applied sciences

MDPI

Article

Attention-Guided Network Model for Image-Based Emotion Recognition

Herag Arabian [1,*], Alberto Battistel [1], J. Geoffrey Chase [2] and Knut Moeller [1,2,3]

[1] Institute of Technical Medicine (ITeM), Furtwangen University, 78054 Villingen-Schwenningen, Germany
[2] Department of Mechanical Engineering, University of Canterbury, Christchurch 8041, New Zealand
[3] Department of Microsystems Engineering, University of Freiburg, 79110 Freiburg, Germany
* Correspondence: h.arabian@hs-furtwangen.de; Tel.: +49-(0)7720-307-4637

Featured Application: This work is being developed as part of a closed-loop system to be used in the therapeutic treatment of people with autism spectrum disorder.

Abstract: Neural networks are increasingly able to outperform traditional machine learning and filtering approaches in classification tasks. However, with the rise in their popularity, many unknowns still exist when it comes to the internal learning processes of the networks in terms of how they make the right decisions for prediction. As a result, in this work, different attention modules integrated into a convolutional neural network coupled with an attention-guided strategy were examined for facial emotion recognition performance. A custom attention block, AGFER, was developed and evaluated against two other well-known modules of squeeze–excitation and convolution block attention modules and compared with the base model architecture. All models were trained and validated using a subset from the OULU-CASIA database. Afterward, cross-database testing was performed using the FACES dataset to assess the generalization capability of the trained models. The results showed that the proposed attention module with the guidance strategy showed better performance than the base architecture while maintaining similar results versus other popular attention modules. The developed AGFER attention-integrated model focused on relevant features for facial emotion recognition, highlighting the efficacy of guiding the model during the integral training process.

Keywords: attention module; emotion recognition; deep learning; digital health; mental well-being; network prediction analysis

Citation: Arabian, H.; Battistel, A.; Chase, J.G.; Moeller, K. Attention-Guided Network Model for Image-Based Emotion Recognition. *Appl. Sci.* **2023**, *13*, 10179. https://doi.org/10.3390/app131810179

Academic Editors: Paula Viana, Pedro Carvalho and Teresa Chambel

Received: 17 August 2023
Revised: 7 September 2023
Accepted: 8 September 2023
Published: 10 September 2023

1. Introduction

Neural networks have shown the ability to outperform traditional approaches in different computational tasks, especially classification. Popular pre-trained network models include AlexNet [1], VGG16 [2], and residual networks such as ResNet50 [3]. However, with the rise in their popularity there are still many unknowns when it comes to the integral learning processes of the network when making appropriate decisions, which is why they are typically regarded as black boxes [4]. To try and understand the decision process, explainable artificial intelligence (XAI) models have been introduced. These models use prediction visualization techniques to visualize the regions of influence from the image on the predicted outcome at different layers within the architecture.

XAI helps in understanding which region the network based its decision upon, and based on this knowledge changes in the hyper-parameters can be made to improve the results. This fine-tuning of parameters is an exhaustive time-consuming process for the identification of optimal settings. However, with the introduction of attention modules in recent years [5–9], a shortcut to extensive parameter tuning for better robustness and regions of focus has become available. These attention modules can be easily integrated

into any existing network architecture to improve network representation and regions of focus by concentrating on informative features and diminishing less important ones [5,6].

One of the well-known and basic applications of machine learning is emotion intelligence [10]. In this study, an attention module developed for facial emotion recognition (FER) is examined for the performance improvement of a base convolutional neural network (CNN) model. The face is chosen for emotion recognition as it is estimated that 65% of emotions are expressed nonverbally via facial expressions (55%) and physiological signals (10%) [11]. In this case, an efficient, robust FER system is desired to create a feedback variable for integration in a closed-loop system to help treat people with autism spectrum disorder (ASD) [12]. While facial expressions are often emphasized as primary nonverbal communicators of emotion, it should be mentioned that whole-body expressions also hold equal importance in understanding others' emotional states, offering distinct advantages such as recognizing emotions when faces are occluded and powerfully conveying action intentions, as shown in previous research [13,14].

The dynamics of emotional stimuli, particularly their role and influence in clinical contexts, has evolved in terms of our understanding over the years. It was widely accepted that emotional stimuli, especially those perceived as threats, wielded considerable power in directing selective attention, prioritizing their processing, and consequently inducing automatic reactions [15,16]. However, recent research offers an intriguing perspective, noting individuals' reactions to such stimuli are incited primarily when they align with personal objectives. This result suggests a context-dependent effect of facial emotions [17,18].

For instance, a previous study [18] explored this concept through a go/no-go task, revealing that emotional cues influenced participants' behavior only when pertinent to the task at hand. Parallel findings have been observed in other domains, such as reaching arm movements and critical executive functions like inhibitory control [17,19]. These revelations underscore an essential nuance: the attentional system's interaction with emotions is intricately tied to its immediate context and the goals of the individual. Consequently, the impact of information on emotion recognition might be contingent upon its relevance to the task in focus. However, given this distinction, the focus of this article is aligned to staged emotions rather than spontaneous reactions to stimuli.

This treatment concept is reinforced in [20], in which FER was used in a closed system that provided a positive perspective towards utilizing such a concept for ASD children's interactions and behavioral monitoring. A small pilot study conducted in [21] showed that the use of a closed-loop virtual reality environment had encouraging results, suggesting such a system is potentially beneficial for the support of ASD patients in building communication skills. However, to date, these systems lack feedback of the subject's response to the stimulus input, while a robust, efficient FER system can provide this feedback.

The key to any good prediction model is its ability to interpret any input and provide the correct output regardless of external influences, such as lighting, orientation, background noise, and color combinations. In short, this can be summarized as model robustness. In some studies, robustness has become associated with adversarial robustness, which is the ability of the network to maintain its decisions when the input data are distorted or perturbed [22]. In this study, robustness is defined as the ability of a trained network classifier in identifying emotions from data not seen during the training process, and thus novel to the system yet still containing the same emotions. This robustness definition is also synonymous with the network's ability to generalize.

The pre-trained VGG16 [2] architecture modified for FER was used as the base model for comparison in this study. Three separate models were created from the base by incorporating three different attention modules. The first model adopted a custom FER attention module, the AGFER, which was developed specifically for guiding the training process for robust emotion classification. The second model, the SEFER, was integrated with the squeeze and excitation block (SE) [5]. Finally, the third model, the CBAMFER, was integrated with a convolution block attention module (CBAM) [6].

The well-accepted emotion database of OULU-CASIA [23] was used for training and validation, while the FACES [24] database was used for testing and robustness analysis. The use of two databases in this way ensures external influences are different and thus tests robustness better than using a subset of similarly obtained data. The input images were first pre-processed according to the method in [12] and split into a five-fold cross-validation [25] scheme, with 80% training and 20% validation sets thus created for each segment.

This study aims to show that attention modules help guide the decision process of the network to focus on areas of significance to the classification task, thereby achieving better overall accuracy and efficacy.

Hereafter, Section 2 defines the methods for the attention module, network architecture, image pre-processing, and analysis criteria. Section 3 presents the results, Section 4 presents discussion of these results, and Section 5 performs the ablation study. Finally, Section 6 presents the conclusion.

1.1. Related Work

1.1.1. Facial Emotion Recognition (FER)

In [26], the performance of traditional machine learning techniques involving K-nearest neighbors and support vector machines with the histograms of oriented gradients as a feature extractor were examined against the performance of an AlexNet [1] CNN for FER. The results obtained showed that the traditional methods performed just as well as neural networks in terms of performance. In [12], the focus area of the networks' decision was studied by visualizing the predictions through the implementation of gradient-weighted class activation mapping (Grad-CAM) [27] on different image inputs. Results showed that enhanced image pre-processing made the network more robust in terms of focusing on the areas of particular significance to emotion classification.

Different approaches and models for emotion recognition have been studied over the years. In [28] a study was conducted on the FER2013 [29] dataset and achieved state-of-the-art results by using the VGGNet architecture and conducting extensive fine-tuning of network hyper-parameters. In [30], a shallow two-network model architecture was studied, where one network removed background data and the second generated point features on the remaining face image, where accuracies of up to 96% were recorded using a combined dataset. In [10], a two-network strategy was adopted where one network was trained on a sequence of images and the second on the geometry-based trajectory computation of facial landmarks, with a joint fine-tuning method subsequently proposed that achieved better performance and state-of-the-art results on the OULU-CASIA [23] database (reaching accuracies of 81.46% in a 10-fold cross-validation). In [31], the novel approach of a peak-piloted network was proposed, where the peak and non-peak image frames from a sequence were considered as a paired input to a network based on the GoogleNet [32] architecture, where tests on the OULU-CASIA [23] database showed accuracies of up to 84.59%.

1.1.2. Attention Mechanisms

Learning mechanisms have recently shown improvement in CNN representations by capturing spatial correlations between features [5]. In [5], the relations between channels in network design were investigated and, as a result, an SE block was introduced to improve the performance of networks by computing the interdependencies between the channels. The tests were conducted on the ImageNet 2012 [33] dataset using the architectures of VGGNet [2] and different residual networks. The results showed the SE blocks incorporated into the architecture outperformed the baseline models while slightly affecting computational performance. In [6], the CBAM was introduced, in which the work extended SE blocks by focusing on the spatial as well as the channel information, arguing spatial attention is important in deciding where the network must focus. The tests were conducted on the ImageNet 1K [34] database using ResNet50 [3] as the base architecture. The results showed better performance over both the base and the SE–base integrated models with the CBAM having finer attention than SE.

In [7], two attention modules were proposed for spatial and channel features, which learn the relations between the inter-channels and inter-spatial dimensions sequentially and output refined features. Residual networks were used as the framework of the models with an additive angular margin loss function during training. Performance on the test dataset revealed enhanced performance. However, this strategy increased the computational time and power required. In [8], the attention branch network (ABN) is defined; an attention map from XAI is generated, which represents the regions of significance in an image; and the effective weight of the attention mechanism is extracted in a supervised learning approach. Results show that combining an ABN with base models improved performance and enhanced capabilities when incorporated with other attention-laced networks. This structure is somewhat complicated to add to an existing network and also adds more computational load.

1.1.3. Facial Emotion Recognition with Attention

In [9], a VGG16 combined with bidirectional LSTM and an attention mechanism was studied on single 2D images and multi-viewpoint images. The results showed increased performance against other approaches, reaching 87.62% and 80.73% accuracy on datasets for single and multi-viewpoint images, respectively. In [35], the authors propose slide-patch and whole-face models that use attention mechanisms with SE blocks. State-of-the-art performance was achieved on multiple datasets, and cross-database experiments showed the improved performance and generalization ability of the models.

Based on the improvements noted from the SE and CBAM block applications over base models, and given the simplicity in incorporating them into an existing network with limited hindrance on computational performance, they were selected for use in this study's FER modeling.

2. Materials and Methods

2.1. Base Network Architecture

The VGG16 was selected as the base network model, as a deep representation depth is considered important in classification results [2]. It is composed of five convolutional blocks and three fully connected (FC) layers. The first and second convolution blocks contain two convolution layers with rectified linear unit (ReLU) activation functions and a pooling layer. Convolution blocks three, four, and five have three convolution layers with ReLU activation and a pooling layer at the end. The architecture takes 224 × 224-pixel RGB images as input and is composed of 41 layers with a total of 138 million parameters. For this study, the model was trained from scratch without pre-trained weights.

To boost robustness and guide the models' decision making during training, attention modules were added throughout the architecture. In this study, the integration strategy of [5] was adopted for all the proposed attention modules of AGFER, SEFER, and CBAM-FER. They are placed after the second, third, and fourth convolutional blocks prior to the last pooling operation of each of the blocks. Following the work of [36], the last pooling layer ("pool5") was also removed from the architecture to increase the spatial dimension for better classification results and assessment of the decision visualization. This impact can be seen in the ablation study performed in Section 5.1, which assesses change in performance when specific parts of the CNN are removed. Figure 1 represents the network architecture and the implementation of the attention block strategy.

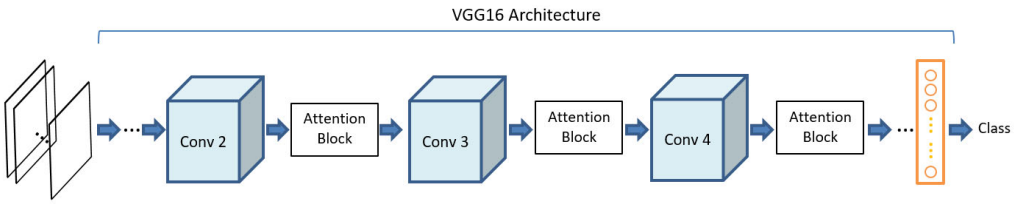

Figure 1. Schematic of the base network architecture and the integration of attention modules.

2.2. Attention Modules

Attention blocks have had a significant impact on conventional CNNs' performance. They have been adopted for different tasks, such as image classification [5–9] and image segmentation [37]. They are credited for their ability to drive the learning process of networks to focus on areas of particular importance to the task, thereby improving network efficiency and efficacy.

2.2.1. Squeeze and Excitation (SE)

Squeeze and excitation (SE) blocks are simple in structure and implementation. They enhance the representation power of the network by refining the channel-wise features without drastically affecting computational performance [5].

In SE blocks, the features generated by the convolution layers are first squeezed to find the channel statistics, which is achieved by implementing global average pooling (GAP) on the spatial features across the channel dimension. An advantage of GAP is that it enforces correlation between feature maps and the classes, making them class-agnostic confidence maps. Equally, since there are no optimization parameters at this layer, over-fitting can be avoided [38]. A gating mechanism is then added by means of two FC layers to bottleneck the data. To reduce the dimensionality of the information, the first FC layer was assigned to output a feature vector with a size of C/R, where R is a reduction coefficient and C is the total number of channels. The output then passes through an ReLU activation function to threshold each feature element. Afterward, it passes through the second FC layer, where the feature space is scaled back up to the input dimension, the excitation phase, and ends with a Sigmoid activation function before being multiplied channel-wise with the input feature space to form the newly refined features. Figure 2 shows a schematic drawing of the SE block architecture.

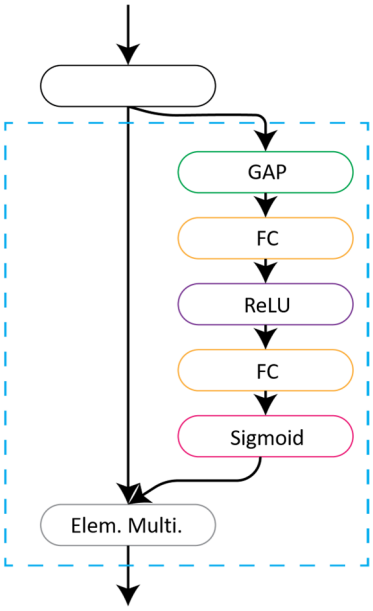

Figure 2. Schematic of the SE attention block (blue dashed lines).

2.2.2. Convolutional Block Attention Model

Convolution block attention models (CBAMs) are lightweight and can also be integrated easily into pre-existing network architectures with minimal impact on the number of parameters. They have a wide application range and have shown improvements in classification tasks [6]. The CBAM attention module extends the concept of SE blocks by considering the spatial information's importance in deciding "where" to focus, as well as the inclusion of a global max pooling (GMP) operator to gather features to infer finer channel attention [6].

A CBAM block is composed of two steps. The first step is where channel-wise attention is computed and processed. In the second step, the spatial attention is calculated.

In the original CBAM architecture channel-wise attention module, the features of the convolution layer are fed into both GAP and GMP in parallel. The output features are then fed into weight-shared multi-layer perceptron (MLP) layers with one hidden layer that reduces the dimension and a second which rescales it back to C. This process produces vectors which are then summed and passed through a Sigmoid activation function to form the channel-wise attention vector. The output of this process is then multiplied with the input feature space to form the channel-wise refined features.

Next, for spatial attention, the refined channel-wise features are fed into channel-wise global max pooling (CGMP) and channel-wise global average pooling (CGAP) in parallel. The outputs are then concatenated across the channel dimension and fed into a convolution layer with a filter size of seven before finally passing through a Sigmoid activation function to form the refined spatial features. The refined CBAM features are then the result of element-wise multiplication between the spatial attention features and the refined channel-wise features.

In this study's approach, some modification to the main CBAM architecture was performed. The weights of the MLP were not shared between the FC layers but rather followed the same approach of the SE block, i.e., a sequential approach. This approach was taken for the FC layers so that they could learn independently from each other for a better representation. Figure 3 represents a schematic drawing of the CBAM block used in this study.

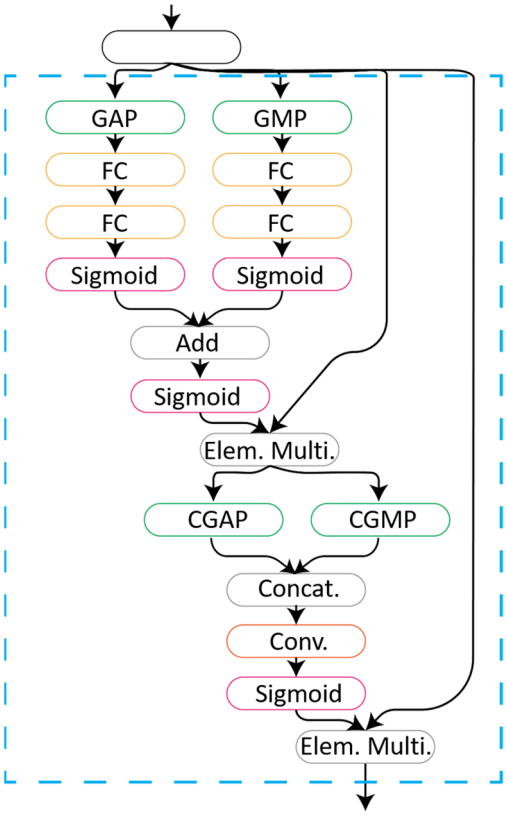

Figure 3. Schematic of the CBAM attention block (inside the blue dashed lines).

2.2.3. Attention-Guided Facial Emotion Recognition

A custom attention module is introduced in this article: the attention-guided facial emotion recognition module (AGFER). The AGFER combines the strengths of both SE and CBAMs by highlighting the feature maps of influence, as well as the focus area within each map. The input features pass through two pipelines in parallel, where the first extracts the

maps of influence by passing through GAP followed by two FC layers, one with an ReLU activation having a size of C/R, where R is a reduction coefficient and C is the total number of channels, and one with a Sigmoid activation function and size of C. The second pipeline determines the position of high concentration within the feature space, which is obtained by applying CGAP followed by two convolution layers. The first layer involves ReLU activation and has a filter size of 3, with the number of channels equal to $m + 1$, where m is the total number of classes. The second convolution uses a Sigmoid activation function with a filter size of 3 and a number of channels equal to 1.

The output from both pipelines is then merged using element-wise multiplication followed by a Sigmoid activation and is subsequently fused with the input feature space. Figure 4 represents a schematic of the AGFER attention block.

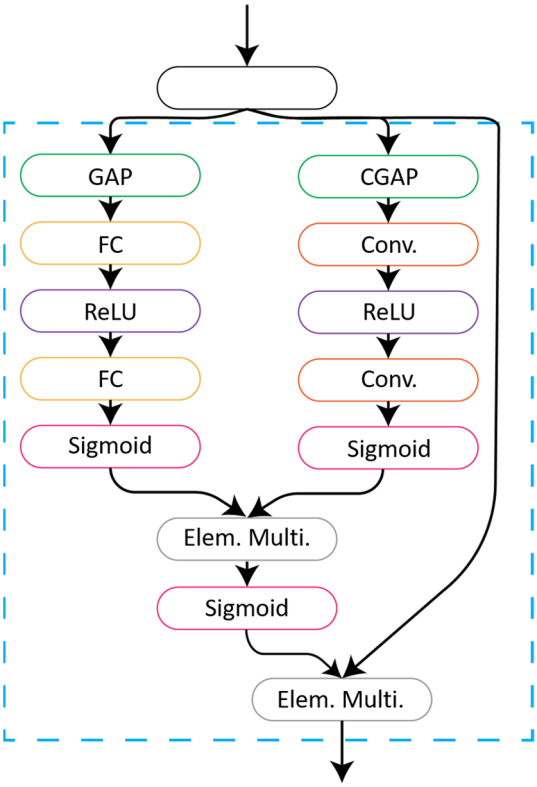

Figure 4. Schematic of the AGFER attention block (inside the blue dashed lines).

2.3. Guidance Strategy

To boost the model's classification performance and its focus on regions of impact for emotion recognition, a guidance strategy was adopted. First, a binary mask representing facial areas of special interest was created. This mask was used only during the training stage, where the input was fused with the mask and specific weighting for the model to focus on these areas and thus reduce the influence of the rest. Equation (1) represents the mask–image fusion and the output I_{Re} from the model input layer. I_{In} is the image input, M is the binary mask, and α is the weighted parameter, set to $\alpha = 0.125$ in this study. Figure 5 provides a graphical illustration of the mask–image fusion.

$$I_{Re} = I_{In} * (M + \alpha) \tag{1}$$

In neural networks, when deployed for facial emotion classification, the attention mechanism serves as a critical component to prioritize certain regions of the input image, making sure the network focuses on emotionally salient features, such as the eyebrows, eyes, and mouth. These regions often contain pivotal cues for emotion recognition. The guidance strategy, in this context, directs the attention mechanism towards these emotion-

ally significant regions. Without this guidance, the attention mechanism might give undue importance to less relevant facial areas, potentially missing or diluting the emotionally charged features that are essential for accurate classification.

Figure 5. Graphical illustration of the mask–image fusion output of the model input layer. Original pre-processed image input (**left**), pre-defined binary mask weighted with α (**middle**), and output of model input layer (**right**).

By introducing a guidance strategy, we provide the network with prior knowledge or a hint about where the most informative regions are likely to be located. This proactive approach ensures that the attention mechanism does not waste computational resources and focus on irrelevant areas. As a result, the model is better equipped to recognize subtle emotional cues, leading to improved accuracy in emotion classification. In the realm of facial emotion classification, where subtle changes in facial expressions can differentiate one emotion from another, ensuring that the network consistently focuses on the right areas can make a significant difference in performance.

Furthermore, a class-weighted cross-entropy loss function was used to compensate for any imbalance in class distribution. Weights were calculated as follows:

$$w_c = \frac{N}{m * s_c} \tag{2}$$

where w_c represents the weights of a certain class c, N is the total number of observations, m stands for the total number of classes, and s_c is the number of observations for a specific class c.

2.4. Database Descriptions

The well-accepted OULU-CASIA [23] database was chosen for the initial training and validation of the network models. The database was generated using two image acquisition systems with infrared and visible light, each taken under three illumination settings (strong, weak, and natural light). The database is composed of image sequences of 80 different subjects expressing the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise). A fraction of the entire database was chosen for the analysis, specifically the non-cropped RGB images of visible light with a strong illumination setting. This dataset was composed of 10,379 image frames with a quality of 320 $\times$ 240 pixels.

A second database was selected to test the generalization and robustness of the trained models, providing an entirely independent dataset for this purpose. For this task, the database of FACES [24] was chosen, which is made of images of facial portraits from varying subject ages. This database is composed of two images per emotion expression, expressing the six emotions of anger, disgust, fear, happiness, neutral, and sadness. It contains a total of 2052 images with a quality of 2835 $\times$ 3543 pixels. Thus, per our definition of robustness and generalization, this database is significantly different in terms of external influences.

Since the FACES database contains an extra emotion class (neutral) not present in the OULU-CASIA dataset, the first three frames of a subject's emotion image sequence were taken as the neutral set, and the rest were taken to represent the specific labelled emotions.

2.5. Image Pre-Processing

To highlight the face of the person in the image and reduce background noise, significant image pre-processing was performed. This step was crucial in model design as the

goal of learning models is to learn from features relevant to the task, emotion recognition in this case. The images from the OULU-CASIA dataset included a large background covering up to 50% of the total frame. In [12], when the image was taken as in the dataset with no pre-processing, the network focused on areas and features with no significance to emotion classification. This study also showed that the more the image was cropped to the exact facial characteristics without background noise, the better the focus areas of the network were in making correct decisions.

Based on the results of [9], a finer image pre-processing model was adopted for this analysis. Utilizing the object detection algorithm of [39] combined with the approach of [40], the face of the subjects in the images were first segmented. After segmentation, the algorithm was developed to extract the left eye and mouth regions from the face images using the approach of [41], with the bounding box locations of the detected regions the utilized as guidelines for further cropping. The dimensions of the left eye boundaries were optimized so that the eyebrow is included and then extended to encompass the right eye and eyebrow region as well.

The bottom boundary of the mouth region was set as the cropping edge for the lower limit of the image, while the upper boundary of the eye pair region was set as the upper limit. The right and left edges were set as the top right endpoint location of the left eye bounding area and the range of the eye pair region. The image processing algorithm developed effectively removed the background noise, focusing solely on the face of the subject. The model was developed so that if any of the regions failed detection then the image was excluded and removed from further analysis. Figure 6 describes the outcome from the main stages of this image pre-processing algorithm.

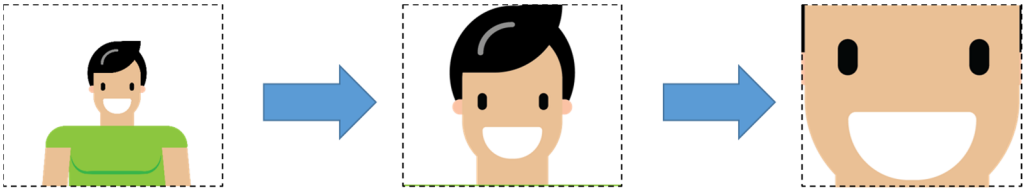

Figure 6. Image pre-processing steps from input to output. Image captured from a camera (**left**), first face detection (**middle**), output of the image pre-processing algorithm (**right**).

2.6. Performance Criteria

To effectively evaluate the performance of the models and the concentrations the attention modules bring, a criterion for evaluation was first defined. The data for analysis were based on the K-fold validation scheme, where images of the dataset were first partitioned into five segments, with each choosing a random training and validation set of images that was different from the next. This approach guarantees that all images are utilized for both training and validation. The performance criterion for the respective model is determined by taking the average of the true-positive (TP) accuracies from the predicted validation sets. Training accuracy was used to assess if the model was over-fit to its dataset of images by comparing it with the test or validation set accuracy. A limit of 5% difference between the validation and training accuracy was set, where any value less than this limit was considered to be a model that was not over-fit to its data.

2.7. Training Options

All models were executed in a MATLAB 2022a environment on a desktop with an AMD Ryzen Threadripper PRO 3955WX 16-Core @3.90 GHz, 512.00 GB memory (RAM), and the 64-bit Windows 10 operating system with an NVIDIA RTX A6000 graphics card. A constant learning rate of 0.0001 was set and run for 100 epochs. The stochastic gradient descent with momentum (SGDM) optimization function was used with a batch size of 128.

3. Results

3.1. Dataset Distribution

Table 1 summarizes the class distribution of both datasets. After passing the data through the image pre-processing stage, a loss of 10.41% and 8.77% of images was detected for the OULU-CASIA and FACES datasets, respectively, due to the inability of the algorithm to locate the defined regions of interest. The emotion classes are nearly equally distributed with a mean of 14.29% ± 1.45 and 16.67% ± 1.04 for OULU-CASIA and FACES datasets, indicating there is no particular bias to a certain emotion class. The lack of a surprise class in the FACES dataset will pose a challenge and demonstrate the capability of the model in prediction generalization.

Table 1. Class distribution of both OULU-CASIA and FACES datasets before and after Image Pre-Processing.

Emotion	OULU-CASIA		FACES	
	Original	After Image Pre-Processing	Original	After Image Pre-Processing
Anger	1790	1315	342	292
Disgust	1633	1195	342	292
Fear	1796	1503	342	303
Happiness	1668	1502	342	338
Neutral	N/A	1219	342	330
Sadness	1668	1200	342	317
Surprise	1701	1365	N/A	N/A
Total	10,379	9299	2052	1872

3.2. Model Performance

Figure 7 shows that the attention-integrated models significantly outperformed the base model. The neutral class showed the weakest performance in the attention-guided networks with a mean of 58.32%, 57.95%, and 58.53% for each of the AGFER, SEFER, and CBAMFER models, respectively, while the remaining emotion classes had greater than 90% classification accuracy. The proposed AGFER model had a mean accuracy of 89.19% ± 0.75, matching the performance of the SEFER and CBAMFER models with 88.73% ± 0.91 and 89.77% ± 0.70, respectively. The base model's low performance was recorded with a mean of 19.76% ± 2.49 over all emotion classes and the five segments.

To assess the ability to generalize, the trained models were tested against the independent FACES dataset. Table 2 shows the results separated according to the different age groups present. The model integrated with the proposed AGFER attention block performed the best out of the different trained models, achieving a mean classification of 46.50% ± 1.39. The AGFER model outperformed the others with increases of 27.97%, 6.02%, and 1.19% compared to the base, SEFER, and CBAMFER models, respectively. The young age group had the highest mean classification accuracy across each of the models, except for the base case. It is important to note that the models were entirely naïve to the FACES dataset.

In Figures 8 and 9, the confusion matrices for the AGFER (a) and CBAMFER (b) models are depicted for the OULU-CASIA validation and FACES dataset over all five segments combined (summed). The matrix reveals the struggle of the models to classify the neutral class, where a strong misclassification of the fear class in the OULU-CASIA validation set was observed. For the confusion matrix of the FACES datasets, the models were misclassifying fear as surprise.

Table 2. Mean classification accuracy (%) with the FACES dataset for all emotions across each segment for each model. Values in bold represent the best performance of the stride dimension.

Age Group	AGFER	Base	SEFER	CBAMFER
Young	**49.60 ± 0.95**	17.35 ± 0.45	43.95 ± 8.89	48.81 ± 3.36
Middle-aged	**48.50 ± 1.30**	18.07 ± 0.43	41.80 ± 7.57	46.57 ± 4.71
Old	**41.40 ± 1.92**	20.17 ± 0.50	35.70 ± 4.66	40.56 ± 6.92
Mean	**46.50 ± 1.39**	18.53 ± 0.46	40.48 ± 7.04	45.31 ± 5.00

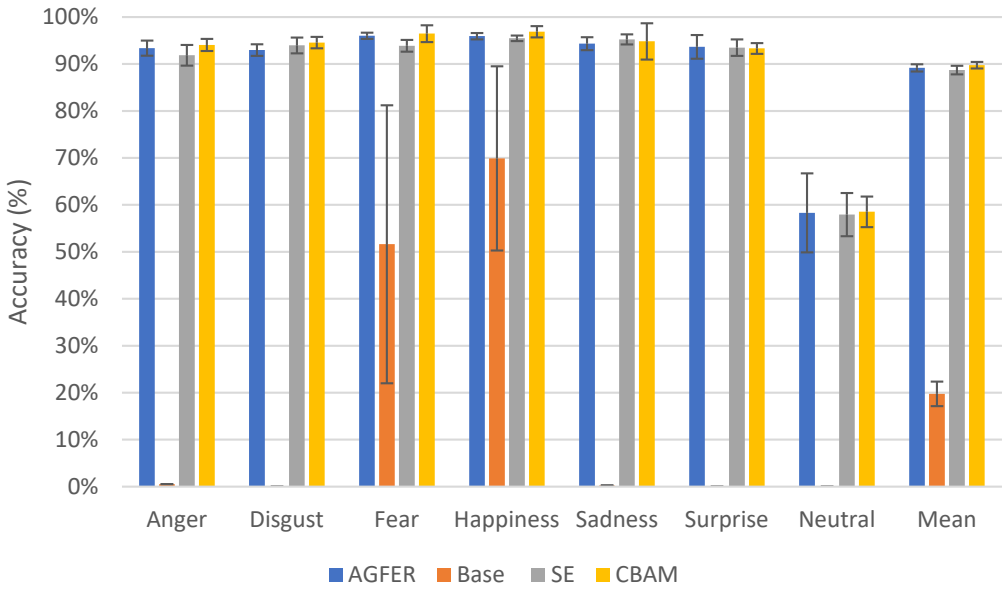

Figure 7. Mean classification accuracy for each emotion and the mean of all emotion classes of the OULU-CASIA validation dataset over the five segments for each model. The error bars represent the standard deviation over the five segments.

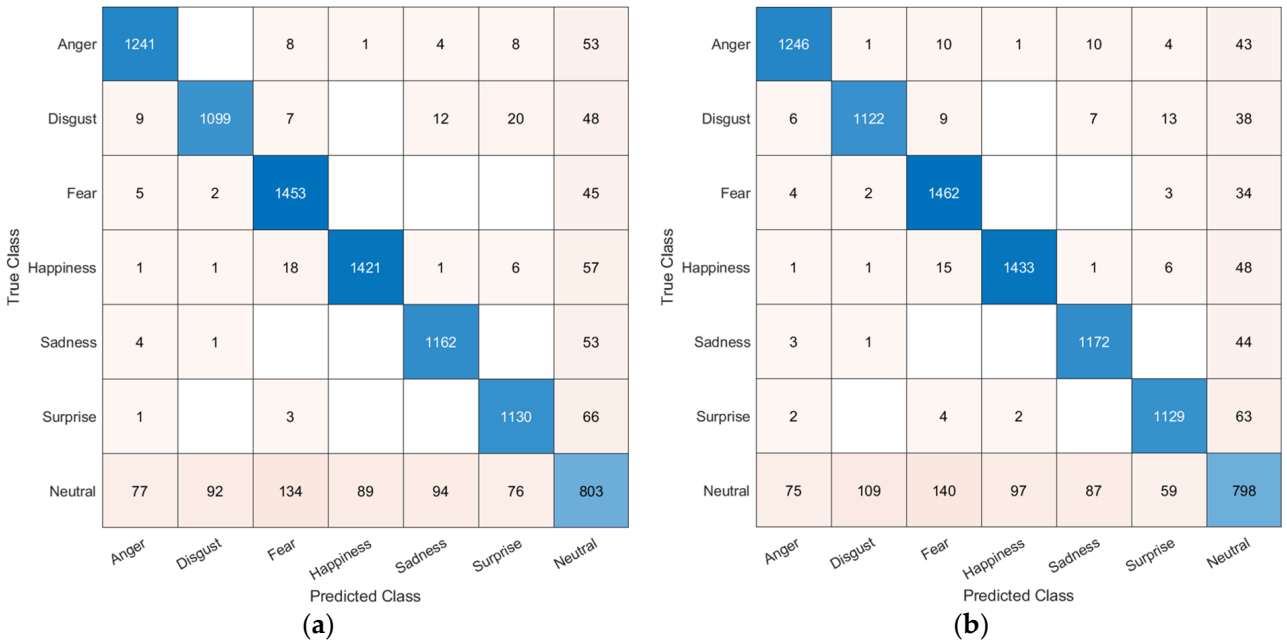

Figure 8. Confusion matrix results from the five segments for the OULU-CASIA validation set on the (**a**) AGFER model and (**b**) CBAMFER model. Blank boxes are equivalent to 0 or no predictions.

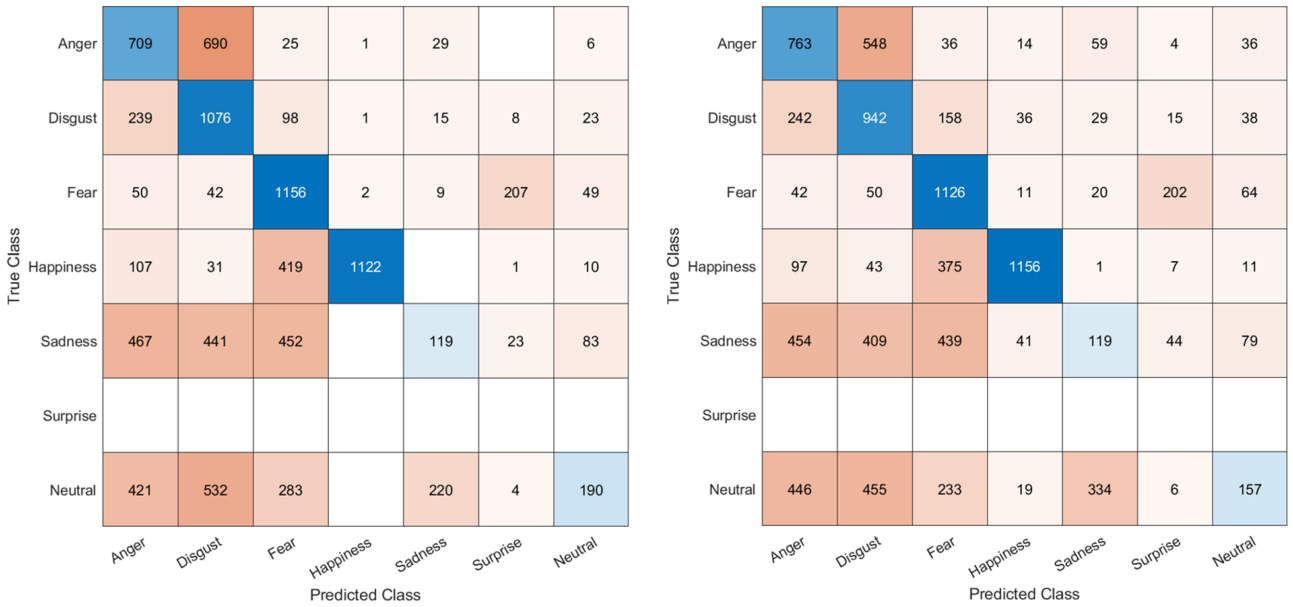

Figure 9. Confusion matrix results from the five segments for the FACES database on the (**a**) AGFER model and (**b**) CBAMFER model. Blank boxes are equivalent to 0 or no predictions.

4. Discussion

The results from Figure 7 reveal that the attention module-integrated network models outperformed the base architecture in all emotional classes given the training settings defined. The models did not achieve comparable performance for the neutral class compared to the other emotional classes. This difference is linked to feature representation, making it difficult for the model to distinguish neutral from the other classes, due in part to limited inter-class variability. This issue can be seen in the confusion matrix results of Figure 8, where the neutral class was misclassified the most.

The base model's performance showed the inability of the model to learn the required features with the given constraints, as evident from its performance for non-neutral emotion classes. An increase in the number of epochs (to reach 450) for the base model did yield similar performance to the other models trained at 100 epochs; however, this level of training uses much more resources and the training time required was three times longer.

The AGFER model was able to achieve comparable standards with the other attention models of SEFER and CBAMFER, highlighting its ability to refine features for a conclusive output. The CBAM model showed a small performance advantage of 0.57% $\pm$ 0.05 over the AGFER model due to the design structure of the CBAM module in learning and embedding more information. This slight improvement comes at the cost of more resources, with an increase of 0.1% in the number of learning parameters for the CBAMFER model compared to the AGFER model. The inference rate was also higher for the CBAMFER model, with a rate of 25 Hz compared to that of the AGFER (30 Hz). This increase is attributed to the depth of the architecture, where the AGFER model has a total of 80 layers while the CBAMFER has 92 layers. In terms of training time, an average of 2 h 35 min was observed for AGFER, while an average of 3 h 1 min was observed for CBAMFER. The error bars also indicate that the attention-based models all worked within a short range across all five segments, which emphasizes the efficiency and stability of the model guidance process.

While the choice of the OULU-CASIA database was informed by its credibility and prevalent use in related studies, a cross-database analysis was conducted to bolster model generalizability and counteract the pitfalls of reliance on a single dataset, reflecting a broader spectrum of facial expressions and conditions and thus underscoring the model's adaptability and performance in diverse real-world scenarios.

Table 2 reveals a weakness in the model when it comes to classifying emotions from other sources with different participant demographics and different external influences. The models were not able to maintain the same high performance rate as they had for the training and validation set from the OULU-CASIA dataset. This anomaly is linked to key factors including the quality of the input image, the age variation, and the different emotional expressiveness between participants in the OULU-CASIA and FACES datasets.

In particular, image quality has a strong impact on prediction as the OULU-CASIA dataset was captured at a low resolution, while the FACES dataset contains high-resolution images. The age of the subjects plays a small role, which can be noticed in Table 2, where the young age group images performed better than the others. This outcome is likely due to the OULU-CASIA dataset consisting mostly of participants from a young age demographic. The facial expressions for the defined emotion classes are also different between the datasets, creating a challenge for any classification model. This issue is evident in the results of the confusion matrices from Figure 9, where the AGFER and CBAMFER models showed difficulty in identifying and distinguishing neutral and sadness images from anger, disgust, and fear in the FACES dataset. This latter issue highlights the limitation associated with the large array of human facial expressions compared to the relatively limited images in these datasets, as well as the potential need for much larger labelled datasets.

A salient difference between the two datasets lies in how they represent emotional intensity. The FACES dataset is characterized by its focus on overt and distinct emotional expressions, capturing what can be described as well-defined emotional states. In contrast, the second dataset presents a more intricate range, spanning from the subtlest of emotional undertones to pronounced expressions. This depth mirrors the diverse ways emotions manifest in real-life settings, where their intensity can vary widely.

Moreover, the cultural representation in each dataset is also markedly different. FACES is primarily anchored in Western contexts, echoing emotional norms and expression styles commonly associated with European societies. Conversely, the OULU-CASIA dataset is more globally inclusive, incorporating data from varied regions such as Asia and Europe, thereby providing a wider lens into cultural emotional expressions.

This diverse blend of cultural variations, demographics, emotional intensities, and image quality contributes to the observed effects on model performance during cross-database evaluation.

Comparing the AGFER-integrated model with other research approaches, it can be seen that the proposed attention–guidance pipeline yielded state-of-the-art results with the OULU-CASIA dataset. Table 3 represents the evaluation results of different research approaches on the OULU-CASIA dataset.

Table 3. Comparison of mean classification accuracy results from different methods with the OULU-CASIA dataset. Values in bold represent the best performance.

Approach	Mean Accuracy (%)
Jung H. et al. [10]	81.46
Haddad J. et al. [42]	84.17
Zhao X. et al. [31]	84.59
Yu Z. et al. [43]	84.72
Yu Z. et al. [44]	86.23
Ding H. et al. [45]	87.71
Yu Z. et al. [43] (with LSTM)	88.98
Proposed AGFER approach	**89.19**

The increase in dimensional space coupled with the proposed attention module and guidance strategy demonstrated effectiveness in improving classification performance for FER. The attention module highlighted the regions of relevance and extracted the descriptive features to find a comprehensive pattern, capturing facial expressions effectively.

The weighted loss balanced the class presence distribution during training so that each emotion class was represented equally.

A crucial aspect warranting discussion is the relationship between the proposed AGFER attention module and the self-attention mechanism foundational to the Transformer architecture. The Transformer's self-attention mechanism has garnered acclaim for its ability to weigh the significance of different input elements in relation to a given element, allowing for dynamic recalibration based on context. Our attention module highlights the feature maps of influence, as well as the focus area within each map, and is coupled with the adopted guidance strategy, thus directing the attention mechanism towards regions of emotional significance. By understanding the underpinnings of both mechanisms, it becomes evident that, while they share foundational similarities in attending to different components of input, the differences in our module are geared towards providing the network with prior knowledge about where the most informative regions are likely to be located. This proactive approach ensures the model's ability to recognize subtle emotional cues, leading to improved accuracy in emotion classification. Drawing these parallels and distinctions helps contextualize our module's design choices and its potential advantages in the broader landscape of attention mechanisms.

Adversarial robustness, particularly in the face of imperceptible perturbations, is in-deed an indispensable aspect of model evaluation. Perturbations, even those that might seem negligible or imperceptible to the human eye, can significantly skew model predictions, revealing vulnerabilities. When the models were subjected to such perturbed images, based on the work of [46], a consistent pattern of performance degradation was observed with an average decrease of 24% for each model. This observation accentuates the challenges these perturbations present, underscoring the need for continuous refinements. Drawing insights from [46], it becomes evident that achieving a balance between generalizability and adversarial resilience remains an intricate step in model optimization. While the current framework of AGFER showcases promising results with standard datasets, the adversarial landscapes underscore areas for further exploration and enhancement.

As demonstrated in this study, attention-integrated models outperformed the base model given the defined constraints, parameters, and datasets. As in any research, some study limitations were highlighted. Limitations include no hyper-parameter tuning for the reduction parameter R of the attention blocks, neglecting the time-domain factor from the training data sequences, and not performing model explicability analysis. Future work will focus on tackling some of these limitations by tuning the reduction parameter via an optimization function. The time domain will be taken into consideration through the use of long short-term memory (LSTM) networks or Transformer models. To evaluate the explainability of the model, a class-dependent evaluation metric will be established for a quantitative measurement. Other approaches, such as multi-input feature fusion, are also considered along with larger dataset acquisition for more comprehensive model development.

In the scope of the comparisons, the proposed AGFER attention module was evaluated against two distinct attention modules and a foundational model architecture. However, the study did not encompass a broader assessment against leading state-of-the-art attention mechanisms in facial emotion recognition. This limitation confines the robustness of the results and might limit the full appreciation of the relative significance of the AGFER attention module.

5. Ablation Study

5.1. Model Stride Reduction

The network architecture's stride reduction was analyzed for performance assessment of the proposed models. Any reduction in the overall stride will increase the computational demand; therefore, to keep the training options fixed to the defined settings, a reduction of two was possible given the hardware restrictions. Thus, the models were trained with a reduction of two in the overall network stride to determine the effects of the increase in the

spatial dimension on the outcomes. The final pool layer prior to the FC layer was removed to increase the feature space from 7 × 7 × C to 14 × 14 × C, where C is the number of channels at the given level.

Table 4 highlights the results of the model stride reduction experiment. As can be noted, the lower the overall stride, the better the performance observed in each model. This improvement is related to the increase in the spatial dimension, which provides a more accurate interpretation of the learned features.

Table 4. Mean classification accuracy over all segments and emotion classes for each model at both overall network architecture stride dimensions with the OULU-CASIA dataset. Values in bold represent the best performance of the stride dimension.

Dataset	Overall Stride	AGFER	SEFER	CBAMFER
OULU-CASIA	32	86.81% ± 0.64	86.00% ± 0.74	86.84% ± 1.69
	16	**89.19% ± 0.75**	**88.73% ± 0.91**	**89.77% ± 0.70**

6. Conclusions

In this article, the advantages of using attention-integrated modules for facial emotion recognition are highlighted and a new attention module with a guided training pipeline is proposed. The proposed attention-guided facial emotion recognition (AGFER) module couples spatial and channel importance and works with the guidance strategy to emphasize key descriptors in the images for a more robust outcome. The attention-infused models were able to outperform the base model with a margin greater than 60% in terms of mean accuracy given the training settings applied. The AGFER model achieved comparable performance to the squeeze and excitation (SE) and convolutional block attention module (CBAM)-integrated models with the OULU-CASIA validation set while outperforming them with a mean margin of 3.60% with the FACES dataset. The proposed attention–guidance strategy showed the capabilities of attention-laced networks in improving performance by focusing on areas of relative importance towards emotion classification and reducing the time allotted for training, thereby saving resources.

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

Funding: This research is partially funded by the German Federal Ministry of Research and Education (BMBF) under grant CoHMed/PersonaMed-A 13FH5I06IA and LESSON FKZ: 3FH5E10IA, as well as grants from KOMPASS funded by the Ministerium für Wissenschaft, Forschung und Kunst (MWK) of Baden-Wuerttemberg Germany, ERAPERMED2022-276—ETAP BMG FKZ 2523FSB110, and the Deutscher Akademischer Austauschdienst (DAAD) under grant AIDE-ASD FKZ 57656657.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The databases used in this study were the (OULU-CASIA NIR and VIS facial expression database) and the (FACES database). The OULU-CASIA NIR and VIS facial expression database dataset (https://www.oulu.fi/en/university/faculties-and-units/faculty-information-technology-and-electrical-engineering/center-machine-vision-and-signal-analysis (accessed on 6 October 2020)) and the FACES database (https://faces.mpdl.mpg.de/imeji/ (accessed on 27 August 2021)) are available from the respected publishers upon request.

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

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
2. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
3. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
4. Samek, W.; Wiegand, T.; Müller, K.-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv* **2017**, arXiv:1708.08296.
5. 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, 18–22 June 2018; pp. 7132–7141.
6. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19, ISBN 978-3-030-01233-5.
7. Ling, H.; Wu, J.; Wu, L.; Huang, J.; Chen, J.; Li, P. Self Residual Attention Network for Deep Face Recognition. *IEEE Access* **2019**, *7*, 55159–55168. [CrossRef]
8. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10705–10714.
9. Sepas-Moghaddam, A.; Etemad, A.; Pereira, F.; Correia, P.L. Facial Emotion Recognition Using Light Field Images with Deep Attention-Based Bidirectional LSTM. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3367–3371.
10. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2983–2991.
11. Mehrabian, A. Communication without Words. In *Communication Theory*; Mortensen, C.D., Ed.; Routledge: London, UK, 2017; ISBN 978-1-351-52752-1.
12. Arabian, H.; Wagner-Hartl, V.; Chase, J.G.; Möller, K. Image Pre-Processing Significance on Regions of Impact in a Trained Network for Facial Emotion Recognition. *IFAC-Pap.* **2021**, *54*, 299–303. [CrossRef]
13. de Gelder, B. Why Bodies? Twelve Reasons for Including Bodily Expressions in Affective Neuroscience. *Philos. Trans. R. Soc. B Biol. Sci.* **2009**, *364*, 3475–3484. [CrossRef]
14. de Gelder, B.; Van den Stock, J.; Meeren, H.K.M.; Sinke, C.B.A.; Kret, M.E.; Tamietto, M. Standing up for the Body. Recent Progress in Uncovering the Networks Involved in the Perception of Bodies and Bodily Expressions. *Neurosci. Biobehav. Rev.* **2010**, *34*, 513–527. [CrossRef]
15. Lang, P.J.; Bradley, M.M. Emotion and the Motivational Brain. *Biol. Psychol.* **2010**, *84*, 437–450. [CrossRef]
16. Vuilleumier, P. How Brains Beware: Neural Mechanisms of Emotional Attention. *Trends Cogn. Sci.* **2005**, *9*, 585–594. [CrossRef]
17. Mancini, C.; Falciati, L.; Maioli, C.; Mirabella, G. Happy Facial Expressions Impair Inhibitory Control with Respect to Fearful Facial Expressions but Only When Task-Relevant. *Emotion* **2022**, *22*, 142–152. [CrossRef]
18. Mirabella, G.; Grassi, M.; Mezzarobba, S.; Bernardis, P. Angry and Happy Expressions Affect Forward Gait Initiation Only When Task Relevant. *Emotion* **2023**, *23*, 387–399. [CrossRef] [PubMed]
19. Mancini, C.; Falciati, L.; Maioli, C.; Mirabella, G. Threatening Facial Expressions Impact Goal-Directed Actions Only If Task-Relevant. *Brain Sci.* **2020**, *10*, 794. [CrossRef]
20. Leo, M.; Del Coco, M.; Carcagni, P.; Distante, C.; Bernava, M.; Pioggia, G.; Palestra, G. Automatic Emotion Recognition in Robot-Children Interaction for ASD Treatment. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, 7–13 December 2015; pp. 537–545.
21. Ravindran, V.; Osgood, M.; Sazawal, V.; Solorzano, R.; Turnacioglu, S. Virtual Reality Support for Joint Attention Using the Floreo Joint Attention Module: Usability and Feasibility Pilot Study. *JMIR Pediatr. Parent.* **2019**, *2*, e14429. [CrossRef]
22. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *arXiv* **2019**, arXiv:1903.12261.
23. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; Pietikäinen, M. Facial Expression Recognition from Near-Infrared Videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [CrossRef]
24. Ebner, N.C.; Riediger, M.; Lindenberger, U. FACES—A Database of Facial Expressions in Young, Middle-Aged, and Older Women and Men: Development and Validation. *Behav. Res. Methods* **2010**, *42*, 351–362. [CrossRef] [PubMed]
25. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2009; ISBN 978-0-262-01319-2.
26. Arabian, H.; Wagner-Hartl, V.; Geoffrey Chase, J.; Möller, K. Facial Emotion Recognition Focused on Descriptive Region Segmentation. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Virtual, 1–5 November 2021; pp. 3415–3418.
27. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
28. Khaireddin, Y.; Chen, Z. Facial Emotion Recognition: State of the Art Performance on FER2013. *arXiv* **2021**, arXiv:2105.03588.

29. Challenges in Representation Learning: Facial Expression Recognition Challenge. Available online: https://kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge (accessed on 9 August 2021).
30. Mehendale, N. Facial Emotion Recognition Using Convolutional Neural Networks (FERC). *SN Appl. Sci.* **2020**, *2*, 446. [CrossRef]
31. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-Piloted Deep Network for Facial Expression Recognition. *arXiv* **2017**, arXiv:1607.06997.
32. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842.
33. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *arXiv* **2015**, arXiv:1409.0575. [CrossRef]
34. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
35. Li, Y.; Lu, G.; Li, J.; Zhang, Z.; Zhang, D. Facial Expression Recognition in the Wild Using Multi-Level Features and Attention Mechanisms. *IEEE Trans. Affect. Comput.* **2020**, *14*, 451–462. [CrossRef]
36. Vardazaryan, A.; Mutter, D.; Marescaux, J.; Padoy, N. Weakly-Supervised Learning for Tool Localization in Laparoscopic Videos. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*; Stoyanov, D., Taylor, Z., Balocco, S., Sznitman, R., Martel, A., Maier-Hein, L., Duong, L., Zahnd, G., Demirci, S., Albarqouni, S., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 169–179.
37. Fernandez, P.D.M.; Pena, F.A.G.; Ren, T.I.; Cunha, A. FERAtt: Facial Expression Recognition with Attention Net. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; IEEE: Long Beach, CA, USA, 2019; pp. 837–846.
38. Lin, M.; Chen, Q.; Yan, S. Network In Network. *arXiv* **2014**, arXiv:1312.4400.
39. Viola, P.; Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I–I.
40. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
41. Castrillón, M.; Déniz, O.; Guerra, C.; Hernández, M. ENCARA2: Real-Time Detection of Multiple Faces at Different Resolutions in Video Streams. *J. Vis. Commun. Image Represent.* **2007**, *18*, 130–140. [CrossRef]
42. Haddad, J.; Lezoray, O.; Hamel, P. 3D-CNN for Facial Emotion Recognition in Videos. In *Advances in Visual Computing*; Bebis, G., Yin, Z., Kim, E., Bender, J., Subr, K., Kwon, B.C., Zhao, J., Kalkofen, D., Baciu, G., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12510, pp. 298–309, ISBN 978-3-030-64558-8.
43. Yu, Z.; Liu, G.; Liu, Q.; Deng, J. Spatio-Temporal Convolutional Features with Nested LSTM for Facial Expression Recognition. *Neurocomputing* **2018**, *317*, 50–57. [CrossRef]
44. Yu, Z.; Liu, Q.; Liu, G. Deeper Cascaded Peak-Piloted Network for Weak Expression Recognition. *Vis. Comput.* **2018**, *34*, 1691–1699. [CrossRef]
45. Ding, H.; Zhou, S.K.; Chellappa, R. Facenet2expnet: Regularizing a Deep Face Recognition Net for Expression Recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; IEEE: Washington, DC, USA, 2017; pp. 118–126.
46. Zhang, W.; Li, D.; Min, X.; Zhai, G.; Guo, G.; Yang, X.; Ma, K. Perceptual Attacks of No-Reference Image Quality Models with Human-in-the-Loop. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2916–2929.