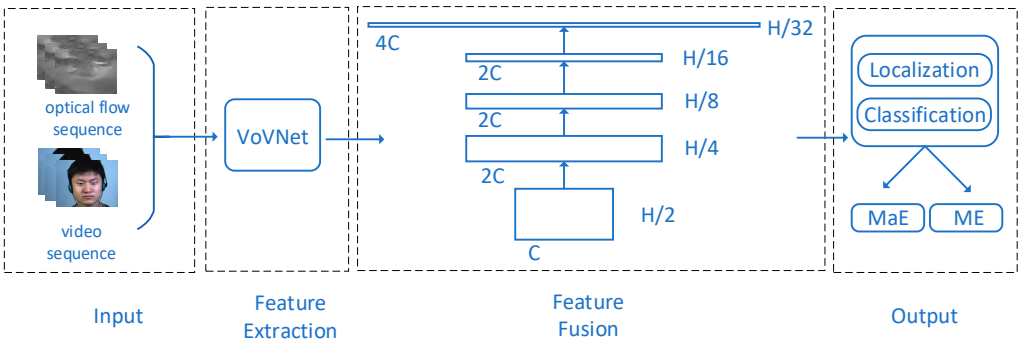

Figure 1. Structure of the micro-expression spotting model.

3.2. Optical Flow

Optical flow is the displacement of pixels due to the motion of objects in a continuous sequence of image frames. Calculating the optical flow between different frames can obtain the motion trajectory of the object in the image sequence because the micro-expressions occur as tiny details and are not easy to find; however, optical flow has good performance for the estimation of motion in a small range. Tiny movements in specific areas of the face can be detected by calculating optical flow. Micro-expressions are continuous actions, and optical flow can extract rich features from continuous image frames and capture the temporal correlation of local areas in the image. Compared with static images, optical flow can provide dynamic change information and capture the motion information of an image sequence. By combining optical flow features with raw video, micro-expressions can be spotted more accurately.

The optical flow method is based on three assumptions: (1) that the illumination remains constant between two frames; (2) that the motion of the same pixel between two frames is small; and (3) that the motion of adjacent pixels is similar. Let I(x,y,t) be the brightness value at the position (x,y) at time t, and the distance the pixel moves in dt time be (dx,dy). Because the brightness value between two frames is unchanged, we can achieve Equation (1).

$$I(x, y, t) = I(x + dx, y + dy, t + dt), \tag{1}$$

Equation (1) is expanded by the Taylor series, and Equation (2) is obtained by removing the general terms and dividing by dt.

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0, \tag{2}$$

If p,q are the horizontal and vertical directions of the pixel to obtain the velocity component then:

$$\mathrm{p} = \frac{dx}{dt}\ ,\ \mathrm{q} = \frac{dy}{dt} \tag{3}$$

By bringing p and q into Equation (2), the optical flow change of each pixel of the picture can be obtained.

According to Liong S T et al. [27], the TVL1 optical flow method is more robust and accurate than other methods in the study of micro-expression. Therefore, this article also uses the TVL1 method.

3.3. VoVNet Module

Related studies have proven that features with multiple receptive fields can capture richer visual information [28–30]. Since the features are inconspicuous when micro-expressions occur, they are mainly manifested in the weak intensity of facial muscle changes and sparse facial action units. Therefore, to improve the extraction capability of micro-expression features, VOVNet is used. By fusing the features of different receptive fields, VoVNet can extract the relevant features of the long-range facial action unit and improve the performance of micro-expression spotting. VoVNet is mainly composed of One-Shot Aggregation (OSA) modules, as shown in Figure 2. The OSA module consists of multiple convolution layers, each of which is bi-directionally connected. One is used to connect to the next convolution layer to generate features with a larger receptive field, and the other is used to connect to the last layer to achieve feature splicing and reuse. This structure is designed to enhance the feature extraction capability of the network by fusing features with different receptive fields. It does not cause redundancy of features and improves the efficiency of the model.

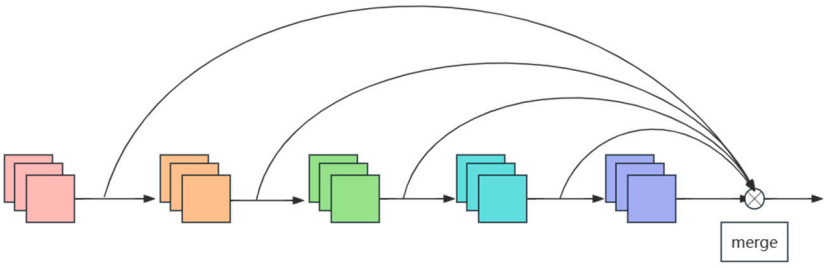

Figure 2. OSA module structure.The convolved feature graphs are changed and finally fused.

Memory Access Cost (MAC) is an important measure of model processing speed. VoVNet is not only strong in feature extraction, but also fast in computation, mainly because of the small MAC. The MAC is calculated as shown in Equation (4). Let the number of input and output channels of one convolution layer of the OSA module be c_1 and c_2, respectively, and the size of the feature map is h × w; then, the FLOPs of 1 × 1 convolution is B = h × w × c_1 × c_2. Equation (5) is derived from the mean inequality and the MAC is minimized when $c_1 = c_2$. Therefore, when the number of input and output channels in the middle layer of the OSA module is the same, the model MAC is minimized and the model processing speed is fastest.

$$\mathrm{MAC} = \mathrm{h} \times \mathrm{w} \times (c_1 + c_2) + c_1 \times c_2 \,, \tag{4}$$

$$\mathrm{MAC} \geq 2\sqrt{h \times w \times B} + \frac{B}{h \times w,} \tag{5}$$

VoVNet consists of three convolution layers and four OSA modules. Each OSA module consists of five convolution layers with the same input and output channels to minimize the value of the MAC. The number of feature channels is gradually increased by superimposing multiple OSA modules, and the superposition of feature maps of different sizes makes the features more abundant. Feature fusion is performed after feature extraction. The specific network structure is shown in Table 1.

Table 1. VoVNet network architecture.

Type	VoVNet
Inception	3 × 3conv, 64, stride = 2 3 × 3conv, 128, stride = 2
OSA 1	3 × 3conv, 64, ×5 concat: 1 × 1conv, 128
OSA 2	3 × 3conv, 80, ×5 concat: 1 × 1conv, 256
OSA 3	3×3conv, 96, ×5 concat: 1 × 1conv, 384
OSA 4	3 × 3conv, 112, ×5 concat: 1 × 1conv, 1

3.4. FPN Module

In object detection, feature fusion can effectively improve model performance and generalization ability. FPN, as a common feature fusion module in object detection, improves detection accuracy by constructing a feature pyramid structure, extracting and fusing multi-scale features from different levels.

The feature pyramid consists of multiple levels, each with a different resolution and receptive field. This design enables the model to analyze and process micro-expression sequences at different scales. The bottom pyramid layers contain features at lower levels,

while the features at higher levels are more abstract and semantic. Combining these features can obtain more comprehensive and accurate features, which can help to improve the accuracy of micro-expression detection. Each level will extract the features of the micro-expression sequence, which can improve the ability to understand and analyze the micro-expression sequence by transferring information. Low-level features can provide background and global information for higher-level features so that the model can better detect micro-expressions. Because features at different levels obtain information at different scales, they are robust in detection.

The pyramid model can obtain feature maps of different scales through one-dimensional convolution, and rich multi-scale and multi-level feature representations can be obtained through multi-level feature extraction and combination. Thus, the accuracy of micro-expression spotting is improved. The specific network structure is shown in Table 2.

Table 2. FPN network architecture.

Type	FPN
layer 1	3 × 1conv, 512 × 64, stride = 2
layer 2	3 × 1conv, 1024 × 32, stride = 2
layer3	3 × 1conv, 1024 × 16, stride = 2
layer4	3 × 1conv, 1024 × 8, stride = 2
layer5	3 × 1conv, 2048 × 4, stride = 2

3.5. Loss Function

The loss function can calculate the difference between the predicted result and the true label. The lower the loss value, the stronger the ability of micro-expression training. Classification loss, boundary frame loss, and IOU loss are calculated, respectively, in micro-expression spotting.

The first is classification loss, which is a measure of the ability to classify the target species. Because there are more macro-expressions than micro-expressions in the video sample, focal loss [31] is used to solve the category imbalance problem. Focal loss introduced w and pt to adjust the weights of samples to be unequal, reduce the emphasis on easily identifiable samples, and increase the emphasis on samples that are difficult to classify. The alph is a balancing parameter standing at 0.25 and w changes the weight of the sample. The proportion of different samples in the equation is different, which makes the model pay more attention to the small number of samples. Pt reduces the weight of samples that are easy to classify and increases the weight of samples that are difficult to classify by calculating probabilities. This makes the model pay more attention to those samples that are difficult to classify, improving the model's learning ability for difficult samples. Equation (6) is the calculation equation for focal loss.

$$\text{Focal\_loss} = -\text{w}(1-pt)^2 \times log(pt), \tag{6}$$

$$\text{pt} = \begin{cases} p & positive\ sample \\ 1-p & negative\ sample' \end{cases} \tag{7}$$

$$\text{w} = \begin{cases} alph & positive\ sample \\ 1-alph & negative\ sample\ ' \end{cases} \tag{8}$$

In addition to classification losses, positioning losses are used to measure the difference between the predicted bounding box and the true bounding box. This difference is optimized to better regulate the location of the predicted bounding box. When the absolute difference between the predicted value and the target value is large, the smooth L1 loss function adopts the square function, and the loss growth rate slows down. It is more

robust in the face of outliers and large error boundary boxes. The equation for the smooth L1 loss function is as follows:

$$\text{L1\_loss} = \begin{cases} 1/2x^2 & t < 1 \\ |x| - 0.5 & t \geq 1 \end{cases}, \tag{9}$$

Finally, IOU loss is a simple and intuitive method to calculate the overlap between the predicted bounding box and the real bounding box. It is not affected by the shape and size of the target and only considers the overlap degree of the two bounding boxes, which is suitable for different targets. By minimizing the IOU loss, the model parameters can be optimized to make the predicted bounding box closer to the real bounding box. Let the left abscissa of the predicted and true bounding boxes be x_{p1} and x_{t1}, and the right abscissa of the bounding boxes be x_{p2} and x_{t2}. Equation (10) is the equation of IOU loss function.

$$\begin{gathered} \text{IOU\_loss} = \text{inter}/\text{union}, \\ \text{Inter} = \min(x_{t2}, x_{p2}) - \max(x_{p1}, x_{t1}, \\ \text{Union} = (x_{p2} - x_{p1}) + (x_{t2} - x_{t1}) - \text{inter}, \end{gathered} \tag{10}$$

$$\text{Loss} = \text{Focal\_loss} + \text{L1\_loss} + \text{IOU\_loss}, \tag{11}$$

The combination of classification loss, positioning loss, and IOU loss can comprehensively evaluate the performance of object detection. Classification loss is used to evaluate the accuracy of the model for target classification. Positioning loss is used to assess the accuracy of the model for the target position. IOU loss assesses the accuracy of the boundary box matching. Equation (11) is the calculation equation of the final loss function. Target spotting usually requires the accurate classification of targets and the accurate location of targets. Combining these loss functions can simplify the model training process. It can improve the stability and convergence of training, and reduce the difficulty of hyper-parameter adjustment so that the model has the ability to perform classification and positioning at the same time.

4. Experiment

4.1. Dataset

Currently, the available micro-expression datasets are very limited and differ in resolution, frame rate, and generation methods. Authoritative datasets that have been released mainly include CASME [32], SMIC [33], CASME II [34], SAMM [35], CAS(ME)^2 [36], SAMM Long Videos [37], MMEW [38], and CASME III [39]. The CASME, CASME II, SMIC, and SAMM only contain micro-expression samples, while the CAS(ME)^2, CASME III, SAMM Long Videos, and MMEW contain not only micro-expression video samples but also macro-expression video samples. CAS(ME)^2 was released by the Chinese Academy of Sciences in 2018. The subjects of CAS(ME)^2 are 22 Asians, and the data are divided into two parts: part A and part B. Part A includes 87 long videos of micro-expressions and macro-expressions. Part B includes 300 cropped macro-expression samples and 57 cropped micro-expression samples. The average duration of each video is 148 s. CASME III manually labeled 1030 micro-expressions and 2264 macro-expressions. The SAMM Long Videos are extended from the SAMM and include a total of 147 long videos. Compared with CAS(ME)^2, SAMM Long Videos have a longer video duration with higher resolution and frame rates. The MMEW was released in 2021 and contains 300 micro-expression video samples and 900 macro-expression video samples. Table 3 shows the details of the commonly used datasets of macro-expressions and micro-expressions.

CAS(ME)^2 and SAMM Long Videos were used in the Facial Micro-Expression (FME) Challenge [40] to validate the micro-expression spotting model. Therefore, in order to ensure the comparability of the results, CAS(ME)^2 and SAMM Long Videos were also selected as the micro-expression spotting dataset in this paper.

Table 3. Micro-expression dataset details.

Dataset		Time	Resolution	Frame Rate	Number of Participants	Number of Samples	Number of Emotions
CAS(ME)^2		2018	640 × 480	30	22	300 (macro) 57 (micro)	4
SAMM Long Videos		2019	2040 × 1088	200	29	343 (macro) 159 (micro)	/
MMEW		2021	1920 × 1080	90	36	900 (macro) 300 (micro)	7
CASME III	A B C	2022	1280 × 720	30	100 116 31	3364 (macro) 1030 (micro)	7

The dataset is divided into two categories: micro-expression and macro-expression. Micro-expressions are extremely brief and tiny changes in human facial expressions, typically lasting between 1/25 and 1/5 of a second. These small facial changes are often very rapid and imperceptible and often occur when people are trying to mask or hide their true feelings. Macro-expression is relative to micro-expression, which refers to the expression changes that are more significant and last longer. Whether micro-expression or macro-expression, both are expressions of the human face in different emotional or psychological states, and all involve the movement and change of the facial muscles. Figure 3 shows micro-expression and macro-expression samples in the SAMM Long Videos dataset. The human eye is difficult to distinguish, and a computer is needed for recognition.

(**a**) (**b**)

Figure 3. (**a**) MMEW dataset macro-expression sample; (**b**) MMEW dataset micro-expression sample.

4.2. Experimental Environment and Hyper-Parameters

The configurations of the computer used for training and validation of the micro-expression spotting model are as follows:

(1) Operating system: 64-bit Ubuntu16.04.1.;
(2) Development environment: PyTorch1.2.0.;
(3) CPU: Intel® Xeon(R) Gold 5218R CPU @ 2.10 GHz × 46;
(4) GPU: Quadro RTX5000;
(5) Memory: 128 GB.

The hyper-parameters of the micro-expression spotting model are as follows:

(1) Optimizer: Adam;
(2) Learning rate: 0.005;

(3) Batch size: the batch size of CAS(ME)^2 is 32, and the batch size of SAMM Long Videos is 2.

4.3. Evaluation Metrics

Intersection over Union (IOU) [41] is used in object detection. IOU is the intersection of the predicted box and the real box divided by their union. When the value of IOU is greater than a certain threshold, it proves that the target is correctly boxed. Equation 12 shows the equation for micro-expression spotting IOU. Where, $W_{spotted}$ is the micro-expression clips obtained by the micro-expression spotting model, $W_{groundTrut}$ is the clips of the real micro-expression, k is the threshold of IOU, which is generally set to a constant. When the intersection of $W_{spotted}$ and $W_{groundTrut}$ divided by their union is greater than k, it proves that the micro-expression clips are detected correctly.

$$\frac{W_{spotted} \cap W_{groundTrut}}{W_{spotted} \cup W_{groundTrut}} \geq k, \tag{12}$$

As shown in Figure 4, AC are the clips where a micro-expression occurs and BD are the clips detected by the micro-expression model. BC is the intersection of $W_{spotted}$ and $W_{groundTrut}$, AD is the union of $W_{spotted}$ and $W_{groundTrut}$.

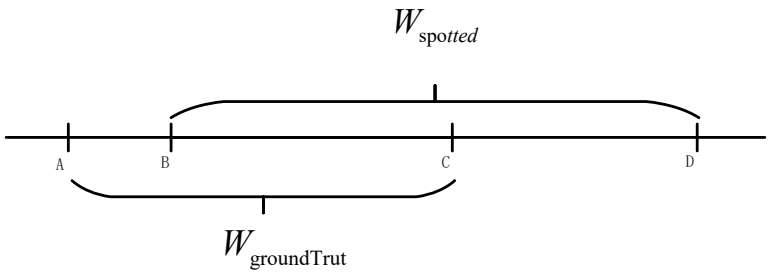

Figure 4. Micro-expression spotting IoU.

The micro-expression spotting performance is evaluated with an F1 score. The equation of the F1 score is shown in Figure 4. TP is the number of positive samples correctly identified. FP is the number of negative samples predicted as positive samples, i.e., the number of false detections. TN is the number of negative samples correctly identified. FN is the number of positive samples detected as negative samples, i.e., the number of missed detections. There are two main reasons for using the F1 score as the evaluation metric of the micro-expression spotting model:

(1) If there is no micro-expression in a single video or no micro-expression is detected in the video, the denominator of recall or precision will be 0. Using the F1 score as the evaluation metric will avoid this situation;

(2) Since the databases are apparently unbalanced, the sample size of micro-expressions is smaller than that of the macro-expressions. An F1 score will give us a fair evaluation of how well the model performs on all the classes rather than biasing only a few certain classes [36].

$$\text{F1-score} = 2 \times \frac{recall \times precision}{recall precision} = \frac{2TP}{2TPFPFN,} \tag{13}$$

$$\text{Recall} = \frac{TP}{TPFN,} \tag{14}$$

$$\text{Precision} = \frac{TP}{TPFP} \tag{15}$$

4.4. Results and Discussion

Following [1], we use Leave-One-Subject-Out (LOSO) cross-validation to report the performance on micro-expression spotting. One micro-expression video sample is taken as the test set and the remaining samples are used as the training set.

The loss value of a model training can intuitively measure the quality of model training. Figure 5 shows the change of loss value when the model is trained on two datasets, respectively. From Figure 5, it can be seen that the loss value of the model decreases and converges as the number of iterations increases, and finally converges to a smooth state. This indicates that the model can reach a smooth convergence state.

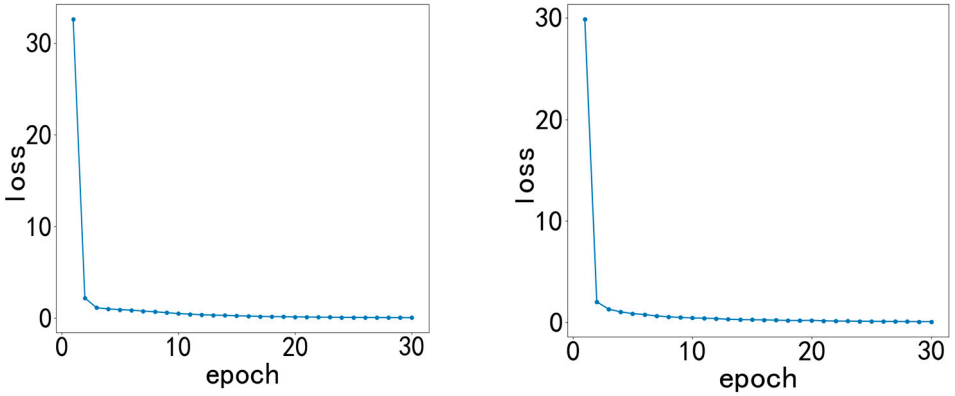

Figure 5. (**a**) Describes the changes in loss for the SAMM Long Video dataset; (**b**) describes the changes in loss for the CAS(ME)^2 dataset.

To evaluate the model performance, the proposed method is compared with the baseline method of the Facial Micro-Expression (FME) Challenge, traditional methods, and deep learning methods. Table 4 provides a detailed comparison of the models. With the SAMM Long Video, compared with the baseline method, the performance of the proposed method is significantly improved for both macro- and micro-expression spotting, with an improvement of 0.1986 and 0.1217 in the F1 score, respectively. With the CAS(ME)^2, compared with the baseline method, the proposed method improves the F1 score of macro- and micro-expression spotting by 0.2597 and 0.0571. In the overall performance of both macro- and micro-expression spotting, the method proposed in this paper outperforms the most popular spotting methods, such as MDMD, STCAN, and SOFTNet.

Table 4. Experimental results.

Model	SAMM Long Video			CAS(ME)^2		
	MaE	**ME**	**Overall**	**MaE**	**ME**	**Overall**
Baseline [11]	0.1863	0.0409	0.1193	0.0401	0.0118	0.0304
MDMD [12]	0.0629	0.0364	0.445	0.1196	0.0082	0.0376
STCAN [42]	0.1469	0.0125	0.1257	0.1250	0.0250	0.1168
SOFTNet [19]	0.2169	0.1520	0.1881	0.2410	0.1173	0.2022
Article	0.3849	0.1626	0.3156	0.2998	0.0689	0.2745

4.5. Ablation Experiment

The ablation experiment in this experiment verifies the influence of different modules on the model and reflects the superiority of the network. We perform replacement experiments on the optical flow module and feature extraction separately.

Optical flow features provide temporal dimension information for micro-expression detection. We compare the results of only the video features with the results of adding optical flow features. In the SAMM Long Video dataset, the F1 score of the method with an optical flow module increased by 0.1986 for macro-expression spotting and 0.2597 for micro-expression spotting. In the CAS(ME)^2 dataset, the F1 score of the method with the optical flow module increased by 0.0115 for macro-expression spotting and 0.0345 for micro-expression spotting. The results show that optical flow improves the micro-expression detection ability. Table 5 shows the effect of the optical flow module.

Table 5. Optical flow module ablation experiment.

Type	SAMM Long Video			CAS(ME)^2		
	MaE	ME	Overall	MaE	ME	Overall
No optical flow	0.2500	0.0696	0.2162	0.2883	0.0344	0.2620
Optical flow	0.3849	0.1626	0.3156	0.2998	0.0689	0.2745

The second stage is feature extraction, which is compared with the feature extraction network with better performance in target detection. ResNet shows good performance in target detection [43]. Residual links help train deeper networks in object detection. They also help solve degradation problems and are more stable in model performance. Dense connectivity in DenseNet enables features from each layer to interact directly with subsequent layers, enabling feature reuse. It can improve the detection performance. ResNet and DenseNet can extract richer and more meaningful features, and their structure is similar to VoVNet. Therefore, VoVNet is compared with ResNet and DenseNet.

A SAMM Long Video sample occupies a large storage space, and DenseNet training requires a large number of parameters. Due to the limitation of server GPU memory, we only used the CAS(ME)^2 dataset for progressive ablation experiments. Table 6 shows the results of VoVNet compared with other models in detail. The results show that VoVNet has the best feature extraction ability in micro-expression detection.

Table 6. Feature extraction module ablation experiment.

Model	CAS(ME)^2		
	MaE	ME	Overall
ResNet	0.2495	0.0421	0.2253
DenseNet	0.3026	0.0459	0.2596
VoVNet	0.2998	0.0689	0.2745

5. Discussion

To address the problem that micro-expressions are difficult to detect, a VoVNet-based micro-expression spotting model driven by multi-scale features is proposed in this paper. VoVNet is used for feature extraction; it integrates the features of different receptive fields to improve the model's performance. The FPN model is used in feature fusion to fuse features of different sizes and achieve deep fusion of fine-grained and semantic features, which reduces the loss of feature information and improves model robustness. Finally, the LOSO cross-validation is used to evaluate the performance of the model. The experimental results show that compared with other popular methods, the micro-expression spotting method proposed in this paper can improve the performance of micro-expression spotting to a certain extent. In addition to micro-expression spotting, the method proposed in this paper can also be applied to video behavior recognition tasks, such as abnormal behavior detection, action recognition, and gesture recognition in surveillance videos. It can also be applied to medical image processing, such as lesion detection, disease classification, and

diagnosis. By extracting and analyzing the features in medical images, it can assist doctors in the diagnosis and treatment of diseases.

Since micro-expression and macro-expression samples are not balanced, we used attention mechanisms and other methods to compensate for this deficiency. Due to the large dataset and many model parameters, our next step will explore the use of a lighter model for feature extraction. What is more, the importance of the three loss functions will be considered. These losses can be weighted, and the weights can be optimized to take advantage of the results.

Author Contributions: Methodology, J.Y.; validation, Z.W.; investigation, J.Y.; writing—original draft preparation, Z.W.; writing—review and editing, Z.W and R.W.; capital, R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "the Open Fund of Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China", grant number 2022ASP-TJ03, and supported by "the Fundamental Research Funds for the Central Universities of China", grant number 3122023011.

Data Availability Statement: Not applicable.

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

References

1. Yu, W.W.; Yang, K.F.; Yan, H.M.; Li, Y.J. Weakly-supervised Micro-and Macro-expression Spotting Based on Multi-level Consistency. *arXiv* **2023**, arXiv:2305.02734.
2. Weinberger, S. Airport security: Intent to deceive? *Nature* **2010**, *465*, 412–416. [CrossRef] [PubMed]
3. Owayjan, M.; Kashour, A.; Al Haddad, N.; Fadel, M.; Al Souki, G. The design and development of a lie detection system using facial micro-expressions. In Proceedings of the 2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), Beirut, Lebanon, 12–15 December 2012; pp. 33–38.
4. Russell, T.A.; Green, M.J.; Simpson, I.; Coltheart, M. Remediation of facial emotion perception in schizophrenia: Concomitant changes in visual attention. *Schizophr. Res.* **2008**, *103*, 248–256. [CrossRef] [PubMed]
5. Yu, W.W.; Jiang, J.; Li, Y.J. LSSNet: A two-stream convolutional neural network for spotting macro-and micro-expression in long videos. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, China, 20–24 October 2021; pp. 4745–4749.
6. Lee, Y.; Hwang, J.; Lee, S.; Bae, Y.; Park, J. An energy and GPU-computation efficient backbone network for real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
7. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
8. Shreve, M.; Godavarthy, S.; Manohar, V.; Goldgof, D.; Sarkar, S. Towards macro-and micro-expression spotting in video using strain patterns. In Proceedings of the 2009 Workshop on Applications of Computer Vision (WACV), Snowbird, UT, USA, 7–8 December 2009; pp. 1–6.
9. Moilanen, A.; Zhao, G.; Pietikäinen, M. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 1722–1727.
10. Patel, D.; Zhao, G.; Pietikäinen, M. Spatiotemporal integration of optical flow vectors for micro-expression detection. In Proceedings of the Advanced Concepts for Intelligent Vision Systems: 16th International Conference, ACIVS 2015, Catania, Italy, 26–29 October 2015; Springer International Publishing: Cham, Switzerland, 2015; pp. 369–380.
11. Li, J.; Soladié, C.; Séguier, R.; Wang, S.-J.; Yap, M.H. Spotting micro-expressions on long videos sequences. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, Lille, France, 14–18 May 2019; pp. 1–5.
12. He, Y.; Wang, S.J.; Li, J.; Yap, M.H. Spotting macro-and micro-expression intervals in long video sequences. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires, Argentina, 16–20 November 2020; pp. 742–748.
13. Xia, Z.; Feng, X.; Peng, J.; Peng, X.; Zhao, G. Spontaneous micro-expression spotting via geometric deformation modeling. *Comput. Vis. Image Underst.* **2016**, *147*, 87–94. [CrossRef]
14. Hong, X.; Tran, T.K.; Zhao, G. Micro-expression spotting: A benchmark. *arXiv* **2017**, arXiv:1710.02820.
15. Nag, S.; Bhunia, A.K.; Konwer, A.; Roy, P.P. Facial micro-expression spotting and recognition using time contrasted feature with visual memory. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2022–2026.

16. Verburg, M.; Menkovski, V. Micro-expression detection in long videos using optical flow and recurrent neural networks. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, Lille, France, 14–18 May 2019; pp. 1–6.
17. Pan, H.; Xie, L.; Wang, Z. Local bilinear convolutional neural network for spotting macro-and micro-expression intervals in long video sequences. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires, Argentina, 16–20 November 2020; pp. 749–753.
18. Yap, C.H.; Yap, M.H.; Davison, A.; Kendrick, C.; Li, J.; Wang, S.-J.; Cunningham, R. 3d-cnn for facial micro-and macro-expression spotting on long video sequences using temporal oriented reference frame. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 7016–7020.
19. Liong, G.B.; See, J.; Wong, L.K. Shallow optical flow three-stream CNN for macro-and micro-expression spotting from long videos. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 2643–2647.
20. Fang, Y.; Deng, D.; Wu, L.; Jumelle, F.; Shi, B. RMES: Real-Time Micro-Expression Spotting Using Phase from Riesz Pyramid. *arXiv* **2023**, arXiv:2305.05523.
21. Li, J.; Dong, Z.; Liu, Y.; Wang, S.-J.; Zhuang, D. A micro-expression spotting method based on human attention mechanism. *Adv. Psychol. Sci.* **2019**, *30*, 2143–2153. [CrossRef]
22. Cao, R. *Micro-Expression Detection for Long Videos Based on Outlier Detection*; South Western University of Finance and Economics: Chengdu, China, 2022.
23. Li Song, Y. *Research on Micro-Expression Spotting and Recognition Based on Convolutional Neural Networks*; Shandong University: Jinan, China, 2021.
24. Liu, L. Inverted Non-maximum Suppression for more Accurate and Neater Face Detection. *arXiv* **2023**, arXiv:2305.10593.
25. Ci, Y.; Wang, Y.; Chen, M.; Tang, S.; Bai, L.; Zhu, F.; Zhao, R.; Yu, F.; Qi, D.; Ouyang, W. UniHCP: A Unified Model for Human-Centric Perceptions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 17840–17852.
26. Oublal, K.; Dai, X. An advanced combination of semi-supervised Normalizing Flow & Yolo (YoloNF) to detect and recognize vehicle license plates. *arXiv* **2022**, arXiv:2207.10777.
27. Liong, S.T.; Gan, Y.S.; Zheng, D.; Li, S.-M.; Xu, H.-X.; Zhang, H.-Z.; Lyu, R.-K.; Liu, K.-H. Evaluation of the spatio-temporal features and gan for micro-expression recognition system. *J. Signal Process. Syst.* **2020**, *92*, 705–725. [CrossRef]
28. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
29. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
30. Lee, Y.; Kim, H.; Park, E.; Cui, X.; Kim, H. Wide-residual-inception networks for real-time object detection. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 758–764.
31. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22 October 2017; pp. 2980–2988.
32. Yan, W.J.; Wu, Q.; Liu, Y.J.; Wang, S.-J.; Fu, X. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–7.
33. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikäinen, M. A spontaneous micro-expression database: Inducement, collection and baseline. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Shanghai, China, 22–26 April 2013; pp. 1–6.
34. Yan, W.J.; Li, X.; Wang, S.J.; Zhao, G.; Liu, Y.J.; Chen, Y.H.; Fu, X. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* **2014**, *9*, e86041. [CrossRef]
35. Davison, A.K.; Lansley, C.; Costen, N.; Tan, K.; Yap, M.H. Samm: A spontaneous micro-facial movement dataset. *IEEE Trans. Affect. Comput.* **2016**, *9*, 116–129. [CrossRef]
36. Qu, F.; Wang, S.J.; Yan, W.J.; Li, H.; Wu, S.; Fu, X. CAS(ME)^2: A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. *IEEE Trans. Affect. Comput.* **2017**, *9*, 424–436. [CrossRef]
37. Yap, C.H.; Kendrick, C.; Yap, M.H. Samm long videos: A spontaneous facial micro-and macro-expressions dataset. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, Buenos Aires, Argentina, 16–20 November 2020; pp. 771–776.
38. Ben, X.; Ren, Y.; Zhang, J.; Wang, S.J.; Kpalma, K.; Meng, W.; Liu, Y.J. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5826–5846. [CrossRef] [PubMed]
39. Li, J.; Dong, Z.; Lu, S.; Wang, S.J.; Yan, W.J.; Ma, Y.; Liu, Y.; Huang, C.; Fu, X. CAS (ME)3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2782–2800. [CrossRef] [PubMed]
40. Li, J.; Soladie, C.; Seguier, R. Local temporal pattern and data augmentation for micro-expression spotting. *IEEE Trans. Affect. Comput.* **2020**, *14*, 811–822. [CrossRef]

41. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.
42. Pan, H.; Xie, L.; Wang, Z. Spatio-temporal convolutional attention network for spotting macro-and micro-expression intervals. In Proceedings of the 1st Workshop on Facial Micro-Expression: Advanced Techniques for Facial Expressions Generation and Spotting, Virtual, China, 24 October 2021; pp. 25–30.
43. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.