symmetry

MDPI

Article

CSINet: Channel–Spatial Fusion Networks for Asymmetric Facial Expression Recognition

Yan Cheng [1,2] and Defeng Kong [3,*]

[1] College of Food Science and Technology, Huazhong Agricultural University, Wuhan 430070, China; chengyan7@webmail.hzau.edu.cn or chengyan7@outlook.com
[2] School of Logistics, Wuhan Technical College of Communications, Wuhan 430065, China
[3] School of Mechanical Engineering, Hubei University of Technology, Wuhan 430068, China
* Correspondence: kongdefeng0916@gmail.com; Tel.: +86-18163510917

Abstract: Occlusion or posture change of the face in natural scenes has typical asymmetry; however, an asymmetric face plays a key part in the lack of information available for facial expression recognition. To solve the problem of low accuracy of asymmetric facial expression recognition, this paper proposes a fusion of channel global features and a spatial local information expression recognition network called the "Channel–Spatial Integration Network" (CSINet). First, to extract the underlying detail information and deepen the network, the attention residual module with a redundant information filtering function is designed, and the backbone feature-extraction network is constituted by module stacking. Second, considering the loss of information in the local key area of face occlusion, the channel–spatial fusion structure is constructed, and the channel features and spatial features are combined to enhance the accuracy of occluded facial recognition. Finally, before the full connection layer, more local spatial information is embedded into the global channel information to capture the relationship between different channel–spatial targets, which improves the accuracy of feature expression. Experimental results on the natural scene facial expression data sets RAF-DB and FERPlus show that the recognition accuracies of the modeling approach proposed in this paper are 89.67% and 90.83%, which are 13.24% and 11.52% higher than that of the baseline network ResNet50, respectively. Compared with the latest facial expression recognition methods such as CVT, PACVT, etc., the method in this paper obtains better evaluation results of masked facial expression recognition, which provides certain theoretical and technical references for daily facial emotion analysis and human–computer interaction applications.

Keywords: facial expression recognition; attention mechanism; channel–spatial information; feature fusion

Citation: Cheng, Y.; Kong, D. CSINet: Channel–Spatial Fusion Networks for Asymmetric Facial Expression Recognition. *Symmetry* **2024**, *16*, 471. https://doi.org/10.3390/sym16040471

Academic Editors: Junaid Baber, Ali Shariq Imran, Sher Doudpota and Maheen Bakhtyar

Received: 13 March 2024
Revised: 5 April 2024
Accepted: 7 April 2024
Published: 12 April 2024

1. Introduction

Human facial expressions are an important way for humans to convey emotional information and they have a wide range of potential applications in human–computer interaction [1], safe driving monitoring [2,3], medical diagnosis [4], and educational counseling [5]. In recent years, deep learning–based facial expression recognition methods have achieved better performance using laboratory data sets, but faces in natural scenes are usually occluded or only part of the face is observed during body movement, and this facial asymmetry leads to poor expression recognition in natural scenes. How to overcome the influence of obstacles such as occlusion and incomplete facial information on the accuracy of expression recognition in natural scenes is a current research hotspot in the field of affective computing and an urgent problem to be solved to improve the quality of human–computer interaction for facial expression recognition in real-world environments [6].

To solve the challenging problem of losing key information by occlusion and posing change, the related research in recent years mainly adopts global feature-based recognition methods [7,8], local feature-based recognition methods [9,10], and two fusion recognition

methods [11]. Global feature-based recognition methods usually input the whole face image into the network and extract the overall features by using reconstruction to occlude part of the image to complement the information or attention mechanism to focus on the key areas. For example, Pan et al. [12] proposed a method for occluded facial expression recognition with the help of nonoccluded facial images, which employs adversarial learning so that the occluded facial image learns the feature distribution of the unoccluded facial image to recognize the global facial image expression. However, recognition methods based on global facial features usually ignore the collaborative relationship to local high-frequency detail information, which is more effective in facial images with complete information but needs to be improved for scenes with severe occlusion and drastic pose changes.

Local feature–based recognition methods focus on local features of the face, such as the eyebrows, eyes, mouth, and other parts of the face, which mainly intercept the local region by cropping, and then extract the local key information to recognize the expression [13]. Psychologist Mehrabian [14] found that 55% of human emotion transmission comes from facial expressions, and the human visual system can utilize both local details and the whole face to perceive the semantic information transmitted by occluded faces in natural scenes. For example, Wang [15] cropped the whole facial image into several regions and used a self-attention network to adaptively obtain the importance of the facial regions for the recognition of occlusion and gesture change expressions, to improve the accuracy of the model in recognizing occluded facial expressions using different local features. Although the local feature-based recognition method pays attention to the detailed information of the key facial regions, it does not take into account the influence of the overall facial features on the correct recognition of expressions, and it is difficult to express the complete global features with the fragmented local information.

Some scholars tried to increase the understanding of the semantics of occluded facial expressions in the natural environment by fusing local and global feature methods [16–18]. Li et al. [9] proposed a convolutional neural network with an attention mechanism, which perceives the occluded region of the face by focusing on the key features of the nonoccluded region of the face, but it is difficult for a single local spatial information to communicate with the global channel information, and single semantics ignores the global contextual information, which results in wrong expression recognition. Wadhawan et al. [19] transferred the learning of local facial signatures in five subnetworks to global expression classification to reduce the impact in predicting facial expressions with extreme poses, lighting, and occlusion conditions. However, the separation of the local and global model training approach hinders the multidimensional details and semantic features from learning from each other, suppressing the impact of local features on the overall facial expression, and the overall facial expression semantics are difficult to accurately express. Yu et al. [20] used a shared shallow module to learn information from local regions and global images, and then constructed a widget-based module to extract local dynamic information related to the overall facial expression. The method of understanding the overall semantics of the facial expression through the learning of local detail features did not distinguish the importance of local features, resulting in the influence of the facial expression recognition of salient features. Therefore, salient regions do not play an important role.

Recently, some researchers have begun to experiment with new underlying network paradigms [21,22]. For example, Chen et al. [23] used graph neural networks (GCNs) to design a subspace streaming learning method for intensity-invariant facial expression recognition, which treats the target task as a node classification problem and learns the streaming representation using two subspace analyses, the locality preserving projection, and the peak-guided locality preserving projection. However, graph neural networks are usually limited to shallow learning, making it difficult to extract highly semantic features and thus failing to accurately recognize facial expressions with small differences. Cheng et al. [24] proposed a novel unified model called the "transformer autoencoder" (TAE) using a transformer-based architecture, aiming to populate modally incomplete data from partially observed data, and while better recognition performance was obtained on

the DEAP and SEED-IV data sets, the transformer architecture needs to address its high computational cost and the problem of requiring large amounts of data for specific tasks.

Existing methods [25–27] do not pay attention to the local details of high-frequency information and global spatial feature learning obstacles; channel information and spatial information fusion is not sufficient; and the significant region of low impact, and so on, leading to the natural environment of the facial expression recognition is difficult, among other issues. Therefore, this paper proposes a channel–spatial fusion network (CSINet) for the natural environment of the multistate human facial expression recognition. The network model contains three main parts as follows: the backbone feature-extraction network with a redundant-information filtering function, the channel–spatial information fusion structure, and the local global feature coordination enhancement module. Specifically, in the backbone feature-extraction network, the network is stacked and deepened by using the attention residual unit [28], which enhances and extracts high-frequency features, such as eyebrows, eye lines, and the corner of the mouth, to obtain the high-semantic features of the facial expression and, at the same time, the attention residual unit filters out a large amount of redundant information in regions that are of little help to the expression recognition such as the hair, cheeks, forehead, etc., to reduce the interference of the low-frequency information on the model and improve the robustness of the algorithm. In the channel–spatial information fusion structure, a method of fusing channel attention and spatial attention feature layers is designed, which first highlights the channel features and spatial features, respectively, and then uses convolutional operation to fuse their extracted important features with the original high-semantic features, which effectively strengthens the information exchange between the channel and the space, and improves the effect of the local details on the global spatial structure. In the local global feature coordination enhancement module, the loss of spatial information of the target is avoided by obtaining the position information and embedding it into the channel attention, and this module can accurately capture the relationship between different channels and improve the accuracy of feature expression.

Compared with existing facial expression recognition networks with global and local feature fusion, the main contributions of this paper are as follows:

(1) An attention residual module emphasizing detailed features is used in the critical backbone feature-extraction session, which is the basis of expression classification.
(2) Avoiding the split between spatial local features and global channel feature learning, and embedding spatial local salient features into optimized channel features to improve the semantic expression accuracy.
(3) Avoiding the separation of the training of the local network and the overall network during the model training process, and adopting the end-to-end one-piece training method to improve the collaboration ability of detail features to semantic features.
(4) This paper architects a new spatial integration model for occlusion or gesture change facial expression recognition channels through the design of different functional modules, which provides new theoretical and technical support for deep learning–based emotion computing methods.

2. CSINet Design

This paper proposes a channel–spatial fusion network that can realize the task of high-precision recognition of occluded facial expressions in natural environments. The network mainly consists of the following three parts: (1) a backbone feature-extraction network; (2) a channel–spatial feature fusion structure; and (3) a coordination and enhancement module for local and global features. The overall network structure is shown in Figure 1. Among them, the backbone feature-extraction network constructed by the attention residual module solves the problem of local and global information redundancy and insignificant important features. The channel–spatial feature fusion structure can make up for the problem of insufficient fusion of single-channel information and spatial information, which leads to the difficulty of recognizing occluded facial features. The local global feature coordination and

enhancement module is used to enhance the representation of locally significant features, and improve the impact of locally important features on global semantics, to recognize the occluded or occluded facial features and identify the occluded or occluded faces. The local global feature coordination enhancement module is used to enhance the local salient feature representation to improve the impact of locally important features on the global semantics, to recognize the real expression of the face in the region of occlusion or posture change. In addition, the CSINet model is not trained in steps and adopts an end-to-end integrated training mode, which avoids mutual learning of multidimensional details and global semantic features and improves the generalization ability of expression recognition.

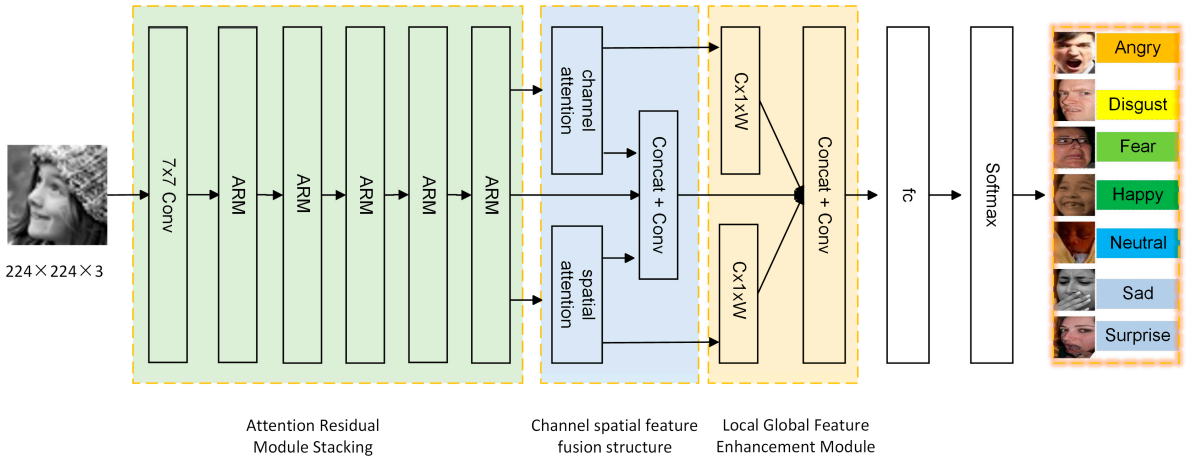

Figure 1. Channel–spatial fusion network structure.

2.1. Attention Residual Module

In the feature-extraction stage, the basic feature-extraction module is composed of three functional blocks of a 3 × 3 convolutional layer, normalization layer, and ReLU activation function layer arranged in sequence and repeated once to constitute the basic feature-extraction unit. The network structure is shown in Figure 2.

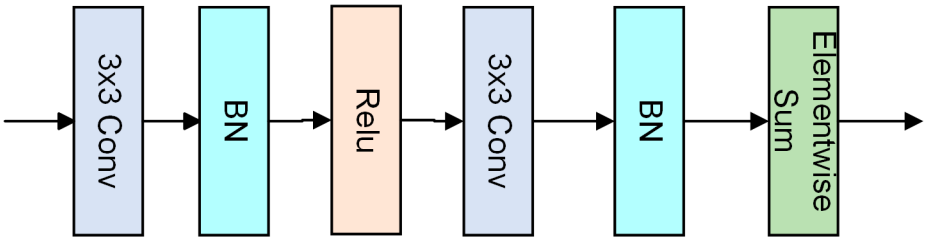

Figure 2. Basic feature-extraction module.

The logical computational relationship of this module can be expressed by Equation (1)

$$T = \delta(BN(f_{3\times3}(I_O))) \tag{1}$$

where I_O denotes the original training image; $f_{3\times3}$ denotes the convolutional layer with a convolutional kernel size of 3; BN denotes the normalization layer; $\delta(\cdot)$ denotes the activation function, and here the ReLU activation function is used; and T denotes the extracted feature vector. The attention residual structure is combined with the basic feature-extraction module to form a new type of feature-extraction module with residual mapping function and attention to the important features, and the structure of the backbone feature-extraction network stacked with this new type of module is shown in Figure 3. The original face image usually contains rich low-frequency information, and the detailed edge features such as eyebrows, eyes, mouth, etc. mainly exist in the high-frequency information, while

the high-frequency information is weak and the detailed features are gradually lost in the process of constant convolution. Therefore, it is necessary to introduce the residual jump-connection branch [29,30], which adds shallow high-frequency information after a certain stage of network learning, and strengthens the expression of the detailed edge features. Although the jump connection of the residual structure strengthens the transmission of high-frequency information to the deep layer, but creates more low-frequency information of color brightness, low-frequency information has a discrete effect on the deep high-level semantic information, which is not conducive to the accurate identification of the target. Therefore, adding the attention mechanism to the jump-connection branch can effectively attenuate the transmission of redundant information. The attention mechanism [3–5] gives more weight to the important information through the learning of features to enhance the expression and transmission of the important features, and the calculation of the expression is shown in Equation (2)

$$T_n = BN(f_{3\times3}(\delta(BN(f_{3\times3}(T_{n-1}))))) + A_{cbam}(T_{n-1}) \quad (2)$$

where T_{n-1} is the input feature of the nth attentional residual module; $f_{3\times3}$ denotes the convolutional layer with convolutional kernel size 3; BN denotes the normalization layer; $\delta(\cdot)$ denotes the ReLU activation function; $A_{cbam}(\cdot)$ denotes the spatial hybrid attentional mechanism of the CBAM [31] channel; and T_n denotes the output feature of the nth attentional residual module. According to Figure 3 and Equation (2), the pseudo-code algorithm is as Algorithm 1.

Algorithm 1: Attention Residual Module Feature-Extraction Algorithm

function ARM(x);
Input: Original low-resolution image
Output: eigenmaps
if x = 0 **then**
    return 0;
else
    $x_1 = A_{cbam}(x)$;
    $x_2 = f_{3\times3}(x)$;
    $x_3 = \mathrm{BN}(x_2)$;
    $x_4 = \delta(x_3)$;
    $x_5 = f_{3\times3}(x_4)$;
    $x_6 = \mathrm{BN}(x_5)$;
    $x_7 = x_1 + x_6$;
    return x_7;
end

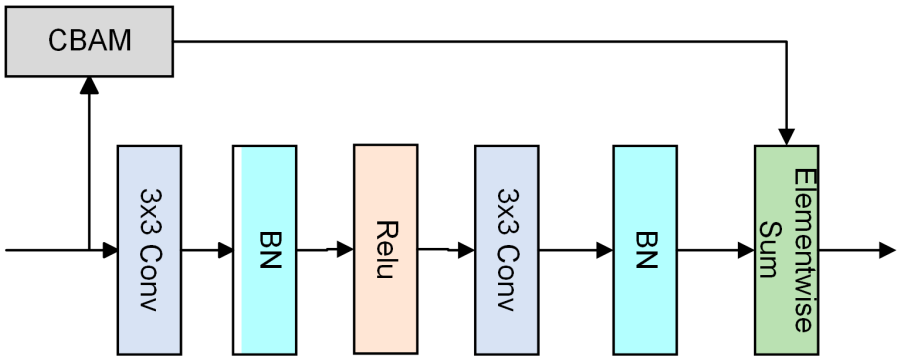

Figure 3. Attention residual module.

In the pseudo-code Algorithm 1, the variable x is a feature map learnable parameter, which is a high-dimensional set of parameters, exactly a feature map with a multidimen-

sional tensor, rather than a specific numerical value, and $x_1 - x_6$ with subscripts likewise refer to the feature maps computed by the different network layers, and the sequence of processing by the different network layers is by the letters of the subscripts from the smallest to the largest, and the final result of the network processing x_7 is returned.

2.2. Channel–Spatial Feature Fusion Structure (CSFF)

The feature map information carriers are usually channel information and spatial information, and channel information refers to the information difference between different color channels (e.g., red, green, and blue channels in an RGB image) for each pixel point in an image [32,33]. Channel information represents the global characteristics of an image, such as the color, lightness, and darkness of the image [34,35]. Spatial information, on the other hand, refers to the spatial relationship between the position of each pixel point in an image and the surrounding pixel points. Spatial information represents the local features of an image, such as texture, edges, etc. In contrast, in fine-grained visual classification tasks, both channel information and spatial information need to be considered because the detail information usually involves both the color and shape of the image [36,37]. Therefore, a fusion of spatial detail features with global features of channels is an effective way to improve the accuracy of visual classification tasks. Figure 4 shows the network structure of the channel–spatial feature fusion method.

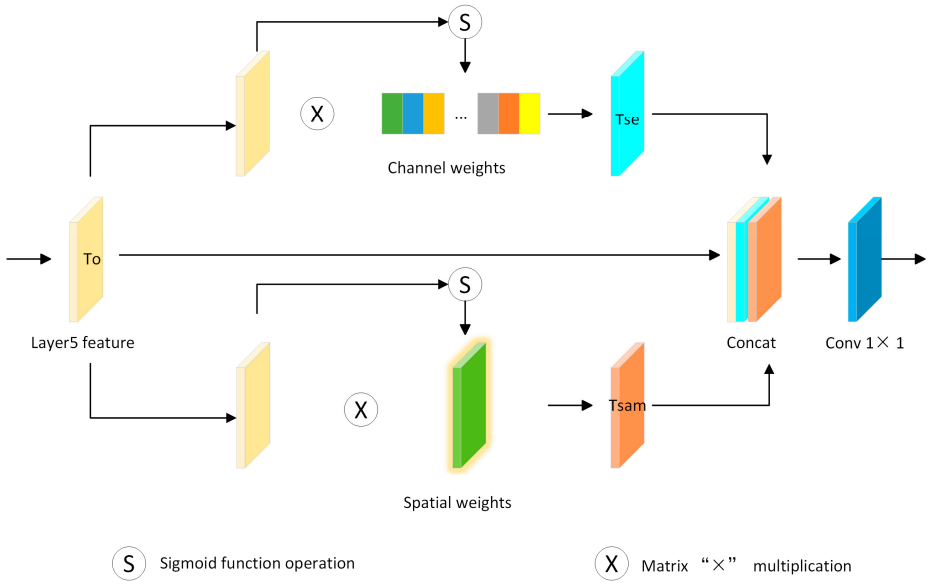

Figure 4. Channel–spatial feature fusion structure.

The feature fusion mechanism proposed in this paper needs to obtain two feature maps with identical dimensions but different weighting parameters at the same time. Combining the observation of Figures 1 and 4, we take the last layer of feature maps after the end of the main feature extraction, and do the channel and spatial attention weighting on the feature map T, respectively, to obtain the weighted channel feature map T_{se} [20] and the spatial feature map T_{sam} [21]. The feature map T_{se}, T_{sam}, and the original feature map T_o are spliced in the channel dimension, and the spatial and channel features are fused and learn useful information from each other by using the size 3×3 filter convolution operation. The feature fusion process is represented by Equation (3)

$$T_n = f_{3\times3}(\delta(Concat[T_o, T_{se}, T_{sam}])) \tag{3}$$

where T_o is the original feature map; T_{se} is the feature map weighted by the SENet channel attention mechanism; T_{sam} is the feature map weighted by spatial attention; $Concat[\bullet]$ denotes the superposition of the different feature layers over the channel dimensions; $\delta(\bullet)$

denotes the ReLU activation function; and $f_{3\times3}$ denotes the convolutional kernel size of three, and the feature map T_n is obtained after the fusion of the spatial features of the channel by the operation in Equation (3).

2.3. Local Global Feature Coordination Enhancement Module (LGFE)

In visual tasks, it is generally believed that local information contains a large amount of spatial location information, which is crucial for capturing object structure in visual tasks, and embedding location information into global channel information will help enhance the representation of objects of interest [36]. In this paper, inspired by the coordinate attention mechanism [38], the feature tensor weighted by spatial attention aggregates features along the H(height) direction, and the feature tensor weighted by channel attention aggregates features along the W(width) direction. In this way, long-range dependencies in one spatial direction can be captured and precise position information in the other spatial direction can be retained at the same time. The network structure is shown in Figure 5.

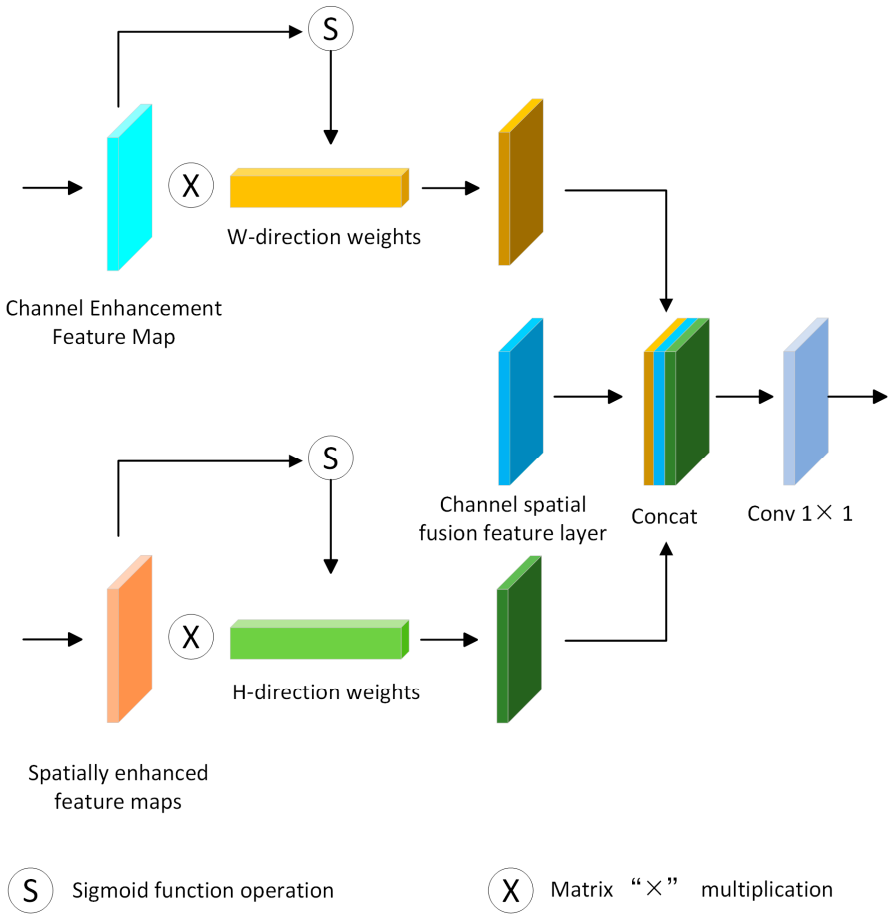

Figure 5. Local global feature coordination enhancement module.

As shown in the figure, two 1D global pooling operations are utilized to aggregate the input features along the vertical and horizontal directions, respectively, while the input feature maps come from the channel-attention-weighted feature map T_{se} [39] and spatial-attention-weighted feature map T_{sam} [40], respectively, to form two separate directionally oriented feature maps $T_{se,\,h}$ in the H-direction and $T_{sam,\,w}$ in the W-direction sensing feature maps. These two feature maps embedded with direction-specific information are encoded as two separate attention maps, each of which captures the long-range dependencies of the input feature maps in one spatial direction, thus preserving positional information. Finally, these two attention maps are applied to the input feature map by multiplication

to emphasize the representation of interest. The network computation process can be represented by Equations (4)–(8)

$$T_{c\times h\times 1} = P_{(1,w)}(T_{se}) \tag{4}$$

$$T_{c\times 1\times w} = P_{(h,1)}(T_{sam}) \tag{5}$$

$$g^h = \delta(T_{c\times h\times 1}) \tag{6}$$

$$g^w = \delta(T_{c\times 1\times w}) \tag{7}$$

$$T_{n+1} = T_n \times g^h \times g^w \tag{8}$$

where, T_{se} is the feature map after channel attentional weighting; $P_{(1,w)}(\bullet)$ denotes pooling along the H-direction with a pooling kernel size of $(1, w)$; T_{sam} is the feature map after spatial attentional weighting; $P_{(h,1)}(\bullet)$ denotes pooling along the W-direction with a pooling kernel size of $(h, 1)$; $\delta(\cdot)$ denotes the Sigmoid activation function; g^h is the attentional weight in the H-direction of the space; g^w is the attentional weight in the W-direction of the space; T_n is the feature map after fusion of the channel space features; and T_{n+1} is the output tensor. According to Equations (4)–(8), the pseudo-code algorithm is as Algorithm 2.

Algorithm 2: Coordinated Enhancement Algorithm for Local Global Features

```
function LGFE(x);
Input: Characterization of the front layer
Output: Local Global Feature Coordination Enhanced Fusion Feature Map
if x = 0 then
    return 0;
else
    if x = T_se then
    T_{c×h×1} = P_{(1,w)}(x);
    end
    if x = T_sam then
    T_{c×1×w} = P_{(h,1)}(x);
    end
    if x = x then
    x = x;
    end
    x_1 = Concat[T_{c×h×1}, T_{c×1×w}, x]
    x_2 = f_{1×1}(x_1)
    return x_2;
end
```

In the pseudo-code Algorithm 2, the variable x involved in the computation is a feature map with a multidimensional tensor. $T_{c\times h\times 1}$, $T_{c\times 1\times w}$, x_1, and x_2 are obtained after specific network layer operations to obtain a new feature layer, when the learnable feature maps go through the different network layers to perform the corresponding rule computation, and finally, the algorithm returns the computed feature map x_2.

3. Data Sets and Experimental Platforms

3.1. Data Set Construction

The images of the data set in the laboratory environment are obtained from volunteers in the background, light source, posture, and facial unobstructed conditions, while most of the samples in the facial expression data set in the natural scene are taken from the real face image material collected from the internet, in which the main factors affecting the recognition of facial expressions, such as posture, lighting, occlusion, and other conditions, well simulate the state of the face in the real environment. Therefore, this paper selects the representative RAF-DB [41] and FERPlus [42] data sets in natural scenes to evaluate the accuracy of the modeling algorithm, and considers the special occlusion and pose factors to

evaluate the performance of facial occlusion and pose variation expression recognition on the following two sub–data sets: Occlusion Datasets and Pose Variation Datasets. Figure 6 illustrates some samples from the RAF-DB, FERPlus, Occlusion Datasets, and Pose Variation Datasets data sets.

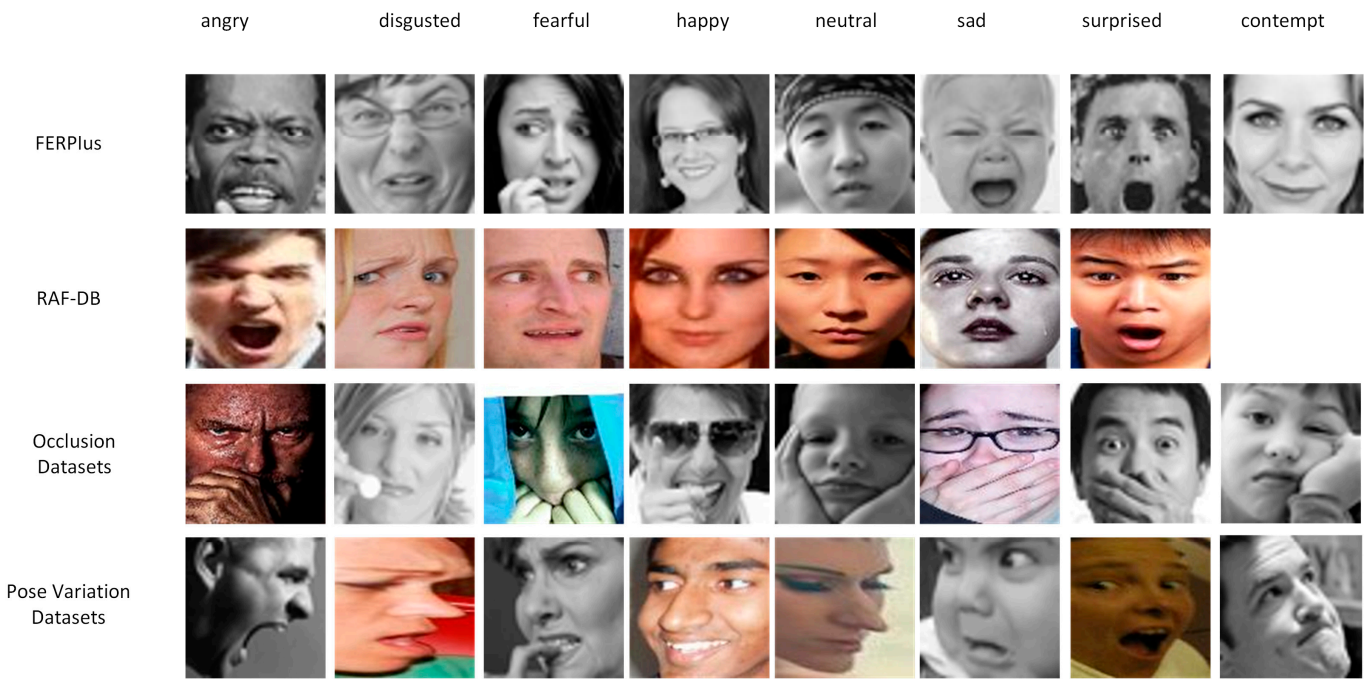

Figure 6. Sample expressions for different types of data sets.

RAF-DB is a data set of facial expressions in natural scenes collected from the internet, which consists of 29,672 diverse facial images. The image faces in the data set are highly varied in terms of age, gender, ethnicity, head pose, illumination conditions, occluders (e.g., eyeglasses, facial hair, or self-obscurations), and postprocessing manipulations (e.g., a wide variety of filters and special effects), and are labeled with seven kinds of FERPlus is an extended version of the FER2013 data set, with 28,709 training samples and 3589 test samples, all of which were manually screened and scaled to the same pixel size. The Occlusion Datasets were collected by Wang et al. [15] from the test sets of RAF-DB and FERPlus as an occlusion data set, with a total of 1340 images. Wang et al. also considered the effect of head pose variation on facial expression recognition, and collected head pose pitch angle and yaw angle images from the RAF-DB and FERPlus test sets, of which 2419 images are larger than a 30° angle, and 1192 images are larger than a 45° angle, which constitutes the Pose Variation Datasets data set. Since the research object of this paper is occlusion or pose variation facial expression recognition in natural scenes, the RAF-DB and FERPlus data sets and their sub–data sets, Occlusion Datasets and Pose Variation Datasets, can validate the reliability and effectiveness of this paper's method.

3.2. Experimental Platform and Parameter Settings

The experimental hardware environment for the model training test in this paper is as follows: the CPU of the computer is 12th Gen Intel® Core™ i5-12600KF 3.70 GHz; the system memory is 16 G; and the graphics card is NVIDIA GeForce RTX 3070 GPU, with 8 GB of video memory capacity. The software environment is as follows: Windows 10 operating system; Pycharm compilation environment; PyTorch1.12 deep learning framework; cuda11.6 accelerated computing platform; Anaconda3.0 environment manager; and the programming language is Python3.8. After many experimental explorations, the channel–spatial fusion network proposed in this paper for facial expression recognition is summarized with

important hyperparameters. The model achieves stable and reliable performance when the model parameters are set as in Table 1.

Table 1. Experimental setup.

Set Item	Parameter
Iteration	200
Batch size	32
Initial learning rate	1×10^{-2}
Min learning rate	$(1 \times 10^{-2}) \times 0.01$
Optimizer	SGD
Momentum	0.9
Weight decay	5×10^{-4}
Learning rate decay type	COS
Thread	4

4. Experimentation and Analysis

4.1. Comparison Experiment with Existing Methods

To verify the performance of facial expression recognition of the CSINet model proposed in this paper, ResNet50 [43] is selected as the baseline model (the backbone feature-extraction network of the model in this paper is improved based on ResNet50) and representative spatially localized feature expression recognition networks (RAN, CVT) are used, as well as the expression recognition network that fuses the channel–spatial information (MA-Net, AMP-Net, VTFF, PACVT) for comparison. The network models in this paper follow the model training settings in Tables 1 and 2 and record the expression recognition accuracies obtained by different network models on different data sets.

Table 2. Correctness of expression recognition on RAF-DB and FERPlus data sets with different network models.

Mold	RAF-DB (%)	FERPlus (%)
ResNet50 (baseline) [30]	76.43	79.31
RAN [15]	86.90	88.55
CVT [44]	88.14	88.81
MA-Net [45]	88.40	-
AMP-Net [46]	89.25	-
VTFF [47]	88.14	88.67
PACVT [47]	88.21	88.72
CSINet (Ours)	89.67	90.83

In Table 2, except for the algorithm proposed in this paper based on the experimental setup, the recognition accuracy data are obtained on RAF-DB and FERPlus data sets, and the recognition accuracy of other network models are referred to the corresponding literature experimental conclusions ("-" indicates that the relevant literature does not provide this experimental data). The experimental results show that compared with the basic ResNet50, the method in this paper has 13.24% and 11.52% substantial improvement, indicating that the simple deep network cannot recognize the expression features well. When comparing with RAN, CVT, and other networks that only use spatially localized features for recognition, the network in this paper improves the recognition accuracy by 1.53–2.77%, which is mainly due to the fact that the CSINet algorithm takes into account the importance of fusing channel global information with spatial local features, and in fine-grained visual classification tasks, it is necessary to consider both channel information and spatial information, because the detail region images contain both colors and edge shapes, and the color information is usually represented by the channel, and the shape contour features are distributed in the image space. Therefore, the interactive fusion of channel–spatial information compensates for the lack of important information about a

single spatial feature in the natural scene applications with a lack of important information, which helps to improve the network generalization ability and the accuracy of recognizing facial expressions. On the RAF-DB and FERPlus data sets, compared with the expression recognition networks based on channel–spatial fusion such as MA-Net, AMP-Net, VTFF, PACVT, etc., the network model in this paper obtains recognition accuracies of 89.67% and 90.83%, respectively, which are higher than those recorded by the existing models. The better recognition performance of the CSINet network is mainly attributed to the attention residuals. The better recognition performance of the CSINet network is mainly due to the design and use of the attention residual module. In the process of shallow feature extraction, detailed high-frequency information is easily lost. The low-frequency information is easily transferred to the deep feature layer due to the residual jump connection, which results in the dispersion of the semantic features. The attention residual module effectively inhibits the transfer of redundant information and enhances the representation of the detailed features.

To verify the performance of the CSINet network for occlusion or pose variation facial expression recognition in the natural environment, this paper makes a comparison with existing related network models on the Occlusion Datasets and Pose Variation Datasets, respectively, and the results are recorded in Table 3.

Table 3. Comparison of recognition accuracies of different network models on the joint data set of Occlusion Datasets and Pose Variation Datasets.

Mold	Occlusion (%)		Pose > 30° (%)		Pose > 45° (%)	
	RAF-DB	FERPlus	RAF-DB	FERPlus	RAF-DB	FERPlus
RAN [15]	82.72	83.63	86.74	82.23	85.20	80.40
FER-VT [48]	84.32	85.24	88.03	88.56	86.08	87.06
CVT [44]	83.95	84.79	87.97	88.29	88.35	87.20
MA-Net [45]	83.65	-	87.89	-	87.99	-
AMP-Net [46]	85.28	85.44	89.75	88.52	89.25	87.57
GE-LA [49]	85.30	86.24	89.94	89.02	89.45	88.80
CSINet (Ours)	85.74	86.49	90.16	89.61	89.60	89.13

Observing Table 3, it can be seen that in the natural scene where the face is under occlusion or lack of facial information for posture change and other unfavorable environments for expression recognition, the recognition accuracy of this paper's method is improved to a different degree compared to other facial expression recognition methods, which indicates that this paper's method has a better generalization ability and robustness for the problem of occlusion and posture facial expression recognition. Unlike RAN, FER-VT, CVT, and other networks that only consider spatial features, the channel–spatial feature fusion structure proposed in this paper combines the spatial structure features of local details with the global channel information, which is weighted by the attention to fuse the learning details and semantic information, thus facilitating the reduction of the negative effects of occlusion and gesture changes. Compared with MA-Net, AMP-Net, and other networks that single focus on channel enhancement of global features while ignoring spatial semantic information, the method in this paper enhances local features from spatial streams and global features from channel streams, respectively, and considers both local detail feature enhancement and global semantic feature expression, which improves the relevance and discriminability of global contextual features. Compared with the GE-LA network, although this network considers the enhanced fusion of channel–spatial features at the same time, it ignores the problem of losing detailed features and spatial location information of the base feature-extraction network, which causes the network to lose texture features of local regions such as eyebrows, corners of eyes, and corners of mouths, etc., and the separation of the spatial location information from the channel information results in the network model's difficulty in capturing the visual task with long-range dependent relationships in the visual task, which results in difficulties in recognizing occlusion and gesture change

facial expressions. In this paper, the local global feature coordination enhancement module embeds local spatial position information into channel information to reduce the negative effect of the convolutional operation on long-range dependency modeling in visual tasks and improves the generalization performance of the network for occlusion and gesture change facial expressions.

4.2. Ablation Experiments

The above experiments have illustrated the good performance of the CSINet network, but have not yet verified that the main improvement modules all play a positive role. This section conducts ablation experiments on the three main part-structural modules of the CSINet network composition—namely, ARM, CSFF, and LGFE—to assess the impact of the different modules on the overall performance of the network. The results of the experiments are recorded in Table 4.

Table 4. Experimental results of ablation of the main improvement modules on the natural scene data set.

Ablation Strategy	ARM	CSFF	LGFE	RAF-DB (%)	FERPlus (%)
(a)	×	×	×	76.43	79.31
(b)	√	×	×	84.29	85.19
(c)	×	√	×	83.84	84.38
(d)	×	×	√	86.74	88.20

The baseline network for the ablation experiments still chooses ResNet50 as the base comparison model because the CSINet network proposed in this paper modifies and adds three main modules to ResNet50. In strategy (a), the ResNet50 network obtains recognition accuracies of 76.43% and 79.31% on the data sets RAF-DB and FERPlus, respectively, and in strategy (b), the Attention Residual Module [50] is replaced with the traditional basic residual structure in ResNet50. The recognition accuracies on the data sets are improved by 7.86% and 5.88%, respectively, which is an increase of a large magnitude that indicates that the ARM module improves the ability to extract important features from the bottom layer of the network. Its redundant information filtering function effectively avoids the discrete effect of a large amount of low-frequency information on high semantic features.

With the addition of the channel–spatial feature fusion structure (CSFF) in strategy (c), the recognition accuracies on the data set are improved by 7.41% and 5.07%, respectively, and the improvement is not as large as that of strategy (b), but it also fully illustrates the importance of the channel–spatial feature fusion structure (CSFF), because any target in an image is composed of color information and spatial structure features, so in the visual recognition, channel information and spatial features must be considered simultaneously in the visual recognition task. In most network models, only the information of a single mode is usually taken into account while ignoring the assistance of other modes. In this paper, we propose the CSFF structure to fuse the enhanced channel features with spatial features and use convolutional operations to make the channel and the space learn each other's important information, to improve the model's ability to recognize the semantics of expressions.

In strategy (d), the Local Global Feature Coordination Enhancement (LGFE) module embeds the local spatial location information into the channel information to avoid the inefficiency of the convolutional operation in capturing the local relationship to the long-range dependence in the visual task, and the recognition accuracy is improved by 10.31% and 8.89% on the RAF-DB and FERPlus data sets, respectively. The improvement of the recognition effect with the addition of the LGFE module is by a larger margin, which shows that the network's ability to capture cross-channel information, as well as orientation-aware and location-sensitive information, affects the accuracy of the model in localizing and recognizing the object of interest. Therefore, the design of the local global feature coordination enhancement module improves the model's ability to recognize expressions that are affected by occlusion and posture change faces over long distances in natural environments.

4.3. Visualization Analysis

To better explain the CSINet effectiveness, this experiment utilizes the GradCam (gradient-weighted class activation mapping) [51] method to visualize and analyze the ResNet50 baseline model, the spatially localized feature expression recognition network FER-VT, the fusion channel–spatial information expression recognition network AMP- Net, and the models in this paper. To test the network model's focus on important regions more broadly, this paper collects normal facial images as well as samples with occlusion and gesture change features from the RAF-DB data set, which contains both full facial samples and occlusion and gesture change sample cases such as illumination, adornment, hair, gesture, head deflection angle, and so on. Figure 7 shows the results of different network model visualizations.

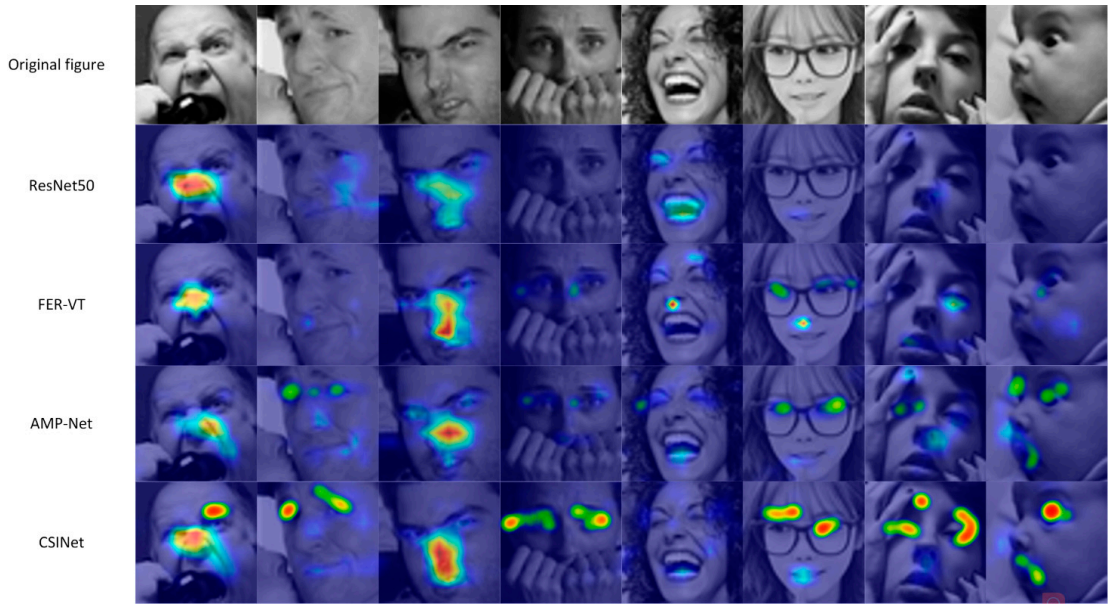

Figure 7. Visualization of different network models for facial expression recognition.

As shown in Figure 7, the first row is the original image, which contains eight kinds of multimodal expressions as follows: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise (left to right). The second row is the visualization of the baseline model ResNet50. The third row is the visualization of the spatially localized feature expression recognition network FER-VT. The fourth row is the visualization of the expression recognition network AMP-Net that fuses the channel–spatial information. Finally, the fifth row is the visualization of the CSINet network proposed in this paper. It can be seen that the spatial location of the face concerned by the baseline ResNet50 network in the second row is not in the key feature expression region, the focus range is diffuse and unfocused, and the occluded and offset part of the face is unfocused, which results in the phenomenon of recognition difficulties. The third row of spatial local features expression recognition network FER-VT considers local detail features, the heat map focus begins to focus on the eyes, eyebrows, mouth, and other key local areas, but the focus is small and there is a large offset situation. The fourth line AMP-Net network fuses channel information with spatial information, which enhances the network's focus on the important areas of the face and semantic expression ability. From the heat map, it can be seen that the focus on the eyes, eyebrows, and mouth focus areas have been focused on learning by the network, the focus on the region is more complete, and the area is significantly larger. The fifth line shows that the attention residual structure uses the redundant-information network filter, such as the hair, cheeks, chin, ears, and other regions are ignored, without extensive attention to the region that is not important to the semantic recognition of the expression. The channel and

the spatial attention information fuse so that the network can learn the important details of the texture features and the global grayscale information heat map focus of attention on the main emotional expression region. The spatial location information embedded in the channel location information makes the network locate the important feature locations more accurately and comprehensively, and the heat map focuses on more comprehensive focus areas with increased area coverage and smaller deviation, indicating that the spatial location information introduced by the local global feature coordination enhancement module assists the network in locating the focus areas. In the field of image recognition, the confusion matrix can be used to evaluate the model's performance in recognizing different objects. Figure 8 shows the confusion matrix [52] of the CSINet model on the RAF-DB data.

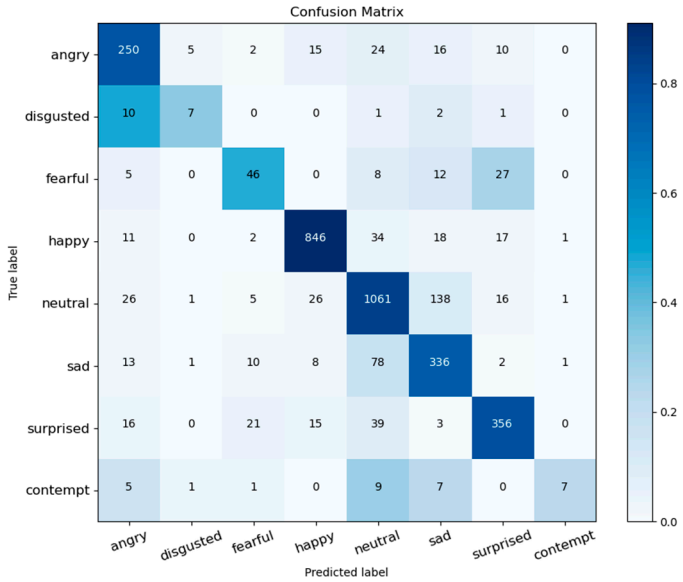

Figure 8. CSINet confusion matrix on the RAF-DB data.

The diagonal lines of the matrix indicate the number of correctly categorized categories, and the off-diagonal points are cases of incorrect categorization. Figure 8 shows that the vast majority of categories have darker diagonal lattice colors and lighter off-diagonal lattice colors, indicating that the CSINet model has good performance on the multicategory classification task and that the "disgusted" and "contested" categories have relatively lighter main diagonal lattice colors, which can be explained by the fact that these two expressions vary in appearance performance. Categories with relatively lighter main diagonal lattice colors can be interpreted as these two expressions do not show significant changes in appearance, and thus are easily misclassified.

4.4. Performance Experiments under Occlusion Environment

To verify that the algorithm model in this paper can still maintain good performance in real-life scenes where the face is under occlusion, this paper randomly selects images of real people's faces taken in random scenes and performs artificial occlusion of the face at different positions to simulate the occlusion that may exist in a natural scene. As shown in Figure 9, this paper adopts four masking methods as follows: masking the upper region of the image, masking the middle region of the image, masking the lower region of the image, and random masking. The AMP-Net network, which performs better in the above experiments, is selected as a comparison, and the recognition accuracy of the proposed model under different occlusion methods is recorded in Table 5.

Figure 9. Face images under different occlusion methods.

Table 5. Comparison of recognition accuracy between the models AMP-Net and CSINet under different occlusion methods.

Mould	Mask Method	Accuracy (%)	
AMP-Net	Upper Mask	83.54	84.82 (average)
	Middle block	85.19	
	Lower block	84.62	
	Random Mask	85.91	
	No Mask	88.75	
CSINet	Upper Mask	86.37	87.69 (average)
	Middle block	88.19	
	Lower block	87.53	
	Random Mask	88.68	
	No Mask	90.83	

The selected photographs originally had occlusion and posed change problems, and when the occlusion region is artificially added again, it reduces the effective information in the image for recognizing that the expression is greatly reduced, thus creating a greater recognition challenge for the network model. As seen in Table 4, it is found that when there is no occlusion, both the CSINet network and the AMP-Net network in this paper have better performance realizations, but the model in this paper is still a little better. Comparing the occlusion situation again, the average decrease of the AMP-Net network is 3.93%, the average decrease of the CSINet network is 3.14%, and the decrease of AMP-Net recognition accuracy is significantly larger than that of CSINet, which indicates that the algorithm of this paper still has a better robustness and generalization ability under the condition of extreme effective information incompleteness. From another perspective, among different masking methods, the recognition accuracy decreases the most when masking part of the human face, because the upper part contains important regions such as eyes and eyebrows, which have more key information and are very important for expression recognition. However, comparing from the dimension of upper part masking, this paper's model is 2.83% higher than AMP-Net's recognition accuracy, reaching 86.37%, which indicates that this paper's network still has good expression recognition accuracy under extremely difficult occlusion environments, and the performance is more stable compared to other model algorithms,

and more accurately recognizes the semantic information of the occluded regions with long-distance correlation.

To further verify the recognition performance on nonpublic data sets, facial expression images in real natural environments were collected for testing the recognition effect of the CSINet network and AMP-Net network in this paper, and some of the test facial expressions are shown in Figure 10.

Figure 10. Real-time natural environment facial expression test image.

The real-time real natural environment facial expression test images are partly from the network, and partly taken by an iPhone12. There are 40 test images for each kind of expression, and the test data set has a total of 320 images. Forty images of the same kind of expression are inputted into the converged CSINet and AMP-Net networks for prediction, respectively, and the number of correctly predicted images is divided by the total number of images (40 images) and then the percentage (%) is calculated, which is the recognition accuracy of each expression. All test images are independent of the RAF-DB and FERPlus public data sets, so that the algorithm can be verified in real-time real natural environment expression recognition performance. Observing the test results in Figure 11, the accuracy of this paper's algorithm CSINet in all types of expression recognition is higher than the comparison algorithm AMP-Net, which indicates that the algorithm still has better robustness and generalization ability in real test images of nonpublic data sets.

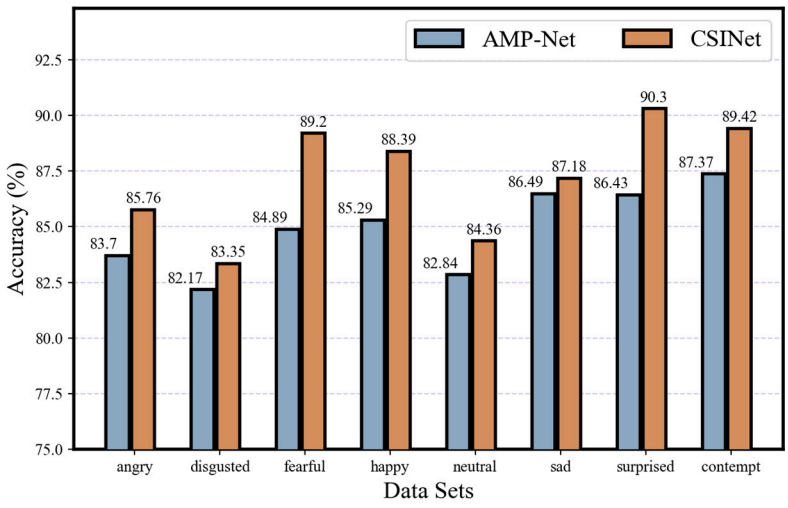

Figure 11. Facial expression recognition accuracy of CSINet and AMP-Net in the real-time natural environment.

The above experimental results show that the method in this paper proves its good performance on the RAF-DB and FERPlus public data sets, as well as on real-time real natural environment test image sets, and further research in the future will combine the software and hardware, which is of practical significance in the fields of human–computer interaction, safe driving, or mental health counseling. The method of this paper has a higher recognition accuracy compared with other methods in this field. It cannot be deployed

in edge computing devices for underlying tasks such as sentiment analysis in practical engineering applications.

5. Statement of Conclusions and Limitations

In this paper, we focus on the task of facial expression recognition, especially the problems of recognition difficulty and low accuracy when the difficulty of expression recognition increases when the face is occluded or the posture changes, and construct a channel–spatial fusion network (CSINet) for facial expression recognition. Starting from the basic ResNet50 network model, the network designs the attention residual module for the extraction of important detail features and the filtering of redundant information; proposes the channel–spatial feature fusion algorithm for the fusion learning of detailed texture and color location information of facial expressions and enhances the generalization ability of occlusion and posture change facial expressions; and establishes the local global feature coordination and enhancement mechanism and embeds spatial location information into the channel information to enhance the generalization ability of occlusion and posture change facial expressions. The local global feature coordination enhancement mechanism is established to embed the spatial location information into the channel information to improve the model's ability to express long-range dependency relations, thus improving the recognition performance of facial expression semantics in occlusion and posture change regions. After the experimental validation of the facial expression data sets RAF-DB and FERPlus obtained from natural scenes and the real-time test image set of real natural scenes, compared with the AMP-Net algorithm, which has the best performance, the recognition accuracy of this paper's algorithm is improved by 0.42% and 0.87%, and in the real test samples, the recognition accuracy of this paper's algorithm is improved by 2.35%, which shows that this paper's algorithm is more accurate in the public data set and the real-life scene facial expression. The public data set as well as the real-life scenarios have better generalization performance and robustness on the task of facial expression recognition, which provides a new network architecture design reference for asymmetric facial expression recognition in natural scenarios. The research of facial expression recognition is the basis of engineering applications in many fields such as human–computer interaction, safe driving, medical diagnosis, etc. The CSINet model can provide certain theoretical and technical references for future development and engineering applications in this field.

In this paper, the network model is not deployed in edge computing devices for testing the actual performance. In addition, it should be emphasized that the eight types of facial expression features (surprise, fear, disgust, happiness, sadness, anger, contempt, and neutrality) are sufficiently different, even if the characteristics of the distinctive expression of the current method of recognition accuracy still does not reach a sufficiently high level, which has a lot of room for improvement. In addition, facial expression recognition is a complex and difficult visual recognition task. This paper, as well as many academic articles, does not fully consider the impact of different races, countries, regions, ages, skin color, and other factors, so the actual generalization performance is still to be proved. The next stage of the task will be to self-construct the facial expression data set and train the network model proposed in this paper, taking into account the adaptability of human–computer interaction as well as the inference efficiency.

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

Funding: This research was funded by Hubei Province Key R&D Program of China, grant number 2022BBA0016; 2022 Special Tasks of Philosophy and Social Science Research of Hubei Provincial Department of Education, grant number 22Z299.

Data Availability Statement: The data that support the findings of this study are openly available in [CSINet] at [https://github.com/jackong180/CSINet.git].

Acknowledgments: We thank "Hubei Province Key R&D Program of China (2022BBA0016) for the support of experimental equipment and materials, and "2022 Special Tasks of Philosophy and Social Science Research of Hubei Provincial Department of Education (22Z299)" for help with experimental techniques.

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

References

1. Huang, X.; Romano, D.M. Coral Morph: An Artistic Shape-Changing Textile Installation for Mindful Emotion Regulation in the Wild. *Int. J. Hum. Comput. Interact.* **2024**, 1–17. [CrossRef]
2. Jeong, M.; Ko, B.C. Driver's facial expression recognition in real-time for safe driving. *Sensors* **2018**, *18*, 4270. [CrossRef] [PubMed]
3. Shafi, I.; Hussain, I.; Ahmad, J.; Kim, P.W.; Choi, G.S.; Ashraf, I.; Din, S. License plate identification and recognition in a non-standard environment using neural pattern matching. *Complex Intell. Syst.* **2022**, *8*, 3627–3639. [CrossRef]
4. Revina, I.M.; Emmanuel, W.S. A survey on human face expression recognition techniques. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 619–628. [CrossRef]
5. Guo, Y.; Huang, J.; Xiong, M.; Wang, Z.; Hu, X.; Wang, J.; Hijji, M. Facial expressions recognition with multi-region divided attention networks for smart education cloud applications. *Neurocomputing* **2022**, *493*, 119–128. [CrossRef]
6. Kortli, Y.; Jridi, M.; Al Falou, A.; Atri, M. Face recognition systems: A survey. *Sensors* **2020**, *20*, 342. [CrossRef] [PubMed]
7. Yang, B.; Wu, J.; Ikeda, K.; Hattori, G.; Sugano, M.; Iwasawa, Y.; Matsuo, Y. Face-mask-aware facial expression recognition based on face parsing and vision transformer. *Pattern Recognit. Lett.* **2022**, *164*, 173–182. [CrossRef] [PubMed]
8. Xu, R.; Huang, A.; Hu, Y.; Feng, X. GFFT: Global-local feature fusion transformers for facial expression recognition in the wild. *Image Vis. Comput.* **2023**, *139*, 104824. [CrossRef]
9. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [CrossRef] [PubMed]
10. Zhang, L.; Verma, B.; Tjondronegoro, D.; Chandran, V. Facial expression analysis under partial occlusion: A survey. *ACM Comput. Surv.* **2018**, *51*, 1–49. [CrossRef]
11. Zhang, H.; Huang, B.; Tian, G. Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognit. Lett.* **2020**, *131*, 128–134. [CrossRef]
12. Pan, B.; Wang, S.; Xia, B. Occluded facial expression recognition enhanced through privileged information. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 566–573.
13. Du, H.; Chen, Y.; Shu, Z. Facial Expression Recognition Algorithm Based on Local Feature Extraction. In Proceedings of the 2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA), Shenyang, China, 26–28 January 2024; pp. 113–118.
14. Mehrabian, A.; Russell, J.A. A verbal measure of information rate for studies in environmental psychology. *Environ. Behav.* **1974**, *6*, 233.
15. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef]
16. Shi, G.; Mao, S.; Gou, S.; Yan, D.; Jiao, L.; Xiong, L. Adaptively Enhancing Facial Expression Crucial Regions via a Local Non-local Joint Network. *Mach. Intell. Res.* **2024**, *21*, 331–348. [CrossRef]
17. Tao, H.; Duan, Q. Hierarchical attention network with progressive feature fusion for facial expression recognition. *Neural Netw.* **2024**, *170*, 337–348. [CrossRef] [PubMed]
18. Rizwan, S.A.; Jalal, A.; Kim, K. An accurate facial expression detector using multi-landmarks selection and local transform features. In Proceedings of the 2020 3rd International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 17–19 February 2020; pp. 1–6.
19. Wadhawan, R.; Gandhi, T.K. Landmark-Aware and Part-Based Ensemble Transfer Learning Network for Static Facial Expression Recognition from Images. *IEEE Trans. Artif. Intell.* **2022**, *4*, 349–361. [CrossRef]
20. Yu, M.; Zheng, H.; Peng, Z.; Dong, J.; Du, H. Facial expression recognition based on a multi-task global-local network. *Pattern Recognit. Lett.* **2020**, *131*, 166–171. [CrossRef]
21. Zhao, Z.; Liu, Q. Former-dfer: Dynamic facial expression recognition transformer. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 1553–1561.
22. Liu, S.; Huang, S.; Fu, W.; Lin, J.C.-W. A descriptive human visual cognitive strategy using graph neural network for facial expression recognition. *Int. J. Mach. Learn. Cybern.* **2024**, *15*, 19–35. [CrossRef]
23. Chen, J.; Shi, J.; Xu, R. Dual subspace manifold learning based on GCN for intensity-invariant facial expression recognition. *Pattern Recognit.* **2024**, *148*, 110157. [CrossRef]
24. Cheng, C.; Liu, W.; Fan, Z.; Feng, L.; Jia, Z. A novel transformer autoencoder for multi-modal emotion recognition with incomplete data. *Neural Netw.* **2024**, *172*, 106111. [CrossRef] [PubMed]
25. Zhang, Y.; Wang, C.; Deng, W. Relative uncertainty learning for facial expression recognition. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17616–17627.

26. Ruan, D.; Yan, Y.; Lai, S.; Chai, Z.; Shen, C.; Wang, H. Feature decomposition and reconstruction learning for effective facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 7660–7669.
27. Li, J.; Jin, K.; Zhou, D.; Kubota, N.; Ju, Z. Attention mechanism-based CNN for facial expression recognition. *Neurocomputing* **2020**, *411*, 340–350. [CrossRef]
28. Hu, X.; Kong, D.; Liu, X.; Zhang, J.; Zhang, D. Printed Circuit Board (PCB) Surface Micro Defect Detection Model Based on Residual Network with Novel Attention Mechanism. *Comput. Mater. Contin.* **2024**, *78*, 915–933. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Hu, Y.; Li, J.; Huang, Y.; Gao, X. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3911–3927. [CrossRef]
33. Chai, X.; Song, S.; Gan, Z.; Long, G.; Tian, Y.; He, X. CSENMT: A deep image compressed sensing encryption network via multi-color space and texture feature. *Expert Syst. Appl.* **2024**, *241*, 122562. [CrossRef]
34. Chang, D.; Ding, Y.; Xie, J.; Bhunia, A.K.; Li, X.; Ma, Z.; Wu, M.; Guo, J.; Song, Y.-Z. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Trans. Image Process.* **2020**, *29*, 4683–4695. [CrossRef]
35. Zheng, H.; Gao, W. End-to-End RGB-D Image Compression via Exploiting Channel-Modality Redundancy. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 7562–7570. [CrossRef]
36. Li, X.; Xiao, J.; Zhou, Y.; Ye, Y.; Lv, N.; Wang, X.; Wang, S.; Gao, S. Detail retaining convolutional neural network for image denoising. *J. Vis. Commun. Image Represent.* **2020**, *71*, 102774. [CrossRef]
37. Li, J.; Zhu, S. Channel-Spatial Transformer for Efficient Image Super-Resolution. In Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 2685–2689.
38. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
39. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
40. Kelenyi, B.; Domsa, V.; Tamas, L. SAM-Net: Self-Attention based Feature Matching with Spatial Transformers and Knowledge Distillation. *Expert Syst. Appl.* **2024**, *242*, 122804. [CrossRef]
41. Huang, Z.-Y.; Chiang, C.-C.; Chen, J.-H.; Chen, Y.-C.; Chung, H.-L.; Cai, Y.-P.; Hsu, H.-C. A study on computer vision for facial emotion recognition. *Sci. Rep.* **2023**, *13*, 8425. [CrossRef] [PubMed]
42. Zhao, G.; Yang, H.; Yu, M. Expression recognition method based on a lightweight convolutional neural network. *IEEE Access* **2020**, *8*, 38528–38537. [CrossRef]
43. Izdihar, N.; Rahayu, S.B.; Venkatesan, K. Comparison Analysis of CXR Images in Detecting Pneumonia Using VGG16 and ResNet50 Convolution Neural Network Model. *JOIV Int. J. Inform. Vis.* **2024**, *8*, 326–332. [CrossRef]
44. Ma, F.; Sun, B.; Li, S. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1236–1248. [CrossRef]
45. Zhao, Z.; Liu, Q.; Wang, S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Trans. Image Process.* **2021**, *30*, 6544–6556. [CrossRef] [PubMed]
46. Liu, H.; Cai, H.; Lin, Q.; Li, X.; Xiao, H. Adaptive multilayer perceptual attention network for facial expression recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6253–6266. [CrossRef]
47. Gadekallu, T.R.; Khare, N.; Bhattacharya, S.; Singh, S.; Maddikunta, P.K.R.; Ra, I.H.; Alazab, M. Early Detection of Diabetic Retinopathy Using PCA-Firefly Based Deep Learning Model. *Electronics* **2020**, *9*, 274. [CrossRef]
48. Huang, Q.; Huang, C.; Wang, X.; Jiang, F. Facial expression recognition with grid-wise attention and visual transformer. *Inf. Sci.* **2021**, *580*, 35–54. [CrossRef]
49. Juan, L.; Ying, W.; Min, H.; Zhong, H. Fusion of Global Enhancement and Local Attention Features for Expression Recognition Network. *J. Front. Comput. Sci. Technol.* **2023**, *11*, 1–16.
50. Madarkar, J.; Sharma, P.; Singh, R.P. Sparse representation for face recognition: A review paper. *IET Image Process.* **2021**, *15*, 1825–1844. [CrossRef]

51. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
52. Heydarian, M.; Doyle, T.E.; Samavi, R. MLCM: Multi-label confusion matrix. *IEEE Access* **2022**, *10*, 19083–19095. [CrossRef]