electronics

MDPI

Article

A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition

Saining Zhang [1,*], Yuhang Zhang [1], Ye Zhang [2], Yufei Wang [3,4] and Zhigang Song [4,*]

[1] School of Computer Science Technology, Beijing Institute of Technology, Beijing 100081, China; yuhangzhang@bit.edu.cn
[2] School of Automation, Beijing Information Science and Technology University, Beijing 100192, China; zhangyethu@163.com
[3] College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China; wangyufei221@mails.ucas.ac.cn
[4] Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
* Correspondence: snzhang@bit.edu.cn (S.Z.); songzhigang@semi.ac.cn (Z.S.); Tel.: +86-183-1054-5560 (S.Z.)

Abstract: In recent years, facial expression recognition (FER) has garnered significant attention within the realm of computer vision research. This paper presents an innovative network called the Dual-Direction Attention Mixed Feature Network (DDAMFN) specifically designed for FER, boasting both robustness and lightweight characteristics. The network architecture comprises two primary components: the Mixed Feature Network (MFN) serving as the backbone, and the Dual-Direction Attention Network (DDAN) functioning as the head. To enhance the network's capability in the MFN, resilient features are extracted by utilizing mixed-size kernels. Additionally, a new Dual-Direction Attention (DDA) head that generates attention maps in two orientations is proposed, enabling the model to capture long-range dependencies effectively. To further improve the accuracy, a novel attention loss mechanism for the DDAN is introduced with different heads focusing on distinct areas of the input. Experimental evaluations on several widely used public datasets, including AffectNet, RAF-DB, and FERPlus, demonstrate the superiority of the DDAMFN compared to other existing models, which establishes that the DDAMFN as the state-of-the-art model in the field of FER.

Keywords: MobileFaceNets; coordinate attention; facial expression recognition; MixConv

Citation: Zhang, S.; Zhang, Y.; Zhang, Y.; Wang, Y.; Song, Z. A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition. *Electronics* **2023**, *12*, 3595. https://doi.org/10.3390/electronics12173595

Academic Editor: Jyh-Cheng Chen

Received: 26 July 2023
Revised: 17 August 2023
Accepted: 23 August 2023
Published: 25 August 2023

1. Introduction

Facial expression plays an important role in human communication, serving as a crucial signal for understanding emotions and attitudes. Consequently, it is necessary for computers to acquire the ability to discern and interpret facial expressions.

The relationship between visual perception, environment mapping algorithms, and facial expression recognition (FER) based on biometric authentication computer vision has been clearly illustrated by [1–3], so it makes sense to use deep learning methods to solve FER problems. The prevailing architecture for FER networks typically consists of a backbone and heads. However, the most recent methods predominantly concentrate on the heads or neck regions and merely employ VGG [4] or ResNet [5] as their backbones. It is worth noting that these backbones, which were originally designed for more extensive datasets, may extract redundant information from images, leading to overfitting in relatively smaller datasets. This work proposes an innovative backbone called the Mixed Feature Network (MFN). The MFN is built upon the foundation of MobileFaceNets [6], a renowned lightweight network specifically tailored for face verification tasks. The MFN is enhanced by introducing mixed depthwise convolutional kernels [7], which exploit advantages from different size kernels. Furthermore, coordinate attention [8] is introduced into the MFN architecture to facilitate the capture of long-range dependencies. Thus, meaningful features for FER are extracted.

Moreover, FER encounters two significant challenges: small inter-class difference and large intra-class difference. To address these challenges, it is crucial to establish connections among various facial regions, such as the mouth, eyes, nose, etc. Attention mechanisms offer a potential solution in this regard. Specifically, the Dual-Direction Attention (DDA) head is applied to the proposed method, which aims to construct attention maps based on the extracted feature information. Derived from previous work [8], attention heads that generate attention maps from both vertical and horizontal directions are designed. Subsequently, multiplying the attention map obtained from the Dual-Direction Attention Network (DDAN) with the input feature map results in a new feature map. This feature map undergoes a linear Global Depthwise Convolution (GDConv) layer [6], followed by a reshaping operation. A fully connected layer is employed to generate the conclusive results. By integrating the proposed DDA head and subsequent processing steps, the model's ability can be enhanced.

Finally, this work integrates the MFN and the DDAN, presenting a novel model named the Dual-Direction Attention Mixed Feature Network (DDAMFN).

In order to visually illustrate the efficacy of the DDAMFN, a comparative analysis involving ResNet_50, MFN, and DDAMFN models was conducted. All models were trained on the AffectNet-7 dataset and tested on the same test datum. The Grad-CAM [9] was applied to capture insights into the features extracted by the respective backbone architectures. This technique facilitates the creation of heat maps highlighting important regions in an image for prediction through gradient-based localization. The outcomes of this analysis are presented in Figure 1. A comprehensive evaluation of the results showcases distinctive patterns in attention focus among the models. It is obvious that the MFN focuses on more particular areas than the ResNet_50. For the DDAMFN, the DDAN allows the MFN to locate more appropriate areas.

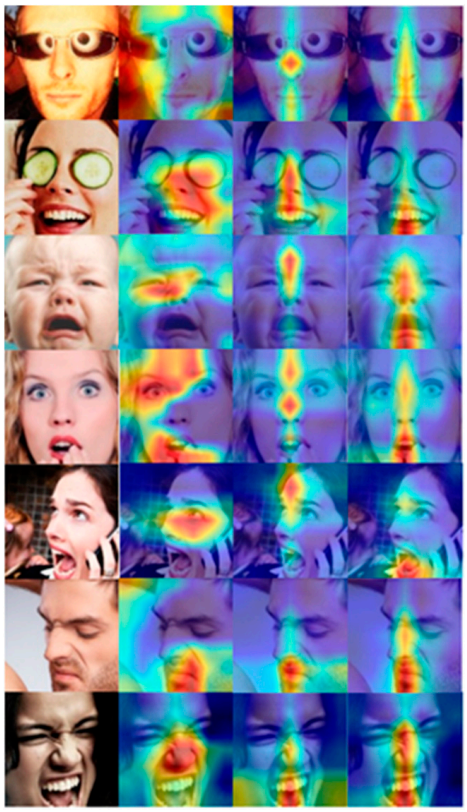

Figure 1. Heat maps on seven images: neutral (Row 1), happy (Row 2), sad (Row 3), surprise (Row 4), fear (Row 5), disgust (Row 6), and angry (Row 7). Column 1: original images. Column 2: ResNet_50. Column 3: MFN. Column 4: DDAMFN (MFN + DDAN). It is obvious that the MFN focuses on more particular areas than ResNet_50. The DDAN allows the MFN to locate more appropriate areas.

Moreover, when extensive experiments were conducted on various benchmark datasets, the DDAMFN model demonstrated a remarkable performance, establishing it as the current state-of-the-art network in FER. The contributions of our research can be summarized as follows:

(1) In order to enhance the quality of extracted features for FER, this work proposes a novel backbone network called the MFN. The MFN capitalizes on the utilization of diverse kernel sizes, thereby facilitating the acquisition of robust features. Additionally, the inclusion of coordinate attention layers within the MFN architecture enables the capture of long-range dependencies, further augmenting its effectiveness in FER tasks.
(2) To effectively detect subtle variations across different facial expressions, the DDAN is introduced. By generating attentions from two distinct directions, the DDAN aims to comprehensively capture relevant facial regions and improve discriminative capabilities for FER.
(3) A novel attention loss mechanism is applied to ensure the attention heads of DDAN are focusing on distinct areas, which leads to a notable enhancement in overall performance and discriminative power of the model.
(4) Extensive evaluations are conducted on prominent FER datasets, including AffectNet, RAF-DB, and FERPlus, to assess the performance of the DDAMFN. The experimental results demonstrate its state-of-the-art performance, with accuracy of 67.03% on AffectNet-7, 64.25% on AffectNet-8, 91.35% on RAF-DB, and 90.74% on FERPlus. These exceptional results highlight the efficacy and superiority of the DDAMFN in the realm of FER.

2. Materials and Methods

This section begins by providing a comprehensive overview of the related works pertaining to two key aspects of FER: backbone architectures and attention mechanisms. Building upon this foundation, this paper then shifts the focus to the method used to address the FER problem.

2.1. Related Works

2.1.1. FER

FER has been a prominent research area for decades. Traditional FER methods rely on handcrafted features or shallow learning techniques, including Non-Negative Matrix Factorization (NMF) [10], Local Binary Patterns (LBPs) [11], and sparse learning [12]. However, these approaches often struggle to effectively handle challenging real-world scenarios characterized by blurring and occlusions.

In recent years, deep learning techniques have revolutionized computer vision, leading to significant advancements in FER. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs) have been employed to tackle the intricate challenges of FER. Notably, state-of-the-art classification models such as VGG [4] and ResNet [5] have served as backbones for FER systems, exhibiting exceptional performance. Building upon this progress, our approach introduces a novel backbone network, the MFN, specifically designed to extract information from various kernel sizes. Notably, the MFN exhibits lightweight characteristics while achieving superior performance in FER tasks, representing a notable advancement in the field. The MFN backbone network holds significant promise for enhancing FER accuracy and effectiveness.

2.1.2. Attention Mechanism

The application of attention mechanisms in various visual tasks has gained considerable attention in recent years. Researchers have explored the integration of attention mechanisms into deep Convolutional Neural Networks (CNNs) to extract more informative features from images. Notable works in this field include the introduction of a squeeze-and-excitation block by [13], which focuses on obtaining useful information from different

channels. Another approach, SGE [14], partitions spatial-wise features into groups to establish spatial connections and enhance feature representation. Furthermore, CBAM [15] combines channel attention and spatial attention mechanisms to capture richer information through a unified framework.

Recently, the Transformer [16] has emerged as a powerful paradigm for numerous tasks, surpassing the performance of traditional Recurrent Neural Networks (RNNs) and CNNs. The Transformer's reliance on multi-head attention mechanisms has enabled it to excel in diverse domains. This recent success has prompted researchers to explore and adapt Transformer-based approaches to address various visual tasks.

Several research papers have introduced attention mechanisms to FER. Notably, [17] proposed the utilization of multiple non-overlapping region attention to capture information from distinct facial regions. Furthermore, [18] explored the application of Transformers in FER tasks. However, these existing methods face limitations in learning comprehensive information from facial images.

Derived from [6], a novel dual-direction attention head was applied to the DDAMFN. This integration facilitates the modeling of long-range dependencies, allowing for the capture of holistic and contextual facial information. By connecting the dual-direction attention head to the MFN backbone, the limitations of existing methods were overcome and more effective learning of comprehensive information from facial images was achieved.

2.2. Method

The architectural overview of the DDAMFN is depicted in Figure 2, comprising two main components: the MFN and the DDAN. Initially, facial images are fed into the MFN, which produces basic feature maps as outputs. Subsequently, attention maps are generated in both the vertical and horizontal directions through the DDAN. Eventually, attention maps are reshaped to specific dimensions, and the expression category of the images is predicted by a fully connected layer.

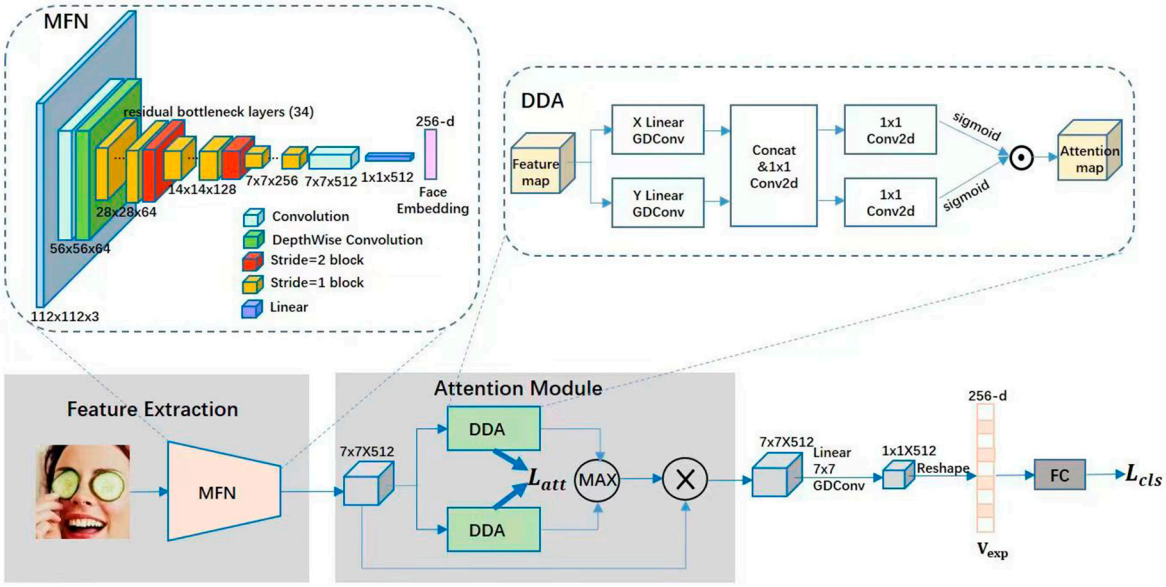

Figure 2. The overall structure of the DDAMFN. Here, "GDConv" refers to Global Depthwise Convolution. The method contains two main novel steps. Firstly, the MFN is applied to extract basic features from facial images. Next, the DDAN is built with multiple DDA heads able to generate attention maps from both horizontal and vertical orientations. Following the DDAN module, the feature undergoes a linear GDConv layer and reshapes the feature map. Finally, a fully connected layer is employed to produce the classification result.

The MFN leverages convolution kernels of varying sizes, as inspired by [7], to capture diverse spatial information from the facial images. On the other hand, the DDAN module

incorporates a series of effective dual-direction attention heads. Each attention head generates an attention map, and a comparative analysis is performed to determine the most informative attention map.

Overall, the DDAMFN framework effectively combines the feature extraction capabilities of the MFN with the discriminative power of the DDAN's attention mechanism. By integrating these components, the DDAMFN could achieve an improved performance in FER tasks.

2.2.1. MFN

In this section, a detailed description of the architecture of the MFN is provided. Considering the potential overfitting issues associated with the use of heavy network architectures on small FER datasets, a lightweight network, MobileFaceNet [6], was adopted as the foundation. As illustrated in Figure 3, a combination of two primary building blocks, a residual bottleneck and a non-residual block, was employed.

The residual bottleneck block was designed to capture complex features and facilitate information propagation within the network. The block leverages residual connections to mitigate the degradation problem and improve the flow of gradients during training. On the other hand, the non-residual block aims to enhance the model's representational capacity by incorporating non-residual connections. This block enables the MFN to capture diverse and discriminative facial features for effective FER.

By employing this architecture, the MFN strikes a balance between model complexity and generalization ability, making it well suited for FER tasks.

The upper-left section of Figure 2 presents the primary structure of the MFN, while Table 1 provides a comprehensive overview of each layer's specifications. Derived from [7], the MixConv operation, which consists of multiple-size kernels arranged as depicted in Figure 4, was integrated into our network's bottleneck. By leveraging this configuration, the MFN can effectively capture diverse and informative features from input images, surpassing the capabilities of the MobileFaceNet architecture. PreLU was also employed as the activation function, performing better than ReLU on extracting facial features [6]. Additionally, this work carefully adjusted the network depth [19] and introduced the coordinate attention mechanism [8] into each bottleneck within the MFN backbone. This attention mechanism facilitates the modeling of long-range dependencies and enables the generation of more accurate positional information compared to the Channel and Spatial Attention Module (CBAM) used in [19].

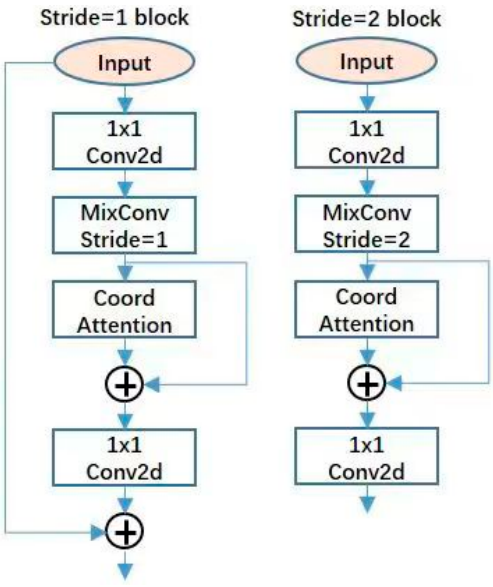

Figure 3. The bottlenecks of the MFN with the left one for stride = 1 block and the right one for stride = 2 block in Table 1.

Table 1. The proposed MFN architecture. For the table, n refers to the number of repetitions, c refers to output channels, t refers to the expansion factor, and s refers to stride.

Input	Operator	t	c	n	S
112 × 112 × 3	conv3 × 3	-	64	1	2
56 × 56 × 64	depthwise conv3 × 3	-	64	1	1
56 × 56 × 64	bottleneck (MixConv 3 × 3, 5 × 5, 7 × 7)	2	64	1	2
28 × 28 × 64	bottleneck (MixConv 3 × 3, 5 × 5)	2	128	9	1
28 × 28 × 128	bottleneck (MixConv 3 × 3, 5 × 5, 7 × 7)	4	128	1	2
14 × 14 × 128	bottleneck (MixConv 3 × 3, 5 × 5)	2	128	16	1
14 × 14 × 128	bottleneck (MixConv 3 × 3, 5 × 5, 7 × 7, 9 × 9)	8	256	1	2
7 × 7 × 256	bottleneck (MixConv 3 × 3, 5 × 5, 7 × 7)	2	256	6	1
7 × 7 × 256	Conv1 × 1	-	512	1	1
7 × 7 × 512	linear GDConv7 × 7	-	512	1	1
1 × 1 × 512	linear	-	256	1	1

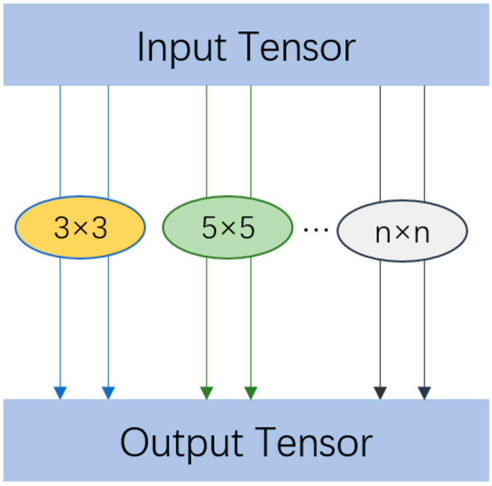

Figure 4. MixConv with multiple-size kernels.

During the preliminary training phase for face recognition, the MFN generated a 256-dimension feature map. However, during the subsequent fine-tuning stage for FER, this work only utilized the pretrained weights before the last two layers. A 7 × 7 × 512 feature map was obtained from the MFN, which serves as input for the DDAN.

2.2.2. DDAN

The DDAN consists of multiple independent DDA heads, each contributing to capturing long-range dependencies within the network. The fundamental structure of coordinate attention [8] was adopted as the basis for the DDAN module.

The detailed structure of the DDAN is depicted in the upper-right portion of Figure 2. Similar to the approach in [8], attention heads initially generate direction-aware feature maps from both the horizontal and vertical directions. However, the average pooling operation is replaced by linear GDConv, which helps to learn very different importances at

different spatial positions [6]. This modification emphasizes the importance of different facial areas, enhancing the discriminative power of the attention mechanism. Subsequently, DDA heads generate two attention maps (x_h, x_w) for both directions using the same structure as outlined in [8]. To obtain the final attention map, we can multiply x_h and x_w element-wise, resulting in an attention map of the same size as the input feature map. This process can be represented as follows:

$$a_i = x_h * x_w \tag{1}$$

Following the generation of attention maps from the multiple dual-direction heads, comparing these attention maps and selecting the one with the highest saliency is a crucial step. This process results in the creation of the optimal attention map a_m, as illustrated in Figure 2. Subsequently, the output of the DDAN is obtained through element-wise multiplication between the input feature map and a_m.

To ensure that each dual-direction head focuses on distinct facial areas, a novel loss function is introduced to the DDAN module, referred to as the attention loss.

Attention loss: The Mean Squared Error (MSE) loss is calculated between each pair of attention maps generated from different dual-direction heads. The attention loss is then defined as the reciprocal of the sum of these MSE losses, which can be mathematically expressed as follows:

$$L_{att} = \frac{1}{\sum_{i=0}^{n} \sum_{k=0}^{n} MSE(a_i, a_k)}, \ (i \neq j) \tag{2}$$

where n is the number of attention heads. a_i and a_k are attentions maps yielded from two different heads.

2.2.3. Loss

As depicted in Figure 2, the feature map of size $7 \times 7 \times 512$, obtained from the DDAN, undergoes a linear GDConv layer and a linear layer. This transformed feature map is then reshaped to a 512 d vector. The class confidence is obtained via a fully connected layer.

Regarding the loss function, the standard cross-entropy loss is employed in the training process. This loss function effectively measures the discrepancy between predicted class probabilities and the ground truth labels, facilitating the optimization of the model's parameters. The overall loss function can be expressed as follows:

$$L = L_{cls} + \lambda_a L_{att}, \tag{3}$$

where L_{cls} stands for standard cross entropy loss and L_{att} is attention loss. λ_a is a hyperparameter. The default of λ_a is 0.1.

3. Results

In this section, comprehensive experimental results, attained by the DDAMFN on three widely used benchmark datasets, AffectNet, RAF-DB, and FERPlus, are displayed. The experimental evaluation began with a series of ablation experiments, where the individual contributions of each component within the DDAMFN were analyzed. This allowed us to assess the significance of each component in enhancing the overall performance of the DDAMFN. Subsequently, a comparative analysis with other state-of-the-art networks was performed to ascertain the superiority of the DDAMFN.

The meticulous experimentation demonstrated the efficacy and effectiveness of the DDAMFN in FER. The results obtained highlight the substantial contributions made by the various components within the DDAMFN. Moreover, the comparison with existing networks substantiates the exceptional performance of the DDAMFN, solidifying its position as the top-performing model in the field of FER.

3.1. Datasets

AffectNet [20]: AffectNet is currently the most extensive publicly available dataset for FER. It consists of two distinct benchmarks: AffectNet-7 and AffectNet-8. AffectNet-7 comprises 283,901 images for training and 3500 images for testing, with seven emotion categories including neutral (74,874), happy (134,415), sad (25,459), surprise (14,090), fear (6378), angry (24,882), and disgust (3803). AffectNet-8 introduces an additional category of "contempt" (3750) and expands the training set to 287,651 images, along with 4000 images for testing.

AffectNet was chosen due to its prominence as one of the largest publicly available FER datasets. Its extensive collection of labeled facial images covers a wide range of expressions, and it contains multiple benchmarks (AffectNet-7 and AffectNet-8) with varying numbers of emotion categories.

RAF-DB [21]: RAF-DB is a real-world facial expression database sourced from the Internet. It encompasses 29,672 facial images annotated with seven basic emotion labels and eleven compound emotion labels. Within RAF-DB, there are 12,271 training samples (surprise (1290), fear (281), disgust (717), happy (4772), sad (1982), angry (705), and neutral (2524)) and 3068 testing samples available for FER.

RAF-DB was chosen for its real-world diversity and challenging conditions. It is a valuable dataset for evaluating FER methods under varying factors such as pose, lighting, and occlusion.

FERPlus [22]: FERPlus derives from the FER2013 dataset [23], originally compiled using the Google image search API. FERPlus incorporates 28,709 training images and 3589 validation images. Each image within FERPlus is annotated with one of the eight emotion categories: neutral, happy, sad, surprise, fear, angry, disgust, or contempt. In this research, 27,298 images were used for training (neutral (9462), happy (7879), sad (3262), surprise (3488), fear (592), disgust (141), and contempt (136)).

FERPlus was selected as it enhances the FER2013 dataset by addressing some of its limitations. The dataset contains more balanced and accurate annotations.

These benchmark datasets serve as fundamental resources for evaluating the performance of FER models, allowing us to assess the effectiveness and generalization capabilities of our method.

3.2. Implementation

In preprocessing, RetinaFace [24] is used to detect faces and landmarks (5 points for two eyes, noses, and two mouth corners) within the AffectNet, RAF-DB, and FERPlus datasets. All face images were aligned to a standardized size of 112 $\times$ 112 pixels. To mitigate overfitting, various data augmentation techniques were employed. Specifically, for AffectNet, horizontal flipping, random affine, and erasing were applied. And horizontal flipping, random rotations, and erasing were set for RAF-DB. For FERPlus, horizontal flipping, color jittering, random rotations, and erasing were utilized. These augmentation techniques enhanced the robustness and generalization ability of DDAMFN during training.

To ensure a fair comparison with other backbone architectures, the MFN backbone was pretrained on the Ms-Celeb-1M dataset [25]. This pretraining step enabled a consistent benchmarking of the DDAMFN.

All experiments were conducted using the PyTorch 1.8.0 toolbox, and the models were trained on a server equipped with a TESLA P40 24G GPU. All tasks were trained for 40 epochs, while the number of attention heads in the DDAN module was set to the default value of 2.

During the training process, the ADAM optimization algorithm was employed to optimize the models' parameters. Specifically, for AffectNet-7 and AffectNet-8, an initial learning rate of 0.0001 and a batch size of 256 were used. For RAF-DB and FERPlus, a larger learning rate of 0.01 was set. These parameter settings were selected to facilitate efficient and effective model optimization across the respective datasets.

The DDAMFN was trained 5 times on each dataset, and the average accuracy (the proportion of correctly predicted instances out of the total instances in the test datum) was calculated to display an overall view of the model's performance.

3.3. Ablation Studies

Effectiveness of the MFN: to evaluate the effectiveness of the MFN backbone, a series of comparative experiments were conducted.

Table 2 presents the performance of different backbones on the RAF-DB dataset. The MFN achieved an accuracy of 90.32% on RAF-DB, surpassing the performance of the other three backbones. Furthermore, the MFN backbone consists of only 3.973 million parameters, which is the second lowest among the compared backbones. It is significantly smaller than ResNet-18 and ResNet-50. Additionally, the computational complexity of the MFN, as measured by the number of floating-point operations (FLOPs), is merely 550.74 million. This is higher than the computational complexity of MobileFaceNet but substantially smaller than that of the other two networks.

Table 2. Evaluation (%) of the MFN and other networks on RAF-DB.

Methods	Accuracy (%)	Params	Flops
MobileFaceNet	87.52	1.148 M	230.34 M
ResNet-18	87.47	16.78 M	2.6 G
ResNet-50	89.63	41.56 M	6.31 G
MFN (our backbone)	90.32	3.973 M	550.74 M
DDAMFN (ours)	91.35	4.106 M	551.22 M

These results demonstrate that the MFN backbone is not only accurate but also lightweight and computationally efficient. It outperformed other backbones in terms of accuracy while maintaining a significantly smaller number of parameters and lower computational complexity. This highlights the effectiveness of the MFN as the preferred backbone for our subsequent experiments.

Furthermore, Table 2 also reveals that even with the addition of the DDAN to the MFN backbone, the number of parameters and FLOPs remained considerably smaller compared to ResNet models. This suggests that the combined model retains its lightweight nature while maintaining high accuracy.

Effectiveness of the DDAN: to ascertain the effectiveness of the DDAN, an ablation study was conducted to evaluate the impact of both the MFN and the DDAN on the RAF-DB and AffectNet-7 datasets (as presented in Table 3).

Table 3. Evaluation (%) of the MFN and the DDAN on RAF-DB and AffectNet-7.

MFN	DDAN	RAF-DB	AffectNet-7
√	-	90.32	66.19
√	√	91.35	67.03

The results shown in Table 3 indicate that the inclusion of the DDAN in the backbone network led to performance improvements of 1.06% and 1.04% for RAF-DB and AffectNet-7, respectively, compared to using the MFN alone. These findings suggest that the DDAN plays a crucial role in enhancing the performance of the MFN by enabling the generation of more comprehensive attention maps from the extracted features. By incorporating DDAN, the model can better focus on relevant regions and capture essential information, resulting in improved recognition accuracy.

These results substantiate the effectiveness of the DDAN in augmenting the capabilities of the MFN and underscore its contribution to the overall performance enhancement in FER tasks.

Number of the attention heads: In order to examine the influence of the number of the DDA heads on the model's performance, experiments with varying numbers of DDA heads on the RAF-DB and AffectNet-7 datasets were conducted. The results are presented in Table 4.

Table 4. Evaluation (%) of influence of number of DDA heads on RAF-DB and AffectNet-7.

-	RAF-DB	AffectNet-7
0	90.32	66.19
1	90.67	66.32
2	91.35	67.03
3	91.11	67.06
4	91.21	67.06

As seen in Table 4, the model with two DDA heads achieved a notably superior performance on the RAF-DB dataset compared to models with different numbers of DDA heads. Furthermore, on the AffectNet-7 dataset, the two-head model performed only slightly worse (0.03% lower) than the models with three or four DDA heads. These findings clearly indicate that the DDAN architecture with two DDA heads outperformed the models with different numbers of DDA heads in terms of recognition accuracy on both datasets.

The results from Table 4 demonstrate that the selection of two DDA heads strikes a favorable balance between capturing sufficient attention information and maintaining optimal model performance. It highlights the significance of the appropriate number of DDA heads in achieving superior performance in FER tasks.

Effectiveness of loss function for the DDAN: To assess the effectiveness of the loss function employed for the DDAN, an evaluation is shown in Table 5. The results indicate that the attention loss function significantly impacts the performance of the DDAN.

Table 5. Ablation studies for the loss function in the DDAN.

Methods	RAF-DB	AffectNet-7
-	90.86	66.39
Attention loss	91.35	67.03

From the obtained results, it is evident that the novel attention loss function plays a crucial role in enhancing the performance of the DDAN. The inclusion of this loss function leads to improved performance, emphasizing the importance of guiding the attention heads to focus on different areas and facilitating better discrimination between facial expressions.

These findings highlight the effectiveness of the attention loss function in optimizing the attention mechanism within the DDAN. By encouraging attention heads to attend to diverse facial regions, the loss function enhances the discriminative power and overall performance of the DDAN architecture in FER tasks.

3.4. Comparison with State-of-the-Art Methods

A comprehensive comparison of the DDAMFN model with other existing models on AffectNet, RAF-DB, and FERPlus datasets is shown in Tables 6–9.

Table 6. Performance comparison for AffectNet-7.

Methods	Accuracy (%)
DLN [26]	63.7
MViT [27]	64.57
DACL [28]	65.20
DAN [17]	65.69
TransFER [18]	66.23
Emotion-GCN [29]	66.46
DDAMFN (ours)	67.03

Table 7. Performance comparison for AffectNet-8.

Methods	Accuracy (%)
RAN [30]	59.50
SCN [31]	60.23
PSR [32]	60.68
MViT [27]	61.40
DAN [17]	62.02
EfficientNet-B2 [33]	63.03
DDAMFN (ours)	64.25

Table 8. Performance comparison for RAF-DB.

Methods	Accuracy (%)
RAN [30]	86.90
SCN [31]	87.03
DACL [28]	87.78
MViT [27]	88.62
PSR [32]	88.98
DAN [17]	89.70
TransFER [18]	90.91
DDAMFN (ours)	91.35

Table 9. Performance comparison for FERPlus.

Methods	Accuracy (%)
PLD [22]	85.10
RAN [30]	88.55
SeNet50 [34]	88.80
RAN-VGG16 [30]	89.16
SCN [31]	89.35
KTN [35]	89.70
TransFER [18]	90.83
DDAMFN (ours)	90.74

The DDAMFN model achieved an outstanding performance on three datasets, attaining an accuracy of 67.03% on AffectNet-7, 64.24% on AffectNet-8, and 91.35% on RAF-DB. These results represent the best performance among existing models on these benchmarks. Notably, on the AffectNet dataset, the DDAMFN outperformed the Emotion-GCN [29] by 0.57% on AffectNet-7 and the EfficientNet-B2 [33] by 1.22% on AffectNet-8. Moreover, on the RAF-DB dataset, the DDAMFN surpassed the previous best result achieved by TransFER [18] by 0.44%. These improvements establish the DDAMFN model as the state-of-the-art approach for FER on these specific benchmarks.

Furthermore, on the FERPlus dataset, the DDAMFN model achieved the second-best result of 90.74%, falling only 0.1% short of the TransFER model. These findings indicate the effectiveness and generalization capabilities of the DDAMFN model across different datasets.

The DDAMFN model has been proven to be the state-of-the-art approach for FER, outperforming existing models on AffectNet and RAF-DB and achieving competitive results on FERPlus. These results demonstrate the effectiveness and robustness of the DDAMFN model in addressing the challenges of FER across diverse datasets.

3.5. K-Fold Cross-Validation

To rigorously evaluate the effectiveness and reliability of the DDAMFN, K-fold cross-validation was conducted on the training dataset. In this process, dataset D was randomly partitioned into k mutually exclusive subsets of equal size. Subsequently, $k - 1$ subsets were employed for training the model and the remaining subset was used for testing. This procedure was repeated for every subset, and the results were collected to calculate the average accuracy. This validation technique ensures that all data points participate in both training and prediction, effectively mitigating the risk of overfitting.

To assess the robustness of the DDAMFN, K-fold cross-validation was performed on the RAF-DB and FERPlus datasets, as these smaller datasets are susceptible to overfitting, potentially leading to instability. The results are presented in Table 10 (K = 10).

Table 10. The results of K-fold cross-validation.

	Accuracy (%)										
	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Fold 6**	**Fold 7**	**Fold 8**	**Fold 9**	**Fold 10**	**Average**
RAF-DB	90.69	90.23	89.89	90.82	89.83	91.02	91.82	90.55	90.55	90.95	90.635
FERPlus	89.78	90.98	90.36	90.7	90.6	90.43	91.11	89.95	90.29	90.6	90.48

Analysis of the results in Table 10 revealed that the DDAMFN achieved an impressive accuracy of 90.635% on RAF-DB and 90.48% on FERPlus. These outcomes demonstrate consistently high performance on both datasets, underscoring the reliability of our approach in FER tasks.

In conclusion, the K-fold cross-validation further confirmed the effectiveness and robustness of the DDAMFN. The notable performances achieved on RAF-DB and FERPlus datasets attest to the high reliability of the DDAMFN in handling FER tasks.

4. Discussion

In this section, the possible explanations behind the experimental results are discussed by examining previous studies and hypotheses. Additionally, the potential areas in which our findings could exert influence are explored. Finally, the future research directions stemming from our investigation of class imbalance and cross-dataset validation are emphasized.

4.1. Possible Explanation and Future Influences

From Figure 1, this work speculates that the remarkable results achieved by the DDAMFN in FER can be attributed to its emphasis on specific facial regions that are highly relevant to human facial expressions. To further substantiate this supposition, an empirically rigorous adversarial attack study aimed at validating the significance of the focused areas for classification was conducted.

First, we trained the DDAMFN on AffectNet-7. Subsequently, the AffectNet-7 test dataset was used as test images. These images were input to the DDAMFN to generate heat maps via the method depicted in Figure 1, which facilitates the extraction of both focused and unfocused areas within the images. In the pursuit of investigating the impact of manipulating these areas, the original unfocused regions were retained without alterations.

However, random noise was introduced to the focused regions through the following transformation:

$$new\_focused\_areas = original\_focused\_areas * (1 - \varepsilon) + random\_noise * \varepsilon \quad (4)$$

Here, ε represents a weight parameter, with values chosen as 0.1, 0.2, 0.3, 0.4, 0.5, and 1.0. To make the experiment more rigorous, the converse scenario, called "negative samples", was tested. This involved ensuring the focused regions remained unchanged while introducing random noise to the unfocused areas, as defined by

$$new\_unfocused\_areas = random\_noise * 1.0 \quad (5)$$

The outcomes of this perturbation process are visually depicted in Figure 5.

Figure 5. Renderings for the adversarial attack study. Left to right: $\varepsilon = 0.1$, $\varepsilon = 0.2$, $\varepsilon = 0.3$, $\varepsilon = 0.4$, $\varepsilon = 0.5$, $\varepsilon = 1.0$ and the negative sample.

After processing, the images were tested on the DDAMFN, and accuracies were recorded. The results are shown in Table 11.

Table 11. The results of the adversarial attack study.

ε	0	0.1	0.2	0.3	0.4	0.5	1.0	Negative Sample
Accuracy	67.03%	62.30%	50.44%	28.47%	20.49%	18.95%	16.95%	62.26%

As seen in Table 11, for the focused areas, as the weight of noise increased, the accuracy of noise decreased. This means that the focused areas are very critical for classification. For unfocused areas, although they were covered with 100% random noise, the accuracy only dropped by 4.77%. The results indicate that the key regions focused by the DDAMFN are more important for FER than other regions, which allowed our model to have a higher performance.

The implications of our findings extend beyond the scope of FER. The DDAMFN holds the potential to influence future research in this field as well as other classification tasks. Moreover, the model exhibits promising applications in areas such as Emotional Robots, Human–Computer Interaction (HCI), Security Monitoring, and Healthcare. Our goal is not only to excel in dataset performance, but also to make meaningful contributions to improving the world we live in.

4.2. Future Directions

In this section, the insights gained from analyzing class imbalance and cross-dataset validation results are leveraged to delineate the directions for future development.

4.2.1. Class Imbalance

In the classification task, the accuracy is not fully persuasive since the performance for each category is not shown. Thus, a confusion matrix (which shows classification variations across different expressions) was leveraged to explore the class imbalance problems in FER. Figure 6a–d presents the confusion matrix of the DDAMFN model tested on four datasets, offering valuable insights into the model's performance.

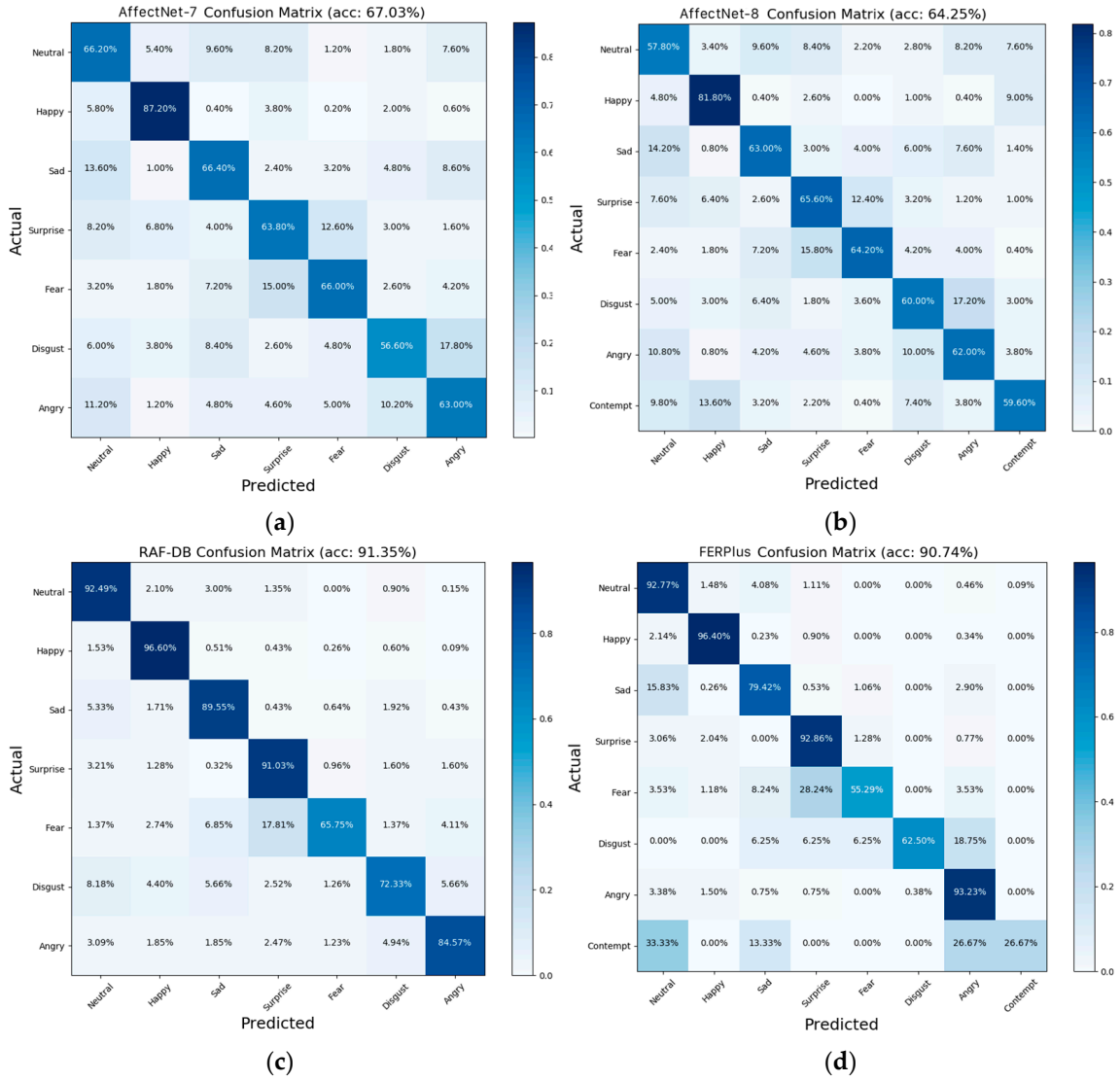

Figure 6. The confusion matrix of the DDAMFN tested on different datasets. (**a**) AffectNet-7; (**b**) AffectNet-8; (**c**) RAF-DB; (**d**) FERPlus.

The analysis of the confusion matrix revealed that the DDAMFN achieved the highest performance for the "Happy" expression category. This indicates that "Happy" expressions are relatively easier for the DDAMFN to recognize accurately. However, for the AffectNet dataset, the DDAMFN had a lower performance for the "Disgust" category, which is often confused with the "Angry" category. Similarly, on the RAF-DB and FERPlus datasets, our model encountered challenges with the "Fear" and "Disgust" categories. These categories exhibit confusion patterns with "Surprise" and "Angry,", respectively. The underlying rationale for the confusion may be attributed to the presence of shared and overlapping signals employed to convey these facial expressions, which results in some expressions possessing perceptual similarities, leading to challenges in accurate differentiation [36].

To further enhance the DDAMFN's performance, future research will focus on maintaining the model's existing strengths while addressing these specific confusion problems, for example, using auxiliary action unit graphs [37] to correct annotations before training. During the face detection process, we could also introduce a multi-point landmark detection method (e.g., 68 landmarks, localizing all facial organs and extracting particular features) to fuse with the features from the backbone, in order to better discriminate between confused expressions.

By addressing the observed confusion problems, this work aimed to achieve a better performance and deeper understanding of the intricacies involved in accurately recognizing facial expressions. Future research efforts will contribute to advancing the field of FER and further improving the performance of the DDAMFN model.

4.2.2. Cross-Dataset Validation

To assess the generalization ability of the DDAMFN, cross-dataset validations on four datasets, AffectNet-7, AffectNet-8, RAF-DB, and FERPlus, were conducted. The datasets were divided into two groups, (AffectNet-7, RAF-DB) and (AffectNet-8, FERPlus), based on the number of facial expression categories. Due to the larger size of AffectNet-7/8 compared to RAF-DB/FERPlus, training was conducted on AffectNet-7/8, and testing was conducted on RAF-DB/FERPlus for the cross-dataset validation. The outcomes of this evaluation are presented in Table 12.

Table 12. The results of cross-dataset validation.

Trained on	Tested on	Accuracy (%)
AffectNet-7	RAF-DB	75.6
AffectNet-8	FERPlus	63.14

Table 12 shows that the DDAMFN achieved 75.6% accuracy on RAF-DB and 63.14% accuracy on FERPlus, showcasing a tendency for universality. After utilizing more images and improving the layers and structures in our network, the DDAMFN exhibited enhanced functionality in FER for fewer or even zero sample tasks. Additionally, the integration of generation models and Contrastive Language–Image Pretraining (CLIP) techniques could further enhance the performance of the DDAMFN on these tasks, unlocking new possibilities for advancing the state of the art in FER. For example, generation models can create synthetic facial expression images that are visually similar to real ones. By generating new samples, particularly for underrepresented classes, the number imbalance issue can be alleviated, leading to better generalization and performance. Also, CLIP's learned representations can be transferred to FER tasks. This transfer enables our model to benefit from the knowledge encoded in CLIP, contributing to better feature extraction and classification.

5. Conclusions

In this paper, a novel and effective approach for FER, termed the DDAMFN, has been proposed. The DDAMFN comprises two key components: the MFN and the DDAN. The MFN leverages the benefits of different-sized kernels to generate comprehensive and discriminative features for expression classification. Meanwhile, the DDAN captures long-range dependencies through newly introduced DDA heads.

Through extensive experiments conducted on four FER datasets (AffectNet-7, AffectNet-8, RAF-DB, and FERPlus), the DDAMFN has demonstrated state-of-the-art performance (67.03% on AffectNet-7, the best, 0.57% surpassing the second; 64.24% on AffectNet-8, the best, 1.22% surpassing the second; 91.35% on RAF-DB, the best, 0.44% surpassing the second; 90.74% on FERPlus, almost same as the best), surpassing existing approaches. These results confirm the effectiveness and superiority of the DDAMFN in the field of FER. By proposing a novel lighter backbone and applying Mixconv and coordinate attention in the model for FER, the DDAMFN will contribute to the advancement of the network structure for FER and serve as a catalyst for future developments in computer vision tasks. Furthermore, we eagerly anticipate the application of the model in diverse domains of artificial intelligence, fostering progress in various applications and facilitating advancements in AI-driven technologies.

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

Funding: The research was funded by the National Key R&D Program of China (2021YFB2800304, 2021YFB2800303) and the National Natural Science Foundation of China (62274153). The APC was funded by the National Key R&D Program of China (2021YFB2800304, 2021YFB2800303).

Data Availability Statement: All data included in this study may be made available upon request via contacting the corresponding author.

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

References

1. Andronie, M.; Lăzăroiu, G.; Karabolevski, O.L.; Ștefănescu, R.; Hurloiu, I.; Dijmărescu, A.; Dijmărescu, I. Remote Big Data Management Tools, Sensing and Computing Technologies, and Visual Perception and Environment Mapping Algorithms in the Internet of Robotic Things. *Electronics* **2023**, *12*, 22. [CrossRef]
2. Pelău, C.; Dabija, D.C.; Ene, I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics on the acceptance of artificial intelligence in the service industry. *Comput. Hum. Behav.* **2021**, *122*, 106855. [CrossRef]
3. Dijmărescu, I.; Iatagan, M.; Hurloiu, I.; Geamănu, M.; Rusescu, C.; Dijmărescu, A. Neuromanagement decision making in facial recognition biometric authentication as a mobile payment technology in retail, restaurant, and hotel business models. *Oeconomia Copernic.* **2022**, *13*, 225–250. [CrossRef]
4. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–14.
5. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
6. Chen, S.; Liu, Y.; Gao, X.; Han, Z. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In Proceedings of the Chinese Conference on Biometric Recognition, Urumqi, China, 11–12 August 2018; pp. 428–438.
7. Tan, M.; Le, Q.V. Mixconv: Mixed depthwise convolutional kernels. In Proceedings of the 30th British Machine Vision Conference 2019, Cardiff, UK, 9–12 September 2019.
8. 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 (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
9. 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.
10. Zhi, R.; Flierl, M.; Ruan, Q.; Kleijn, W.B. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. In Proceedings of the IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Anchorage, AK, USA, 9–12 October 2011; pp. 38–52.
11. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
12. Zhong, L.; Liu, Q.; Yang, P.; Liu, B.; Huang, J.; Metaxas, D.N. Learning active facial patches for expression analysis. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2562–2569.
13. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
14. Li, X.; Hu, X.; Yang, J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv* **2019**, arXiv:1905.09646.
15. 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.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
17. Wen, Z.; Lin, W.; Wang, T.; Xu, G. Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics* **2023**, *8*, 199. [CrossRef]
18. Xue, F.; Wang, Q.; Guo, G. Transfer: Learning relation-aware facial expression representations with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 3601–3610.
19. Li, X.; Wang, F.; Hu, Q.; Leng, C. Airface: Lightweight and Efficient Model for Face Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.

20. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-ex-pression databases from movies. *IEEE Multimed.* **2012**, *19*, 34–41. [CrossRef]
21. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. Image Process.* **2018**, *28*, 356–370. [CrossRef] [PubMed]
22. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 279–283.
23. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Republic of Korea, 3–7 November 2013; pp. 117–124.
24. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-Shot Multi-Level Face Localization in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA; 2020; pp. 5203–5212.
25. Guo, Y.; Zhang, L.; Hu, Y.; He, X.; Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 87–102.
26. Zhang, W.; Ji, X.; Chen, K.; Ding, Y.; Fan, C. Learning a Facial Expression Embedding Disentangled from Identity. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 6755–6764.
27. Li, H.; Sui, M.; Zhao, F.; Zha, Z.; Wu, F. Mvt: Mask vision transformer for facial expression recognition in the wild. *arXiv* **2021**, arXiv:2106.04520.
28. Farzaneh, A.H.; Qi, X. Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 2402–2411.
29. Panagiotis, A.; Filntisis, P.P.; Maragos, P. Exploiting Emotional Dependencies with Graph Convolutional Networks for Facial Expression Recognition. *arXiv* **2021**, arXiv:2106.03487.
30. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *Proc. IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef] [PubMed]
31. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 6897–6906.
32. Vo, T.H.; Lee, G.S.; Yang, H.J.; Kim, S.H. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access* **2020**, *8*, 131988–132001. [CrossRef]
33. Savchenko, A.V.; Savchenko, L.V.; Makarov, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Trans. Affect. Comput.* **2022**, *13*, 2132–2143. [CrossRef]
34. Albanie, S.; Nagrani, A.; Vedaldi, A.; Zisserman, A. Emotion recognition in speech using cross-modal transfer in the wild. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 292–301.
35. Li, H.; Wang, N.; Ding, X.; Yang, X.; Gao, X. Adaptively learning facial expression re-presentation via C-F labels and distillation. *IEEE Trans. Image Process.* **2021**, *30*, 2016–2028. [CrossRef] [PubMed]
36. Luo, Q.; Dzhelyova, M. Consistent behavioral and electrophysiological evidence for rapid perceptual discrimination among the six human basic facial expressions. *Cogn. Affect. Behav. Neurosci.* **2020**, *20*, 928–948. [CrossRef] [PubMed]
37. Liu, Y.; Zhang, X.; Kauttonen, J.; Zhao, G. Uncertain label correction via auxiliary action unit graphs for facial expression recognition. In Proceedings of the 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 777–783.