sensors

MDPI

Article

Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network

Shervin Minaee [1,*], Mehdi Minaei [2] and Amirali Abdolrashidi [3]

[1] Snapchat Inc., Santa Monica, CA 90405, USA
[2] CS Department, Sama Technical College, Azad University, Tonekabon 46817, Iran; mehdiminaei1376@gmail.com
[3] CS Department, University of California, Riverside, CA 92521, USA; amirali.abdolrashidi@email.ucr.edu
* Correspondence: sminaee@snapchat.com

Abstract: Facial expression recognition has been an active area of research over the past few decades, and it is still challenging due to the high intra-class variation. Traditional approaches for this problem rely on hand-crafted features such as SIFT, HOG, and LBP, followed by a classifier trained on a database of images or videos. Most of these works perform reasonably well on datasets of images captured in a controlled condition but fail to perform as well on more challenging datasets with more image variation and partial faces. In recent years, several works proposed an end-to-end framework for facial expression recognition using deep learning models. Despite the better performance of these works, there are still much room for improvement. In this work, we propose a deep learning approach based on attentional convolutional network that is able to focus on important parts of the face and achieves significant improvement over previous models on multiple datasets, including FER-2013, CK+, FERG, and JAFFE. We also use a visualization technique that is able to find important facial regions to detect different emotions based on the classifier's output. Through experimental results, we show that different emotions are sensitive to different parts of the face.

Keywords: convolutional neural network; attention mechanism; spatial transformer network; facial expression recognition

Citation: Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors* **2021**, *21*, 3046. https://doi.org/10.3390/s21093046

Academic Editor: Mariano Alcañiz Raya

Received: 15 February 2021
Accepted: 23 April 2021
Published: 27 April 2021

1. Introduction

Emotions are an inevitable part of interpersonal communication. They can be expressed in many different forms, which may or may not be observed with the naked eye. Therefore, with the right tools, any indications preceding or following them can be subject to detection and recognition. There has been an increase in the need to detect a person's emotions in the past few years and increasing interest in human emotion recognition in various fields including, but not limited to, human–computer interfaces [1], animation [2], medicine [3,4], security [5,6], diagnostics for Autism Spectrum Disorders (ASD) in children [7], and urban sound perception [8].

Emotion recognition can be performed using different features, such as facial expressions [2,9,10], speech [5,11], EEG [12], and even text [13]. Among these features, facial expressions are one of the most popular, if not the most popular, due to a number of reasons; they are visible, they contain many useful features for emotion recognition, and it is easier to collect a large dataset of faces (than other means for human recognition) [2,14,15].

Recently, with the use of deep learning and especially convolutional neural networks (CNNs) [16], many features can be extracted and learned for a decent facial expression recognition system [17,18]. It is, however, noteworthy that, in the case of facial expressions, many of the clues come from a few areas of the face, e.g., the mouth and eyes, whereas other parts, such as the ears and hair, play little parts in the output [19]. This means that, ideally, the machine learning framework should focus only on important parts of the face and should be less sensitive to other facial regions.

In this work, we propose a deep learning-based framework for facial expression recognition, which takes the above observation into account and uses an attention mechanism to focus on the salient part of the face. We show that, by using attentional convolutional network, even a network with a few layers (less than 10 layers) is able to achieve a very high accuracy rate. More specifically, this paper presents the following contributions:

- We propose an approach based on an attentional convolutional network, which can focus on feature-rich areas of the face yet remarkably outperforms recent works in accuracy.
- In addition, we use the visualization technique proposed in [20] to highlight the facial image's most salient regions, i.e., the parts of the image that have the strongest impact on the classifier's outcome. Samples of salient regions for different emotions are shown in Figure 1.

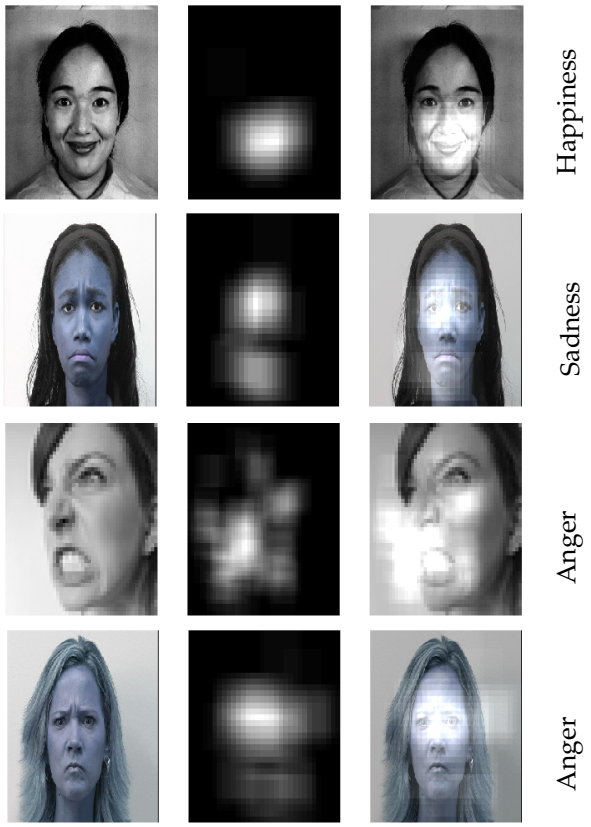

Figure 1. The detected salient regions for different facial expressions using our model. The images in the first and third rows are taken from the FER dataset, and the images in the second and fourth rows belong to the extended Cohn-Kanade dataset.

In the following sections, we first provide an overview of related works in Section 2. The proposed framework and model architecture are explained in Section 3. We then provide the experimental results, overview of databases used in this work, and model visualization in Section 4. Finally, we conclude the paper in Section 5.

2. Related Works

In one of the most iconic works in emotion recognition by Paul Ekman [21], happiness, sadness, anger, surprise, fear, and disgust were identified as the six principal emotions (besides neutral). Ekman later developed FACS [22] using this concept, thus setting the standard for work on emotion recognition ever since. Neutral was also included later on in most human recognition datasets, resulting in seven basic emotions. Image samples of these emotions from three datasets are displayed in Figure 2.

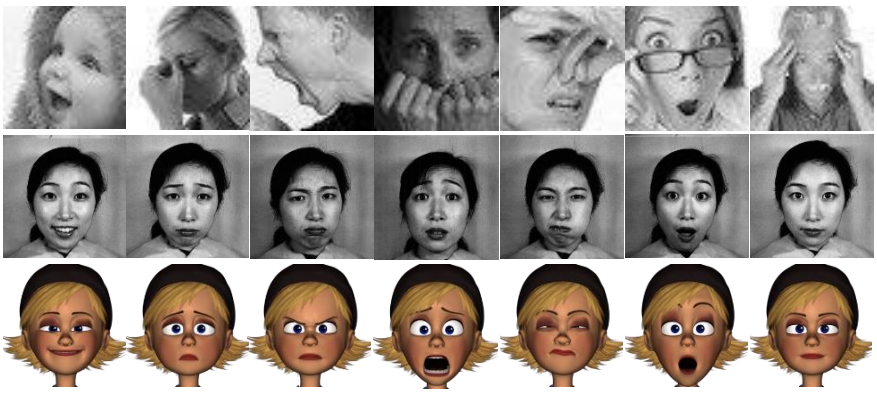

Figure 2. (Left to right) The six cardinal emotions (happiness, sadness, anger, fear, disgust, and surprise) and neutral. The images in the first, second, and the third rows belong to the FER, JAFFE, and FERG datasets, respectively.

Earlier works on emotion recognition relied on the traditional two-step machine learning approach, where in the first step, some features are extracted from the images and, in the second step, a classifier (such as SVM, neural network, or random forest) is used to detect the emotions. Some of the popular hand-crafted features used for facial expression recognition include the histogram of oriented gradients (HOG) [23,24], local binary patterns (LBP) [25], Gabor wavelets [26], and Haar features [27]. A classifier then assigns the best emotion to the image. These approaches seemed to work fine on simpler datasets, but with the advent of more challenging datasets (which have more intra-class variation), they started to show their limitations. To obtain a better sense of some of the possible challenges with the images, we refer the readers to the images in the first row of Figure 2, where the image shows only a partial face or the face is occluded with a hand or eyeglasses.

With the great success of deep learning and more specifically convolutional neural networks for image classification and other vision problems [28–35], several groups developed deep learning-based models for facial expression recognition (FER). To name some of the promising works, Khorrami in [17] showed that CNNs can achieve a high accuracy in emotion recognition and used a zero-bias CNN on the extended Cohn–Kanade dataset (CK+) and the Toronto Face Dataset (TFD) to achieve state-of-the-art results. Aneja et al. [2] developed a model of facial expressions for stylized animated characters based on deep learning by training a network to model the expression of human faces, one for that of animated faces, and one to map human images into animated ones. Mollahosseini [9] proposed a neural network for FER using two convolution layers, one max pooling layer, and four "inception" layers, i.e., sub-networks. Liu in [10] combined feature extraction and classification in a single looped network, citing the two parts' needs for feedback from each other. They used their Boosted Deep Belief Network (BDBN) on CK+ and JAFFE, achieving state-of-the-art accuracy.

Barsoum et al. [36] used a deep CNN on noisy labels acquired via crowd-sourcing for ground truth images. They used 10 taggers to relabel each image in the dataset and used various cost functions for their DCNN, achieving decent accuracy. Han et al. [37] proposed an incremental boosting CNN (IB-CNN) in order to improve the recognition of spontaneous facial expressions by boosting discriminative neurons, which showed improvements over the best methods at the time. Meng in [38] proposed an identity-aware CNN (IA-CNN) that used identity- and expression-sensitive contrastive losses to reduce the variations in learning identity- and expression-related information. In [39], Fernandez et al. proposed an end-to-end network architecture for facial expression recognition with an attention model.

In [40], Want et al. proposed a simple yet efficient Self-Cure Network (SCN) that suppresses uncertainties efficiently and prevents deep networks from overfitting uncertain facial images (due to noisy labels). Specifically, SCN suppresses the uncertainty from two

different aspects: (1) a self-attention mechanism over a mini-batch to weight each training sample with a ranking regularization and (2) a careful relabeling mechanism to modify the labels of these samples in the lowest-ranked group. In [41], Wang et al. developed a facial expression recognition algorithm that is robust to real-world pose and occlusion variations. They proposed a novel Region Attention Network (RAN) to adaptively capture the importance of facial regions for occlusion and pose variant FER. Some of the other recent works on facial expression recognition includes Multiple attention network for facial expression recognition [42], deep self-attention network for facial emotion recognition [43], and a recent survey on facial expression recognition [44].

All of the above works achieved significant improvements over traditional works on emotion recognition, but they are missing a simple method for recognizing important facial regions for emotion detection. In this work, we try to address this problem by proposing a framework based on an attentional convolutional network that is able to focus on salient facial regions.

3. The Proposed Framework

We propose an end-to-end deep learning framework based on an attentional convolutional network to classify the underlying emotion in facial images. Often times, improving a deep neural network relies on adding more layers/neurons, facilitating gradient flow in the network (e.g., by adding skip layers), or better regularizations (e.g., spectral normalization), especially for classification problems with a large number of classes. However, for facial expression recognition, due to the small number of classes, we show that using a convolutional network with less than 10 layers and attention (which is trained from scratch) is able to achieve promising results, presenting better results than state-of-the-art models for several databases.

Given a facial image, it is clear that not all parts of the face are important for detecting a specific emotion, and in many cases, we only need to pay attention to specific regions to get a sense of the underlying emotion. Based on this observation, we added an attention mechanism, through spatial transformer network into our framework to focus on important facial regions.

Figure 3 illustrates the proposed model architecture. The feature extraction part consists of four convolutional layers, with every two followed by a max-pooling layer and a rectified linear unit (ReLU) activation function. They were then followed by a dropout layer and two fully connected layers. The spatial transformer (the localization network) consisted of two convolution layers (each followed by max-pooling and ReLU) and two fully connected layers. After regressing the transformation parameters, the input was transformed to the sampling grid $T(\theta)$, producing the warped data. The spatial transformer module essentially tries to focus on the most relevant parts of the image by estimating a sample over the region of interest. One can use different transformations to warp the input to the output; here, we used an affine transformation, which is commonly used for many applications. For further details about the spatial transformer network, please refer to [45].

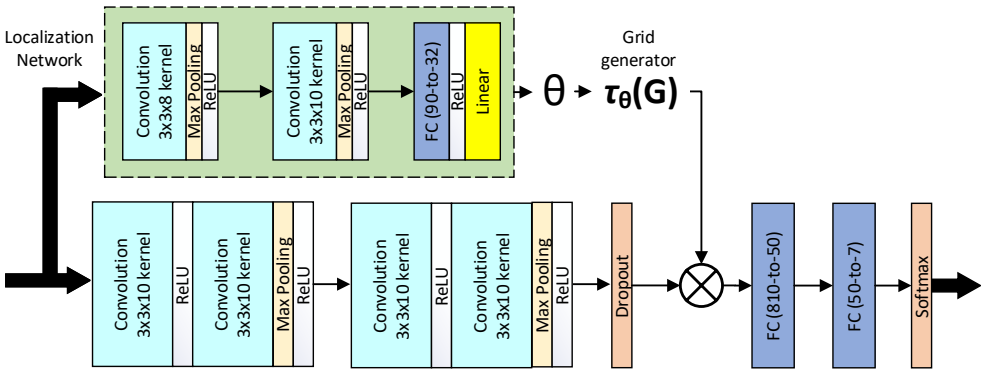

Figure 3. The proposed model architecture.

This model was then trained by optimizing a loss function using the stochastic gradient descent approach (more specifically, the Adam optimizer). The loss function in this work is simply the summation of two terms, the classification loss (cross-entropy), and the regularization term (which is the ℓ_2 norm of the weights in the last two fully-connected layers).

$$\mathcal{L}_{overall} = \mathcal{L}_{classifier} + \lambda \|w_{(fc)}\|_2^2 \quad (1)$$

The regularization weight, λ, is tuned based on the model performance on the validation set to pick the corresponding value that yields the highest performance on the validation set. Adding both dropout and ℓ_2 regularization enables us to train our models from scratch even on very small datasets, such as JAFFE and CK+. It is worth mentioning that we trained a separate model for each of the databases used in this work. We also tried using a network with similar architecture but more than 50 layers, but the accuracy did not improve significantly. Therefore, we decided to use the network with fewer layers, which has a much faster inference speed and is more suitable for real-time applications.

4. Experimental Results

In this section, we provide the detailed experimental analysis of our model on several facial expression recognition databases. We first provide a brief overview of the databases used in this work, then provide the performance of our models on four databases, and compare the results with some of the promising recent works. We then provide the salient regions detected by our trained model using a visualization technique.

4.1. Databases

In this work, we provide the experimental analysis of the proposed model on several popular facial expression recognition datasets, including FER2013 [14], the extended Cohn–Kanade [46], Japanese Female Facial Expression (JAFFE) [15], and Facial Expression Research Group Database (FERG) [2]. Before diving into the results, we give a brief overview of these databases.

FER2013: The Facial Expression Recognition 2013 (FER2013) database was first introduced in the ICML 2013 Challenges in Representation Learning [14]. This dataset contains 35,887 images of 48 × 48 resolution, most of which are taken in wild settings. Originally, the training set contained 28,709 images, and the validation and test sets include 3589 images each. This database was created using the Google image search API, and faces were automatically registered. The faces are labeled as any of the six cardinal expressions as well as neutral. Compared to the other datasets, FER has more variation in the images, including facial occlusion (mostly with a hand), partial faces, low-contrast images, and eyeglasses. Four sample images from the FER dataset are shown in Figure 4.

Figure 4. Four sample images from the FER database.

CK+: The extended Cohn–Kanade (known as CK+) facial expression database [46] is a public dataset for action unit and emotion recognition. It includes both posed and non-posed (spontaneous) expressions. The CK+ comprises a total of 593 sequences across 123 subjects. In most previous work, the last frame of these sequences is taken and used for image-based facial expression recognition. Six sample images from this dataset are shown in Figure 5.

Figure 5. Six sample images from the CK+ database.

JAFFE: This dataset contains 213 images of the 7 facial expressions posed by 10 Japanese female models. Each image has been rated on the six emotional adjectives by 60 Japanese subjects [15]. Four sample images from this dataset are shown in Figure 6.

Figure 6. Four sample images from the JAFFE database.

FERG: FERG is a database of stylized characters with annotated facial expressions. The database contains 55,767 annotated face images of six stylized characters. The characters were modeled using MAYA. The images for each character are grouped into seven types of expressions [2]. Six sample images from this database are shown in Figure 7. We mainly wanted to try our algorithm on this database to see how it performs on cartoonish characters.

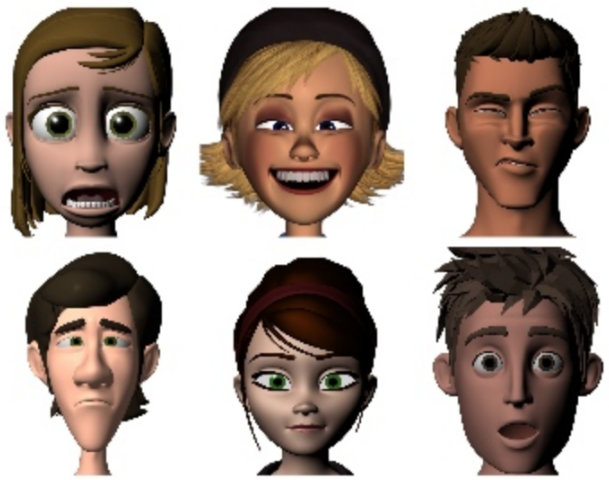

Figure 7. Six sample images from the FERG database.

4.2. Experimental Analysis and Comparison

We now present the performance of the proposed model on the above datasets. In each case, we trained the model on a subset of that dataset, validated it on a validation set, and reported the accuracy over the test set.

Before getting into the details of the model's performance on different datasets, we briefly discuss our training procedure. We trained one model per dataset in our experiments, but we tried to keep the architecture and hyperparameters similar among these different models. Each model was trained for 300 epochs from scratch, on an AWS EC2 instance with a Nvidia Tesla K80 GPU. We initialized the network weights with random Gaussian variables with zero mean and 0.05 standard deviation. For optimization, we used an Adam optimizer with a learning rate of 0.005 (different optimizers were used, including stochastic gradient descents, and Adam seemed perform slightly better). We also added L2 regularization with a weight decay value of 0.001. It took around 2–4 h to train our models on the FER and FERG datasets. For JAFFE and CK+, since there are much fewer images, it took less than 10 min to train a model. For datasets with large class imbalance, we used oversampling on the classes with fewer samples to force the different classes to be of the same order. Data augmentation was used for the images in the training sets to train the model on a larger number of images and to train model for invariances on small transformations. We used flipping, small rotation, and small distortion to augment the data.

As discussed before, the FER-2013 dataset is more challenging than the other facial expression recognition datasets we used. Besides the intra-class variation of FER, another main challenge in this dataset is the imbalanced nature of different emotional classes. Some of the classes such as happiness and neutral have many more examples than others. We used all 28,709 images in the training set to train the model, validated on 3500 validation images, and reported the model accuracy on the 3589 images in the test set. We were able to achieve an accuracy rate of around 70.02% on the test set.

The comparison of the result of our model with some of the previous works on FER 2013 is provided in Table 1.

Table 1. Classification accuracies on the FER 2013 dataset.

Method	Accuracy Rate
Unsupervised Domain Adaptation [47]	65.3%
Bag of Words [48]	67.4%
VGG+SVM [49]	66.31%
GoogleNet [50]	65.2%
FER on SoC [51]	66%
Mollahosseini et al. [9]	66.4%
The proposed algorithm	70.02%
Aff-Wild2 (VGG backbone) [52]	75%

For the FERG dataset, we used around 34,000 images for training, 14,000 for validation, and 7000 for testing. For each facial expression, we randomly select 1000 images for testing. We were able to achieve an accuracy rate of around 99.3%. The comparison between the proposed algorithm and some of the previous works on FERG dataset are provided in Table 2.

Table 2. Classification accuracy on the FERG dataset.

Method	Accuracy Rate
DeepExpr [2]	89.02%
Ensemble Multi-feature [53]	97%
Adversarial NN [54]	98.2%
The proposed algorithm	99.3%

For the JAFFE dataset, we used 120 images for training, 23 images for validation, and 70 images for test (10 images per emotion in the test set). The overall accuracy on this dataset is around 92.8%. The comparison with previous works on the JAFFE dataset is shown in Table 3.

Table 3. Classification accuracy on the JAFFE dataset.

Method	Accuracy Rate
LBP+ORB features [55]	88.5%
Fisherface [56]	89.2%
Deep Features + HOG [57]	90.58%
Salient Facial Patch [58]	91.8%
CNN+SVM [59]	95.31%
The proposed algorithm	92.8%

For CK+, 70% of the images were used as training, 10% was used for validation, and 20% was used for testing (which corresponds to 420, 60, and 113 images for the training, validation, and test sets, respectively). The comparison of our model with previous works on the extended CK dataset is shown in Table 4.

Table 4. Classification accuracy on CK+.

Method	Accuracy Rate
MSR [60]	91.4%
3DCNN-DAP [61]	92.4%
LBP+ORB features [55]	93.2%
Inception [9]	93.2%
Deep Features + HOG [57]	94.17%
IB-CNN [37]	95.1%
IACNN [38]	95.37%
DTAGN [62]	97.2%
ST-RNN [63]	97.2%
PPDN [64]	97.3%
Dynamic cascaded classifier [65]	97.8%
The proposed algorithm	98.0%

4.3. Confusion Matrix

The confusion matrix of the proposed model on the test set of FER dataset is shown in Figure 8. As we can see, the model makes more mistakes for classes with less samples such as disgust and fear.

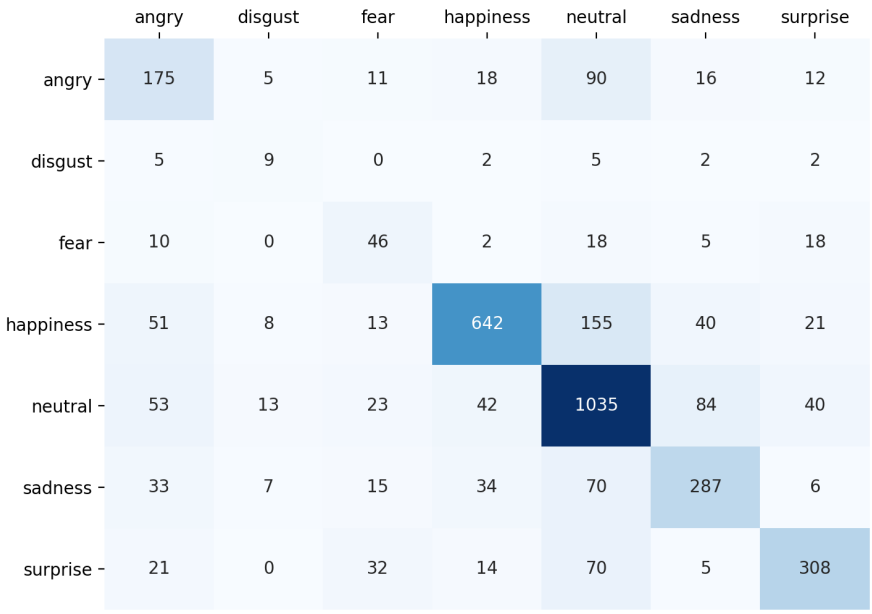

Figure 8. The confusion matrix of the proposed model on the test set of the FER dataset. The number of images for each emotion class in the test set is as follows: angry: 328, disgust: 25, fear: 99, happiness: 930, neutral: 1290, sadness: 450, and surprise: 450.

The confusion matrix of the proposed model on the test set of the FERG dataset is shown in Figure 9.

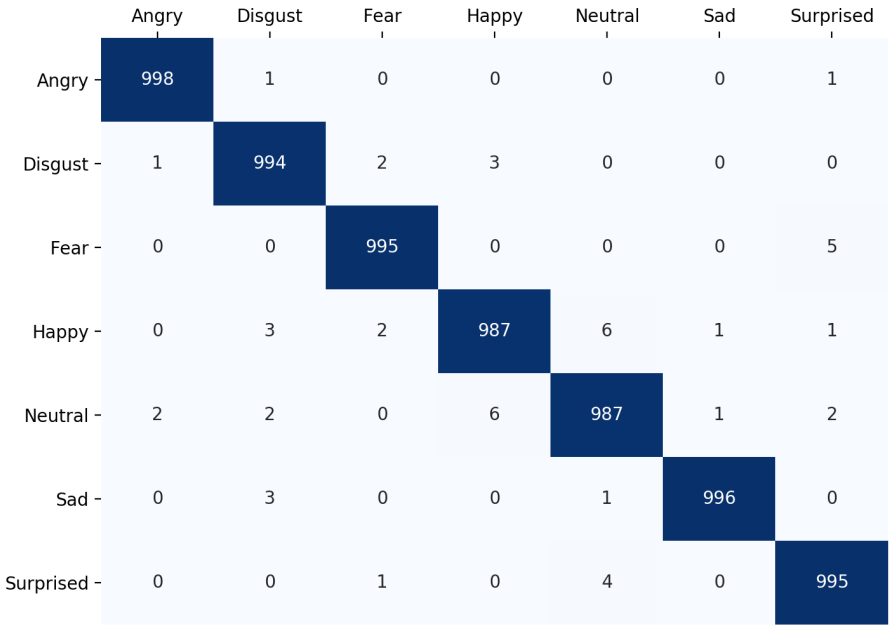

Figure 9. The confusion matrix on the FERG dataset. Each emotion class has 1000 images in the test set.

The confusion matrix of the proposed model on the JAFFE dataset is shown in Figure 10.

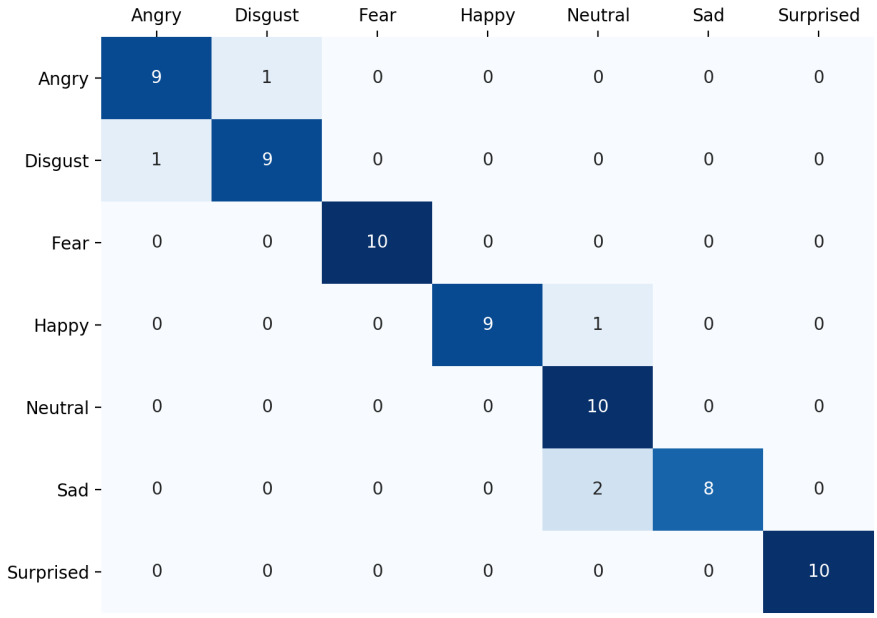

Figure 10. The confusion matrix on the JAFFE dataset. There are a total of seven emotions, and each class has 10 images in the test set.

4.4. Model Visualization

Here, we provide a simple approach for visualizing important regions while classifying different facial expressions, inspired by the work in [20]. We start from the top-left corner of an image, and each time, we zero out a square region of size $N \times N$ inside the image and make a prediction using the trained model on the occluded image. If occluding that region makes the model provide a wrong prediction in terms of facial expression label that region is considered a potential region of importance for classifying that specific expression. On the other hand, if removing that region does not impact the model's prediction, we infer that region as being not very important in detecting the corresponding facial expression. Now, if we repeat this procedure for different sliding windows of $N \times N$, each time shifting them with a stride of s, we can obtain a saliency map for the most important regions in detecting an emotion from different images.

We show nine example cluttered images for a happy and an angry image from the JAFFE dataset and how zeroing out different regions would impact the model prediction in Figure 11. As we can see, for a happy face, zeroing out the areas around the mouth would cause the model to make a wrong prediction, whereas for an angry face, zeroing out the areas around eye and eyebrow makes the model make a mistake.

Figure 12 shows the important regions of seven sample images from the JAFFE dataset, each corresponding to a different emotion. There are some interesting observations from these results. For example, for the sample image with neutral emotion in the fourth row, the saliency region essentially covers the entire face, which means that all of these regions are important to infer that a given image has a neutral facial expression.

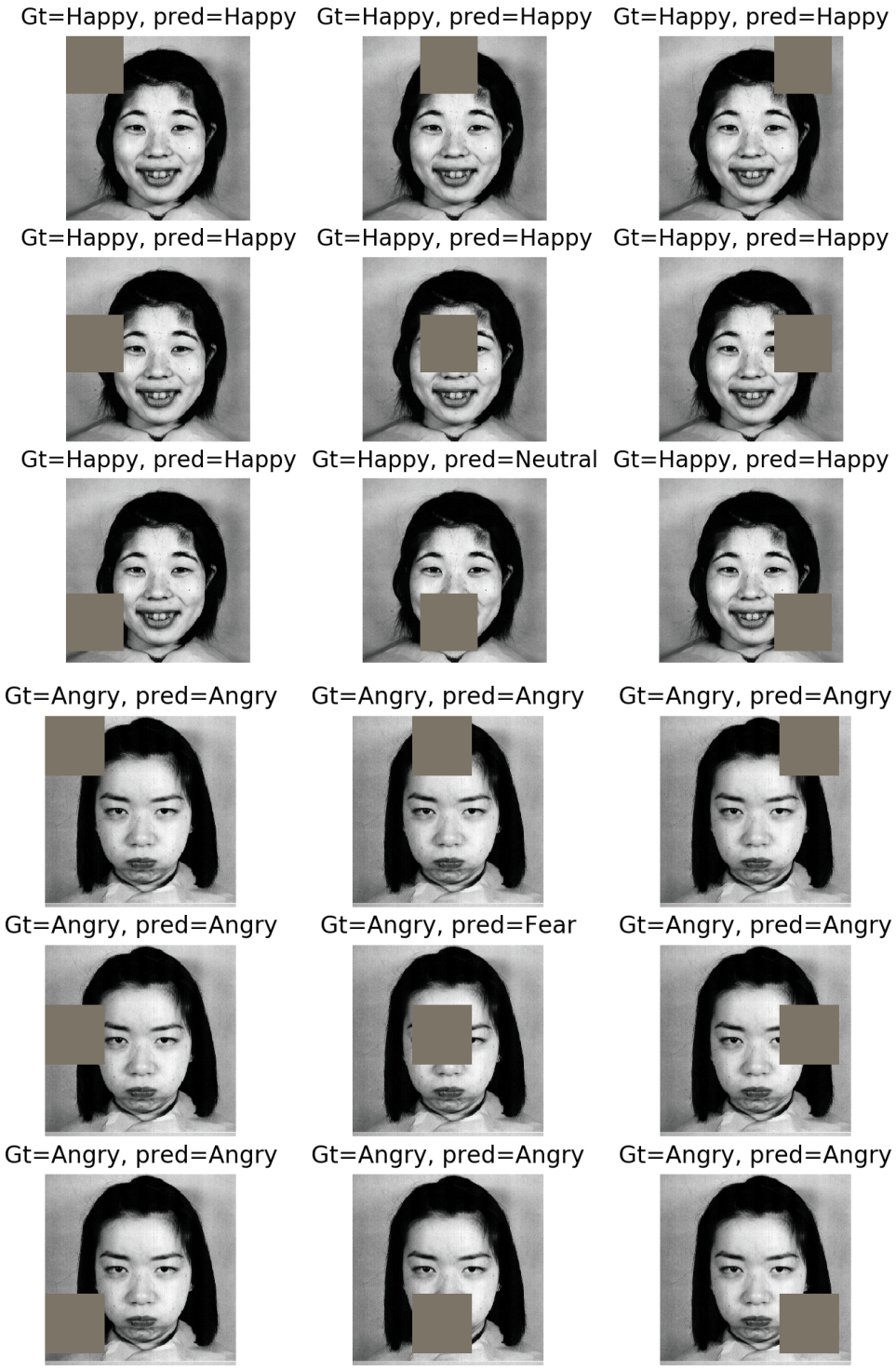

Figure 11. The impact of zeroing out different image parts on the model prediction for a happy face (**the top three rows**) and an angry face (**the bottom three rows**).

This makes sense, since changes in any part of the face (such as the eyes, lips, eyebrows, and forehead) could lead to a different facial expression, and the algorithm needs to analyze all of those parts in order to correctly classify a neutral image. This is however not the case for most of the other emotions, such as happiness, and fear, where the areas around the mouth turns out to be more important than other regions.

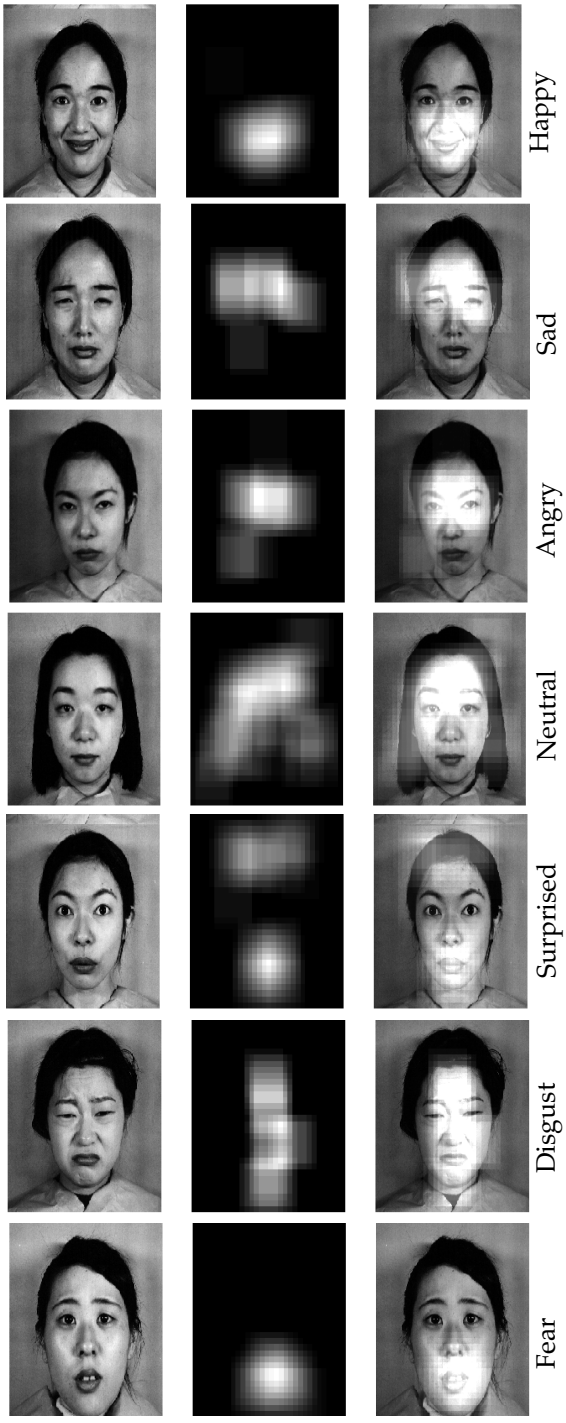

Figure 12. The important regions for detecting different facial expressions. As it can be seen, the saliency maps for happiness, fear, and surprise are sparser than that for the other emotions.

It is worth mentioning that different images with the same facial expression could have different saliency maps due to the different gestures and variations in the image. In Figure 13, we show the important regions for three images with a facial expression of "fear". As seen in this figure, the important regions for these images are very similar when detecting the mouth, but the last one also considers some parts of forehead as important regions. This could be because of the strong presence of forehead lines, which is not visible in the two other images.

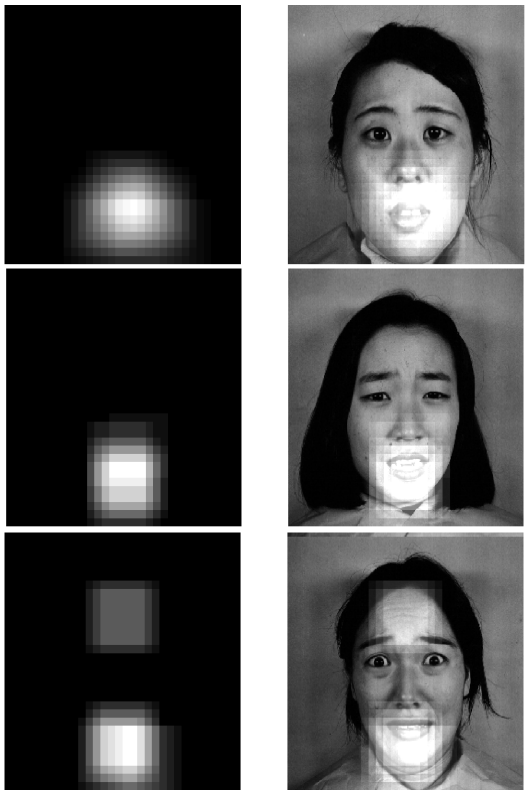

Figure 13. The important regions for three images of fear.

4.5. Model Convergence

In this part, we present the model classification accuracy on the validation set during the training. This helps us obtain a better understanding of the model convergence speed. The model performance on the validation set for the FERG dataset is presented in Figure 14. As seen, the general trend in validation accuracy increases over time but there are some oscillations at some epochs.

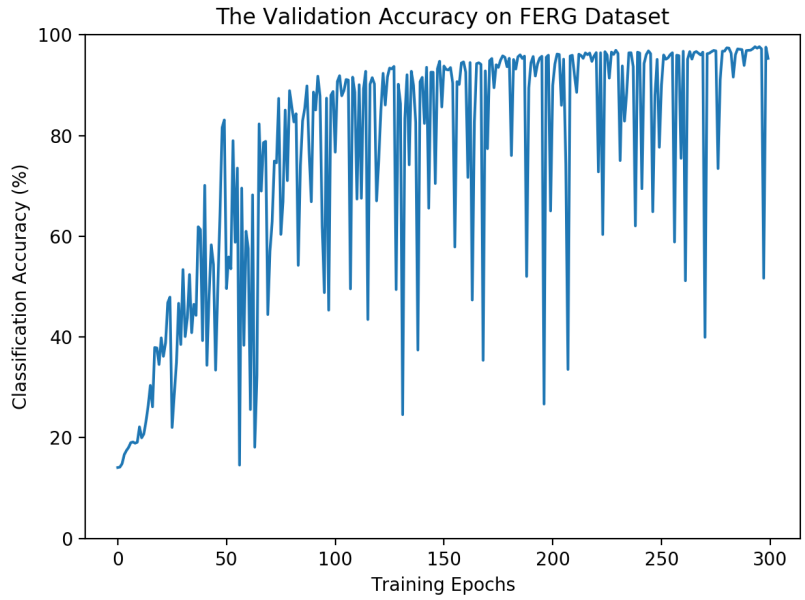

Figure 14. The validation accuracy over different epochs for the model trained on the FERG dataset.

These oscillations could be avoided by choosing a smaller learning rate, but that could lead to slower convergence while training the model. It is worth mentioning that, in the end, the model with the highest validation accuracy is used to report the test error. Through experimental results, we noticed that the choice of learning rate is very important and that choosing a learning rate larger than 0.01 usually leads to model divergence. This could be because of the limited number of samples for some of the datasets.

5. Conclusions

This paper proposes a new framework for facial expression recognition using an attentional convolutional network. We believe attention to special regions is important for detecting facial expressions, which can enable neural networks with less than 10 layers to compete with (and even outperform) much deeper networks in emotion recognition. We also provide an extensive experimental analysis of our work on four popular facial expression recognition databases and show some promising results. Additionally, we deployed a visualization method to highlight the salient regions of face images that are the most crucial parts thereof for detecting different facial expressions.

Author Contributions: S.M. contributed to the algorithm design and also experimental studies, and writing some part of the paper. M.M. contirbuted to the verification of experimental studies, and also revising and re-writing some part of the paper. A.A. contributed to some part of experimental study, and also writing some part of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We express our gratitude to the people at University of Washington Graphics and Imaging Lab (GRAIL) for providing us with access to the FERG database. We also thank our colleagues and partners for reviewing our work and for providing very useful comments and suggestions.

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

References

1. Roddy, C.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80.
2. Deepali, A.; Colburn, A.; Faigin, G.; Shapiro, L.; Mones, B. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 136–153.
3. Jane, E.; Jackson, H.J.; Pattison, P.E. Emotion recognition via facial expression and affective prosody in schizophrenia: A methodological review. *Clin. Psychol. Rev.* **2002**, *22*, 789–832. [CrossRef]
4. Chu, H.C.; Tsai, W.W.; Liao, M.J.; Chen, Y.M. Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning. *Soft Comput.* **2017**, *22*, 2973–2999.
5. Chloé, C.; Vasilescu, I.; Devillers, L.; Richard, G.; Ehrette, T. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Commun.* **2008**, *50*, 487–503.
6. Saste, T.S.; Jagdale, S.M. Emotion recognition from speech using MFCC and DWT for security system. In Proceedings of the IEEE 2017 International Conference o fElectronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 20–22 April 2017; pp. 701–704.
7. Marco, L.; Carcagnì, P.; Distante, C.; Spagnolo, P.; Mazzeo, P.L.; Rosato, A.C.; Petrocchi, S. Computational assessment of facial expression production in ASD children. *Sensors* **2018**, *18*, 3993. [CrossRef]
8. Meng, Q.; Hu, X.; Kang, J.; Wu, Y. On the effectiveness of facial expression recognition for evaluation of urban sound perception. *Sci. Total Environ.* **2020**, *710*, 135484.
9. Ali, M.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the IEEE 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016.
10. Liu, P.; Han, S.; Meng, Z.; Tong, Y. Facial expression recognition via a boosted deep belief network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1805–1812.

11. Kun, H.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014. [CrossRef]
12. Petrantonakis, C.P.; Hadjileontiadis, L.J. Emotion recognition from EEG using higher order crossings. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 186–197.
13. Wu, C.-H.; Chuang, Z.-J.; Lin, Y.-C. Emotion recognition from text using semantic labels and separable mixture models. *ACM Trans. Asian Lang. Inf. Process. TALIP* **2006**, *5*, 165–183.
14. Courville, P.L.C.; Goodfellow, A.; Mirza, I.J.M.; Bengio, Y. *FER-2013 Face Database*; Universit de Montreal: Montréal, QC, Canada, 2013.
15. Akamatsu, M.J.S.L.; Kamachi, M.; Gyoba, J.; Budynek, J. The Japanese female facial expression (JAFFE) database. In Proceedings of the Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 14–16.
16. LeCun, Y. Generalization and network design strategies. *Connect. Perspect.* **1989**, *119*, 143–155.
17. Pooya, K.; Paine, T.; Huang, T. Do deep neural networks learn facial action units when doing expression recognition? In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015.
18. Panagiotis, T.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309.
19. Cohn, F.J.; Zlochower, A. A computerized analysis of facial expression: Feasibility of automated discrimination. *Am. Psychol. Soc.* **1995**, *2*, 6.
20. Zeiler, D.M.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014. [CrossRef]
21. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124.
22. Friesen, E.; Ekman, P.; Friesen, W.; Hager, J. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Psychologists Press: Hove, UK, 1978.
23. Hough, P.V.C. Method and Means for Recognizing Complex Patterns. U.S. Patent 3,069,654, 18 December 1962.
24. Junkai, C.; Chen, Z.; Chi, Z.; Fu, H. Facial expression recognition based on facial components detection and hog features. In Proceedings of the International Workshops on Electrical and Computer Engineering Subfields, Istanbul, Turkey, 22–23 August 2014; pp. 884–888.
25. Caifeng, S.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816.
26. Stewart, M.B.; Littlewort, G.; Frank, M.; Lainscsek, C.; Fasel, I.; Movellan, J. Recognizing facial expression: Machine learning and application to spontaneous behavior. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 568–573.
27. Jacob, W.; Omlin, C.W. Haar features for facs au recognition. In Proceedings of the IEEE FGR 2006 7th International Conference on Automatic Face and Gesture Recognition, Southampton, UK, 10–12 April 2006; p. 5.
28. Alex, K.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.
29. Kaiming, H.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Jonathan, L.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
31. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
32. Minaee, S.; Abdolrashidiy, A.; Wang, Y. An experimental study of deep convolutional features for iris recognition. In Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, PA, USA, 3 December 2016.
33. Shervin, M.; Bouazizi, I.; Kolan, P.; Najafzadeh, H. Ad-Net: Audio-Visual Convolutional Neural Network for Advertisement Detection In Videos. *arXiv* **2018**, arXiv:1806.08612.
34. Ian, G.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *arXiv* **2014**, arXiv:1406.2661.
35. Shervin, M.; Wang, Y.; Aygar, A.; Chung, S.; Wang, X.; Lui, Y.W.; Fieremans, E.; Flanagan, S.; Rath, J. MTBI Identification From Diffusion MR Images Using Bag of Adversarial Visual Features. *arXiv* **2018**, arXiv:1806.10419.
36. 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.
37. Han, Z.; Meng, Z.; Khan, A.-S.; Tong, Y. Incremental boosting convolutional neural network for facial action unit recognition. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 109–117.
38. Meng, Z.; Liu, P.; Cai, J.; Han, S.; Tong, Y. Identity-aware convolutional neural network for facial expression recognition. In Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, Washington, DC, USA, 30 May–3 June 2017; pp. 558–565.

39. Marrero Fernandez, P.D.; Guerrero Pena, F.A.; Ren, T.; Cunha, A. Feratt: Facial expression recognition with attention net. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019.
40. 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, Seattle, WA, USA, 14–19 June 2020.
41. Peng, K.W.; Yang, X.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef]
42. Gan, Y.; Chen, J.; Yang, Z.; Xu, L. Multiple attention network for facial expression recognition. *IEEE Access* **2020**, *8*, 7383–7393.
43. Arpita, G.; Arunachalam, S.; Balakrishnan, R. Deep self-attention network for facial emotion recognition. *Proc. Comput. Sci.* **2020**, *17*, 1527–1534.
44. Shan, L.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**.
45. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015.
46. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA, 13–18 June 2010. [CrossRef]
47. Xiaoqing, W.; Wang, X.; Ni, Y. Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Comput. Intell. Neurosci.* **2018**, [CrossRef]
48. Tudor, R.I.; Popescu, M.; Grozea, C. Local learning to improve bag of visual words model for facial expression recognition. In Proceedings of the ICML Workshop on Challenges in Representation Learning, 2013. Available Online: https://www.semanticscholar.org/paper/Local-Learning-to-Improve-Bag-of-Visual-Words-Model-Ionescu-Grozea/97088cbbac03bf8e9a209403f097bc9af46a4ebb?p2df (accessed on 26 April 2021).
49. Mariana-Iuliana, G.; Ionescu, R.T.; Popescu, M. Local Learning with Deep and Handcrafted Features for Facial Expression Recognition. *arXiv* **2018**, arXiv:1804.10892.
50. Panagiotis, G.; Perikos, I.; Hatzilygeroudis, I. Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013. In *Advances in Hybridization of Intelligent Methods*; Springer: Cham, Switzerland, 2018; pp. 1–16.
51. Vin, T.P.; Vinh, T.Q. Facial Expression Recognition System on SoC FPGA. In Proceedings of the IEEE 2019 International Symposium on Electrical and Electronics Engineering (ISEE), Ho Chi Minh City, Vietnam, 10–12 October 2019.
52. Dimitrios, K.; Zafeiriou, S. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv* **2019**, arXiv:1910.04855.
53. Hang, Z.; Liu, Q.; Yang, Y. Transfer learning with ensemble of multiple feature representations. In Proceedings of the IEEE 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA), Kunming, China,13–15 June 2018.
54. Clément, F.; Piantanida, P.; Bengio, Y.; Duhamel, P. Learning Anonymized Representations with Adversarial Neural Networks. *arXiv* **2018**, arXiv:1802.09386.
55. Ben, N.; Gao, Z.; Guo, B. Facial Expression Recognition with LBP and ORB Features. *Comput. Intell. Neurosci.* **2021**, *2021*, 8828245. [CrossRef]
56. Zaenal, A.; Harjoko, A. A neural network based facial expression recognition using fisherface. *Int. J. Comput. Appl.* **2012**, *59*, 3.
57. Hao, W.; Wei, S.; Fang, B. Facial expression recognition using iterative fusion of MO-HOG and deep features. *J. Supercomput.* **2020**, *76*, 3211–3221.
58. Happy, S.L.; Routray, A. Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affect. Comput.* **2015**, *6*, 1–12.
59. Yoshihiro, S.; Omori, Y. Image Augmentation for Classifying Facial Expression Images by Using Deep Neural Network Pre-trained with Object Image Database. In Proceedings of the ACM 3rd International Conference on Robotics, Control and Automation, Chengdu China, 11–13 August 2018.
60. Salah, R.; Bengio, Y.; Courville, A.; Vincent, P.; Mirza, M. Disentangling factors of variation for facial expression recognition. In *Computer Vision—ECCV 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 808–822.
61. Mengyi, L.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Proceedings of the Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 143–157.
62. Heechul, J.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2983–2991. [CrossRef] [PubMed]
63. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Li, Y. Spatial-temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* **2018**, *99*, 1–9.
64. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-piloted deep network for facial expression recognition. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 425–442.
65. Eleyan, A.M.A.; Akdemir, B. Facial expression recognition with dynamic cascaded classifier. *Neural Comput. Appl.* **2020**, *32*, 6295–6309.