sensors

MDPI

Article

Self-Difference Convolutional Neural Network for Facial Expression Recognition

Leyuan Liu [1,2], Rubin Jiang [1], Jiao Huo [1] and Jingying Chen [1,2,*]

[1] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China; lyliu@ccnu.edu.cn (L.L.); jrubin@mails.ccnu.edu.cn (R.J.); huojiao@mails.ccnu.edu.cn (J.H.)
[2] National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan 430079, China
* Correspondence: chenjy@ccnu.edu.cn; Tel.: +86-135-1721-9631

Abstract: Facial expression recognition (FER) is a challenging problem due to the intra-class variation caused by subject identities. In this paper, a self-difference convolutional network (SD-CNN) is proposed to address the intra-class variation issue in FER. First, the SD-CNN uses a conditional generative adversarial network to generate the six typical facial expressions for the same subject in the testing image. Second, six compact and light-weighted difference-based CNNs, called DiffNets, are designed for classifying facial expressions. Each DiffNet extracts a pair of deep features from the testing image and one of the six synthesized expression images, and compares the difference between the deep feature pair. In this way, any potential facial expression in the testing image has an opportunity to be compared with the synthesized "Self"—an image of the same subject with the same facial expression as the testing image. As most of the self-difference features of the images with the same facial expression gather tightly in the feature space, the intra-class variation issue is significantly alleviated. The proposed SD-CNN is extensively evaluated on two widely-used facial expression datasets: CK+ and Oulu-CASIA. Experimental results demonstrate that the SD-CNN achieves state-of-the-art performance with accuracies of 99.7% on CK+ and 91.3% on Oulu-CASIA, respectively. Moreover, the model size of the online processing part of the SD-CNN is only 9.54 MB (1.59 MB $\times$6), which enables the SD-CNN to run on low-cost hardware.

Keywords: facial expression recognition; difference-based method; self-difference convolutional neural network; facial expression synthesis; facial expression classification

Citation: Liu, L.; Jiang, R.; Huo, J.; Chen, J. Self-Difference Convolutional Neural Network for Facial Expression Recognition. *Sensors* **2021**, *21*, 2250. https://doi.org/10.3390/s21062250

Academic Editor: Marcin Woźniak

Received: 5 February 2021
Accepted: 19 March 2021
Published: 23 March 2021

1. Introduction

Facial expression recognition (FER) aims to classify a static face image or a sequence of face images into one of the typical facial expression classes, such as anger, disgust, fear, happiness, sadness, and surprise [1]. Facial expressions often convey cues about the emotional state and even the intentions of human beings. It is, therefore, unsurprising that automatic FER systems have many practical applications, including, but not limited to, human-robot interaction [2], leaner's interest analysis [3,4], and detection of mental disorders [5]. Thereby, FER has become one of the hottest research topics in the computer vision community.

Benefiting from the development of deep learning, great progresses have been made in the field of FER in the past decade [6]. However, FER is still a challenging problem due to the intra-class variation caused by subject identities. It is rather difficult to transfer face images into a feature space where the distance between two samples of different subjects with the same facial expression is always smaller than the distance between two samples of the same subject with different facial expressions. To better understand the intra-class variation, we picked out 90 samples of 5 randomly selected subjects (each subject with the 6 typical facial expressions and 3 expression intensities) from the CK+ dataset [7],

and visualized the deep features of these samples extracted by the well-known VGG-face network [8] fine-tuned on CK+ using the t-SNE [9]. As shown in Figure 1a, the samples from all the six facial expression classes are entangled with each other, while most of the samples are relatively clustered according to their identities. As most of the samples distribute in the feature space according to the identities rather than facial expressions, it is difficult to design a high-precision classifier for FER.

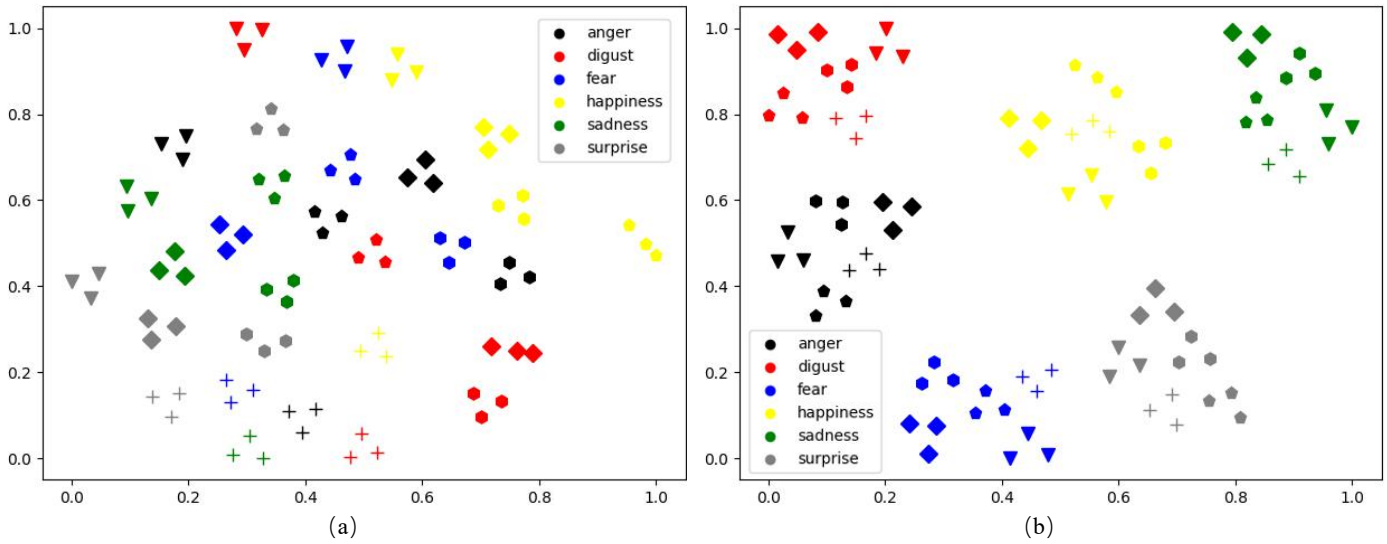

Figure 1. Visualization of deep features output by the fine-tuned VGG-face [8] (**a**) and the self-difference features extracted by the proposed self-difference convolutional network (SD-CNN) (**b**). Each dot represents the deep feature extracted from a face image. Different shapes denote different subject identities, and different colors represent different facial expressions.

To address the intra-class variation issue, several studies [10–12] have tried to recognize facial expression by comparing a subject's expression with the neutral expression of the same subject. In our previous work [10], a deep feature extraction network called deep peak-neutral difference is proposed to use the difference between two deep representations of the fully expressive (peak) and neutral facial expression frames for FER. The peak-neutral-difference-based methods assume that the images with neutral and peak expressions can be discriminated accurately from the facial expression sequences. However, discriminating the neutral or peak expressions is also a tough task. Furthermore, facial expression sequences even may not obtainable in many application scenarios. In order to utilize the peak-neutral-difference approach under such scenarios, Yang et al. [12] treated facial expression as a combination of an expression component and a neutral component (or identity component) of a subject and proposed a De-expression Residue Learning (DeRL) procedure to eliminate the subject identity information. First, the DeRL uses a generative adversarial network to generate the corresponding neutral face image for the input image. Then, given the generated neutral face image, the DeRL learns the expression component deposited in the generative model for FER. Although the DeRL can generate a neutral expression image based on the single input image and tries to remove the identity component, the expression component deposited in the generative model is not powerful enough for facial expression classification.

In this paper, we propose a novel approach called Self-Difference Convolutional Neural Networks (SD-CNN) for FER. Our SD-CNN still uses the feature-level differences between the testing image and the corresponding synthesized expressions of the same subject for FER. However, unlike the existing difference-based FER methods that pick out or synthesize only one reference image with the neutral expression, our SD-CNN uses a conditional Generative Adversarial Network (cGAN) [13] to generate six reference images based on action unit intensities. Each of the generated reference images with one of the six typical facial expressions. Moreover, six compact and light-weighted difference-based

CNNs, called DiffNets, are designed for classifying facial expressions. Each DiffNet extracts a pair of deep features from the testing image and one of the six synthesized reference images, and compares the difference between the deep feature pair. In this way, any potential facial expression in the testing image has an opportunity to be compared with the synthesized "Self"—an image of the same subject with the same facial expression as the testing image. Our insight is that the distance between any potential facial expression and the synthesized "Self" is very small. As the self-difference features extracted from the images with the same facial expression gather tightly in the feature space, the intra-class variation issue is alleviated and, thus, can improve the performance of FER. To confirm this insight, the self-difference features of the 90 samples mentioned above are also visualized using the t-SNE [9]. As illustrated in Figure 1b, the self-difference features of different subjects with the same facial expression are obviously clustered in the low-dimensional visualization space. It means that the self-difference features are discriminative for facial expression classification. Our SD-CNN has been evaluated on two widely-used FER datasets: CK+ [7] and Oulu-CASIA [14]. Experimental results show that our SD-CNN achieves high accuracies of 99.7% on CK+ and 91.3% on Oulu-CASIA, and it outperforms most of the current state-of-the-art FER methods on both the two datasets. Moreover, the model size of the online processing part of our SD-CNN is only 9.54 MB because of the compact and lightweight design of DiffNets.

The main contribution of this paper is three-fold:

(1) We propose a novel approach called SD-CNN for FER. The self-difference feature output by our SD-CNN can significantly alleviate the intra-class variation issue in FER an is discriminative for fical expression classification.
(2) We present a cGAN-base facial expression generator to synthesize reference images with the six typical facial expressions. Conditioned by empirical action unit intensities, the generator can synthesize photo-realistic images with the desired expressions from any input face image.
(3) We design networks called DiffNets to extract the self-difference feature between the testing image and the synthesized expression image. Despite the fact that DiffNets are compact and light-weighted, they achieve state-of-the-art performance on public FER datasets.

The rest of the paper is arranged as follows: the related work is introduced in Section 2, the details of the SD-CNN are described in Section 3, the experiments are presented in Section 4, and the conclusions are given in Section 5.

2. Related Work

FER methods can be divided into two main categories according to the information used: static image-based FER and sequence-based FER. Static image-based methods extract visual features only from a single face image, while sequence-based methods learn representations from a sequence of face images with contiguously changed expressions and may utilize the spatial-temporal information among frames.

2.1. Static Image-Based FER Methods

To mitigate the overfitting problem when training FER models with relatively small-scale facial expression data, many studies fine-tune networks that have been pre-trained on task-related large-scale datasets. Liu et al. [15] proposed to pre-train the VGG-face [8] on a large-scale face-recognition dataset and then fine-tuned the network on a small facial expression dataset for FER. To extend the generalizability of FER models, Mollahosseini et al. [16] fine-tuned a GoogleLeNet-like [17] network on a set of public available facial expression datasets. Ding et al. [18] presented a method called FaceNet2ExpNet to transform visual representations from the face recognition domain into the facial expression domain. The convolutional layers of the FaceNet2ExpNet are first trained and regularized by an off-the-shelf face recognition network, then the whole FaceNet2ExpNet, which consists of the pre-trained convolutional layers and fully-connected layers, is re-trained jointly on facial

expression datasets. The FaceNet2ExpNet achieves an accuracy of 98.6% on the small-scale CK+ dataset when following a 10-fold cross-validation protocol. Although pre-training FER networks on other face-related datasets help improve the performance, the identity information retained in the pre-trained models may have a negative impact on the accuracy of FER [10].

To address the identity retaining problem, several researchers proposed to subtract the identity component from the face image. In order to develop a pose-invariant facial expression recognition algorithm, Zhang et al. [19] presented a GAN-based model to synthesize face images with different poses and expressions. In Reference [20], a GAN-based approach, called Identity-Adaptive Generative model (IA-gen), is proposed to alleviate the identity retaining issue by minimizing the distance between the synthesized prototypic facial expressions and the query image. Based on a CNN fine-tuned on FER datasets, the IA-gen achieves accuracies of 96.75% on CK+ and 88.92% on Oulu-CASIA. Fabiano et al. [21] first utilized a deep-learning-based 3D morphable model to synthesize 3D images with different expressions, and then employed the synthesize 3D images to train deep neural networks for facial emotion recognition on 3D datasets.

Recently, many researchers embedded the attention mechanism into FER networks to make the networks focus on the most expressive facial regions. In order to suppress the uncertainties caused by low-quality face images and ambiguous facial expressions, Wang et al. [22] proposed a Self-Cure Network (SCN) to weight each sample in training dataset with a ranking regularization by embedding a self-attention mechanism into neural networks. The SCN outperforms most of the state-of-the-art methods on several public FER datasets. Wang et al. [23] proposed a Region Attention Network (RAN), which employs a region biased loss to encourage high attention weights for the important face regions, to make the network extracts the most expression-related features for FER. Recently, Gan et al. [24] proposed to simulate the coarse-to-fine attention mechanism of human beings, and developed a multiple attention network to boost the performance of facial expression recognition. However, this multiple attention network only achieves an accuracy of 96.4% on the CK+ dataset.

In order to apply automatic FER systems on smart-phone and other embedded platforms, many efforts have been made to design compact and light-weighted neural networks for FER. In order to train a FER classifier with a lower requirement of computing resources, Zhu et al. [25] designed a neural network called IExpressNet-based on the ResNet-18. The IexpressNet is trained in an incremental manner and, thus, reduces the time consumption by over 80%. Zeng et al. [26] developed a FER network that consists of only 3 convolutional layers. Although this network is rather compact, the best accuracy on the CK+ dataset achieved by it is 95.79%, which is much lower than the deeper FER networks, like FaceNet2ExpNet [18] and DeRL [12].

2.2. Sequence-Based FER Methods

In order to reduce the difficulty of facial expression recognition, most sequence-based methods focus on utilizing frames with the most expressive (peak) expressions for FER. Zhao et al. [11] proposed a peak-piloted deep network (PPDN) that embeds the facial expression evolution from non-peak to peak frames into the network parameters. By using a special-purpose back-propagation procedure called PGS for network optimization, the PPDN drives the intermediate-layer features of the non-peak expression sample towards those of the peak expression sample. Yu et al. [27] proposed a deeper cascade peak-piloted network (CPPN) to employ the most expressive images for boosting the performance of the weak facial expression recognition. Similar to the PPDN [11], the CPPN also uses the peak expression to supervise the non-peak expression of the same subject. Despite the fact that CCPN achieves an accuracy of 99.30% on the CK+ dataset, the network of CCPN is 42 layers deep.

Many studies exploited both the spatial and temporal information to to improve the performance of facial expression recognition. To extract representative features from facial expression sequence, Chen et al. [28] proposed a framework named Facial Motion Prior Networks (FMPN) to employ domain knowledge for FER. The insight of the FMPN is that focusing on facial muscle moving regions can help to extract representative features and, thus, improve the FER performance. In order to improve the performance of FER by exploiting more information, Jung et al. [29] proposed to train two CNN-based networks joinlty to squeeze spatial-temporal information form the facial expression sequences and the temporal facial landmark points, respectively. Kuo et al. [30] proposed a compact CNN-based (CompactCNN) approach that employs the gated recurrent units to exploit temporal information among frames. By exploiting the temporal information, the performance of the CompactCNN is boosted from 97.37% to 98.47% on the CK+ dataset. Recently, Meng et al. [31] developed an end-to-end framework called Frame Attention Networks (FAN) to aggregate the spatial features among frames with temporal information. The FAN first employs a spatial attention module to learn an attention weight for each frame, then aggregate the features extracted from all the frames to form a single representation adaptively based on the attention weight of each frame. As a result, the FAN achieves the state-of-the-art performance on the CK+ dataset with an accuracy of 99.68%. Despite the fact that performance of FER can be improved by squeezing spatial-temporal information among frames, facial expression sequences are not always obtainable in many application scenarios.

3. Methodology

3.1. Overview

The proposed FER method consists of two main modules, i.e., the facial expression generator and the facial expression classifier, as illustrated in Figure 2. Given the empirical Action Unit (AU) vectors of facial expressions, the facial expression generator employs a cGAN [13,32] to synthesize photo-realistic reference images with the six typical facial expressions (i.e., happiness, sadness, surprise, disgust, anger, and fear) from an input face image under an arbitrary facial expression. The facial expression classifier is composed of six DiffNets. Each DiffNet first extracts a pair of deep features from the testing image and one of the synthesized reference images, respectively, by two subnets named the query-subnet and the reference-subnet; then produces a difference vector by comparing the features extracted by these two subnets; and finally outputs a probability distribution over predicted facial expressions. As the six DiffNets output probability distributions over predicted facial expressions for each testing image, a voting scheme is adopted to make the final decision. In our voting scheme, each DiffNet holds a vote and uses a "winner-take-all" scheme to cast its vote. After being voted by all of the six DiffNets, the testing image is classified as the facial expression class that has received the most votes.

The networks in our method are divided into the offline components and the online components according to run-time. As illustrated in Figure 2, only the query-subnets in the DiffNets need to process online, while all the other networks can run offline. Since each query-subnet consists of only 4 light-weighted convolutional layers and 3 narrow fully-connected layers, our method is able to run on devices with low-cost processors and memory.

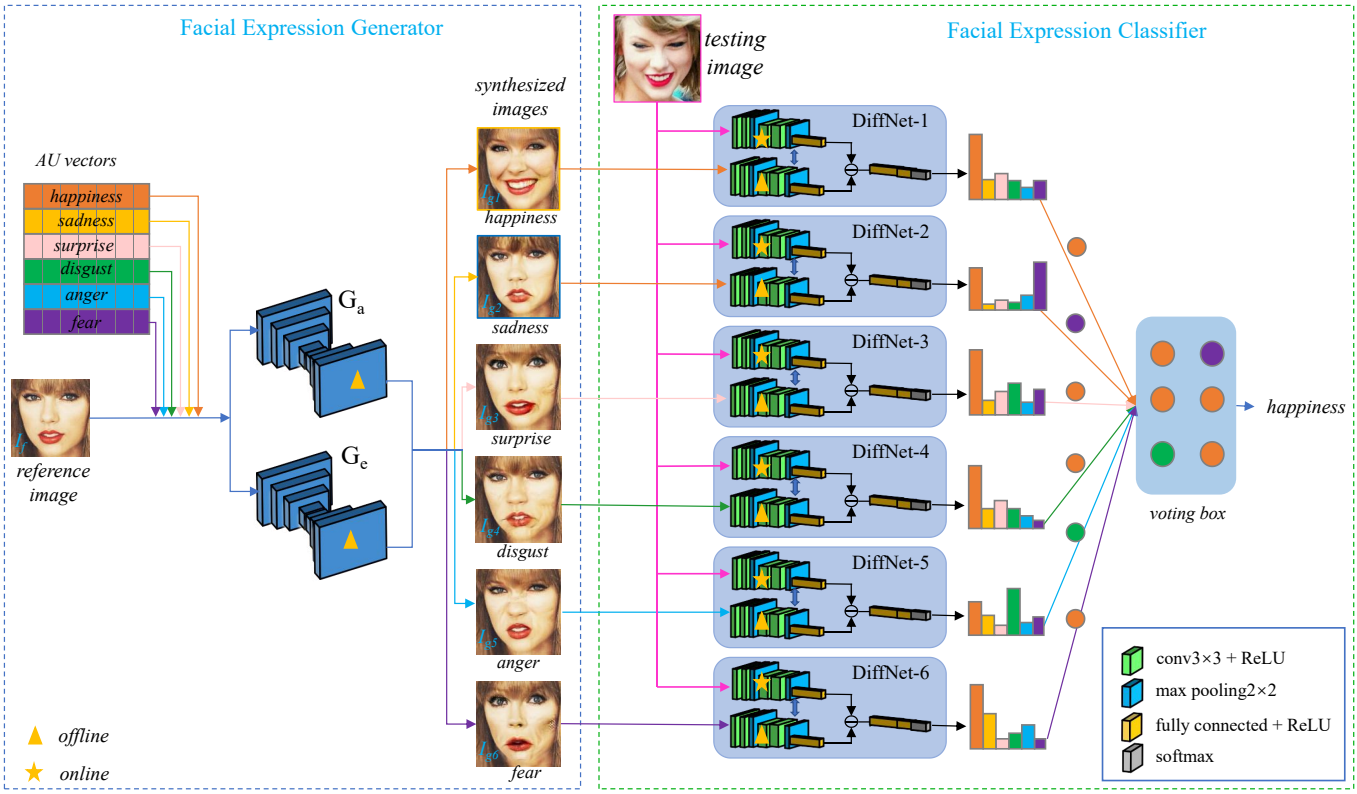

Figure 2. Framework of the proposed facial expression recognition method. The proposed method which at its core consists of two modules, i.e., the facial expression generator and the facial expression classifier. The facial expression generator synthesizes photo-realistic images with the six typical facial expressions from an input face image under an arbitrary facial expression. The facial expression classifier uses six DiffNets for facial expression classification.

3.2. Facial Expression Generator

Facial Expression Representation. The aim of our facial expression generator is to synthesize photo-realistic images with desired facial expressions from an input face image under an arbitrary facial expression. Recently, Pumarola et al. [32] proposed a deep network architecture for synthesizing facial expressions by a novel GAN conditioning scheme based on facial action units. Inspired by this work, we also employ a cGAN conditioned by facial action units to synthesize the six typical facial expressions, including anger, disgust, fear, happiness, sadness, and surprise. For this purpose, each facial expression is encoded by a vector which consists of a set of K action unit intensities:

$$\boldsymbol{V}_i = (v_{i1}, \cdots, v_{ik}, \cdots, v_{iK})^T, \tag{1}$$

where $i \in \{1, \cdots, 6\}$ is the index for each of the six typical facial expressions, and $v_{ik} \in [0, 1]$ denotes the normalized intensity of the kth action unit. According to the Facial Action Coding System (FACS) [33], each of the six facial expressions can be represented by a vector of empirical action unit intensities that :$\tilde{\boldsymbol{V}}_i = (\tilde{v}_{i1}, \cdots, \tilde{v}_{iK})^T$. In our implementation, empirical action unit intensities are calculated from the EmotioNet dataset [34]. To make the generated facial expressions more diverse, we sample an action unit vector for each subject from the empirical action unit intensities using the Gaussian distribution.

$$v_{ik} \sim \mathcal{N}(\tilde{v}_{ik}, \sigma_{ik}^2), \tag{2}$$

where $\mathcal{N}$ denotes the Gaussian distribution, and σ_{ik}^2 is a given variance.

Network Structure. As illustrated in Figure 3, the network structure of our cGAN-based facial expression generator mainly consists of the generative networks G and the discriminative networks D. The generative networks are designed to synthesize a photo-

realistic image $\hat{I}$ with the desired facial expression V_i from the input image I. In order to make the generative networks focus exclusively on the image regions that are responsible of synthesizing the desired facial expression, the attention mechanism is embedded into the generative networks [32]. Concretely, the generative networks are composed of an attention network G_a and an expression synthesis network G_e, that is, $G = (G_a, G_e)$. The output of G_a is a pixel-wise mask that indicates the contribution of each pixel to the desired facial expression. Then, the synthesized image is produced by the outputs of the two networks and the original input image: $\hat{I} = G_a(I|V_i)G_e(I|V_i) + [1 - G_a(I|V_i)]I$. In order to utilize the unsupervised strategy to train the facial expression generator, the generative networks are applied twice. First, we apply the generative networks to transform the input image I to the synthesized image $\hat{I}$ with the desired facial expression V_i: $\hat{I} = G(I|V_i)$; then, we adopt them again to recover the input image from the synthesized image $\hat{I}$ using the ground-truth facial action unit intensities$\tilde{V}_i$: $\tilde{I} = G(\hat{I}|\tilde{V}_i)$, where $\tilde{I}$ denotes the recovered image. By using this strategy, the generator does not require supervision, i.e., no pairs of images of the same person with different expressions, nor the target image $\hat{I}$ are assumed to be known. The discriminative networks D are composed of a reality judgement network D_r, which valuates the synthesized image in its photo-realism, an identity verification network D_i which verifies the input image and the synthesized image are the same individual, and an expression estimation network D_e which estimates the action unit intensities ($\hat{V}$) of the synthesized image.

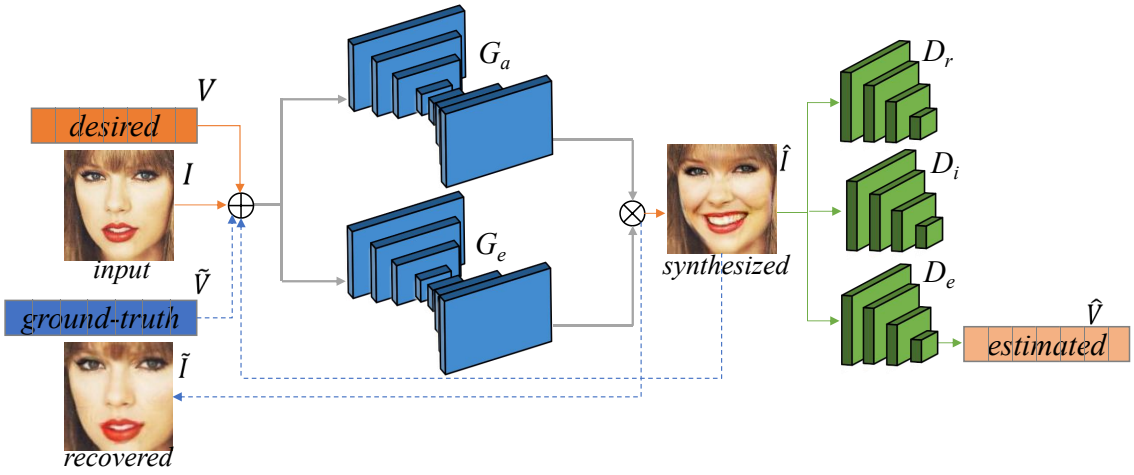

Figure 3. Network structure of the facial expression generator.

Loss Functions Our facial expression generator is trained following a min-max game:

$$G^* = \arg\min_G \max_D \mathcal{L}_G, \tag{3}$$

where $\mathcal{L}_G$ is the loss function which consists of five parts: the adversarial loss ($\mathcal{L}_d$), the attention loss ($\mathcal{L}_a$), the identity loss ($\mathcal{L}_i$), the perceptual loss ($\mathcal{L}_p$), and the expression loss ($\mathcal{L}_e$). The adversarial loss proposed by WGAN-GP [35] is employed:

$$\begin{aligned}\mathcal{L}_d =&(\mathbb{E}_I[D_r(G(I|V_i))] - \mathbb{E}_I[D_r(I)])+\\ &\lambda_1\mathbb{E}_{\dot{I}}[(||\nabla_{\dot{I}}D_r(\dot{I})||_2 - 1)^2],\end{aligned} \tag{4}$$

where $\dot{I}$ is the input image with random noise, and λ_1 is a penalty coefficient. Following Reference [32], the attention loss is defined as:

$$\begin{aligned}\mathcal{L}_a =&\mathbb{E}_I[\sum_{i,j}[(\nabla_x G_a(I|V_i)_{i,j})^2 + (\nabla_y G_a(I|V_i)_{i,j})^2]]+\\ &\lambda_2\mathbb{E}_I[||G_a(I|V_i)||_2],\end{aligned} \tag{5}$$

where ∇ denotes the gradient of an image, and λ_2 is a penalty coefficient. The identity loss ($\mathcal{L}_i$) is defined for identity verification:

$$\mathcal{L}_i = y_1 log(q_1) + (1 - y_1) log(1 - q_1), \tag{6}$$

where $y_1 = 1$ if the input image and the synthesized image are predicted as the same individual by the identity verification network D_i, and q_1 denotes the probability that the input image and the synthesized image the same individual. The perceptual loss [36] is defined by the difference between the input image and the recovered image:

$$\mathcal{L}_p = \mathbb{E}_I[||I - G(G(I|V_i)|\tilde{V}_i)||_2]. \tag{7}$$

The expression loss is defined as:

$$\mathcal{L}_e = \mathbb{E}_I[||D_e(G(I|V_i)) - V_i||_2] + \mathbb{E}_I[||D_e(I) - \tilde{V}_i||_2]. \tag{8}$$

Before training the whole facial expression generator, we pre-train the expression estimation network D_e using the following simple loss function $\mathcal{L}_{e2}$:

$$\mathcal{L}_{e2} = \mathbb{E}_I[||D_e(I) - \tilde{V}_i||_2]. \tag{9}$$

Finally, the whole loss function is built by combining all the four loss functions:

$$\mathcal{L}_G = \lambda_d \mathcal{L}_d + \lambda_a \mathcal{L}_a + \lambda_i \mathcal{L}_i + \lambda_p \mathcal{L}_p + \lambda_e \mathcal{L}_e, \tag{10}$$

where λ_d, λ_a, λ_p, and λ_e are the hyper-parameters.

3.3. Facial Expression Classifier

The DiffNet. The architecture of a DiffNet is given in full detail in Figure 4. A DiffNet consists of two branched subnets, i.e., the query-subnet which is used to extract features from the testing image, and the reference-subnet which is used to extract features from a synthesized reference image. The query-subnet and the reference-subnet are designed as two peer networks with the same structure, but they do not share parameters for each other. A query-subnet/reference-subnet comprises two convolutional units and a fully-connected layer. Both the two convolutional units are composed of two convolutional layers followed by a ReLU and a max pooling, but the numbers of their convolutional kernels are, respectively, 32 and 64. After processed by a subtraction operation, a feature-level difference vector between the testing image and the synthesized expression of the same subject is obtained. Based on the difference vector, the DiffNet uses two fully-connected layers and a Softmax for facial expression classification. Since the DiffNet is designed using a compact and light-weighted structure, the number of parameters, the model size, and the number of operations for a DiffNet are, respectively, 0.395 M, 1.59 MB, and 0.033G FLOPs.

Self-Difference Feature. As mentioned before, the facial expression classifier comprises six DiffNets, each of which compares the difference deep features extracted from the input image and one of the six synthesized reference expression images. Therefore, any potential facial expression in the input image has an opportunity to be compared with the synthesized "Self" facial expression. The distance between any potential facial expression and the synthesized "Self" is usually very small. Thus, the self-difference features extracted from the images with the same facial expression will gather in the feature space, which will alleviate the intra-class variation issue. Although compact and light-weighted DiffNets are employed for facial expression classification, the self-difference feature is discriminative enough.

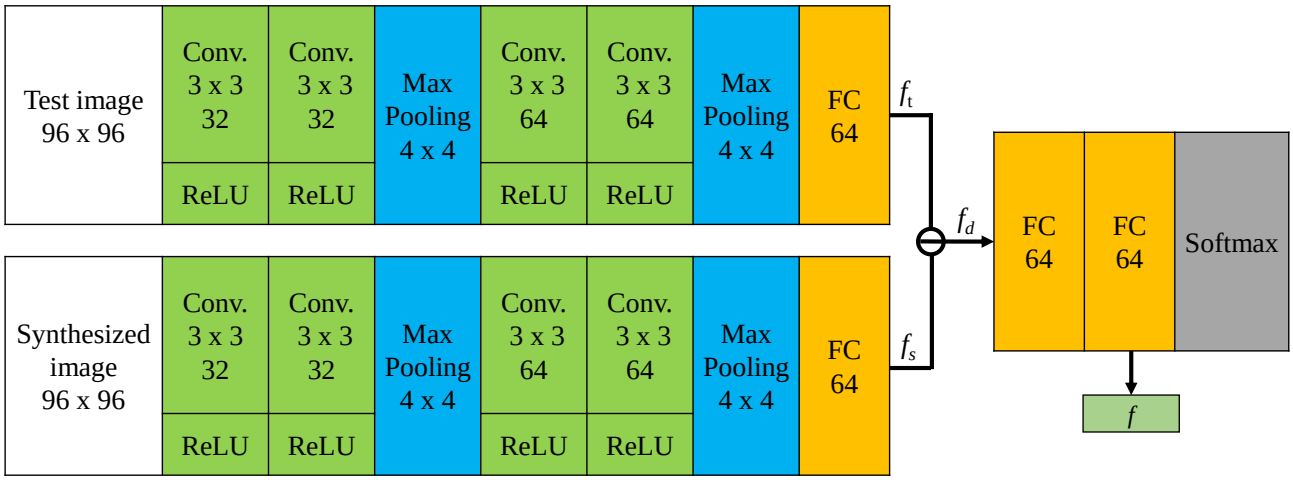

Figure 4. Network structure of the DiffNet for facial expression classification.

Loss Functions. The loss function of each DiffNet is composed of two partial terms: the cross-entropy loss $\mathcal{L}_c$ and the triplet loss $\mathcal{L}_t$. The cross-entropy loss is defined as:

$$\mathcal{L}_c = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{6} y_i^k \log(p_i^k), \tag{11}$$

where M is the number of samples in a mini-batch, $y_i^k \in \{0,1\}$ is the ground-truth label that indicates whether the ith sample belongs to the kth facial expression class, and p_i^k denotes the predicted Softmax probability that the ith sample belongs to the kth facial expression. The triplet loss [37], which is popularly used in the filed of person re-id, is also utilized to optimize our DiffNets. When training a DiffNet with the triplet loss, each randomly selected sample is called as the "anchor". Except for the anchor, we randomly select two additional samples: one belongs to the same facial expression class as the anchor, while the other one with a different facial expression against the anchor. These two samples are, respectively, called as the "positive" and the "negative". The deep features of these three samples output by the last fully-connected layer of a DiffNet are, respectively, denoted as f^a, f^p, and f^n. Thereby, a triplet is formed by $< f^a, f^p, f^n >$, and the triplet loss is defined using this triplet:

$$\begin{aligned}\mathcal{L}_t = -\sum_{i=1}^{P}\sum_{a=1}^{K}[&\max_{p=1\ldots K}||f_i^a - f_i^p||_2 \\ &- \min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq a}}||f_i^a - f_j^n||_2 + m]_+{}',\end{aligned} \tag{12}$$

where m is the margin between the intra-class and inter-class, and $[\cdot]_+ = max(\cdot, 0)$. In each mini-batch, we select P facial expression class and K images from each class. Since the positive pair (f^a, f^p) are extracted from the samples with the same facial expression, while the negative pair (f^a, f^n) are extracted from images with different facial expressions, the purpose of triplet loss is to make the intra-class distance as small as possible while making the inter-class distance as large as possible. Finally, the full loss function is defined by linearly combining the two loss terms:

$$\mathcal{L}_C = \lambda_c L_c + \lambda_t L_t, \tag{13}$$

where λ_c and λ_t are the hyper-parameters that control the relative importance of the two loss terms.

Voting Scheme. As the six DiffNets in our facial expression classifier output six groups of probabilities for each testing image, a voting scheme is designed to make the final decision. In our voting scheme, each DiffNet holds a vote and uses a "winner-take-all" scheme to cast its vote. Formally, let $\{p_j^1, \cdots, p_j^k, \cdots, p_j^6\}_{j=1,\cdots,6}$ denote the probabilities output by the Softmax layer of the jth DiffNet. Then, the jth DiffNet casts its vote to the k^*-th facial expression class:

$$k^* = \arg\max_k \{p_j^1, \cdots, p_j^k, \cdots, p_j^6\}. \tag{14}$$

After being voted by all of the six DiffNets, the testing image is classified as the facial expression class that has received the most votes. If more than one class has received the same votes, then the final decision is made by the maximum Softmax probability output by the DiffNets that cause the tie.

4. Experimental Results

4.1. Implementation Details

The facial expression generator and the DiffNets in the facial expression classifier are trained independently. The expression generator is trained on 500,000 images randomly selected from the EmotioNet dataset [38]. The hyper-parameters for the partial loss functions in the generator are set as $\lambda_d = 1$, $\lambda_a = 0.1$, $\lambda_p = 5$, $\lambda_i = 5$, and $\lambda_e = 4000$.

The six DiffNets are trained on the CK+ [7] and Oulu-CASIA [14] datasets following a 10-fold cross-validation protocol. We do not utilize the "pre-training then fine-tuning" scheme for training our models. In other words, the DiffNets have not been pre-trained on any other dataset. Since both the CK+ and Oulu-CASIA datasets consist of sequences/videos with continuous facial expressions range from neutral to peak, only the images whose sequence number is greater than 8 are captured for training. Each face in the images is cropped and resized to 128×128 pixels. In order to extend the training samples, half of the images in the original training dataset are randomly selected and augmented using FiveCrop [39] and horizontal flipping. The weight parameters of the partial loss functions for the DiffNets are set as $\lambda_c = 1$, and $\lambda_t = 10$. When training the DiffNets, the Adam with a momentum of 0.9 is adopted as the optimizer, and the mini-batch size is fixed to 64. Since each of the fully-connected layers only consists of 64 neurons, dropout is utilized to avoid overfitting. The dropout rate is set as 0.5. Each DiffNet is first trained independently for 1000 epochs, and then all the six DiffNets are jointly trained for another 500 epochs. All the deep neural networks in our method are implemented by PyTorch. It takes about 24 h for training all the six DiffNets on a computer with a single NVIDIA GeForce 2080 Ti GPU.

4.2. Datasets and Performance Metric

Datasets. The proposed SD-CNN is extensively evaluated on two widely-used FER datasets: CK+ [7] and Oulu-CASIA [14]. The CK+ dataset includes 593 sequences from 123 subjects, and involves seven facial expression classes (i.e., anger, contempt, disgust, fear, happiness, sadness, and surprise). In our experiments, the 18 sequences with contempt expression are ignored. The Oulu-CASIA dataset contains 2280 videos with the six typical facial expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise) from 80 subjects. The videos in Oulu-CASIA are captured with two imaging systems, NIR (Near Infrared) and VIS (Visible light), but we only use the VIS videos in our experiments. As the sequences/videos in both CK+ and Oulu-CASIA record faces with facial expressions range form neutral to peak, we cut all the sequences/videos into images and ignore the images with neutral expressions. Totally, we capture 3482 images from the sequences in CK+ and 6540 images from the videos in Oulu-CASIA. As mentioned above, our experiments conduct on CK+ and Oulu-CASIA follow a 10-fold cross-validation protocol. The frames from one sequence are kept only either in the training set or in the testing set. When conducting experiments on CK+, 3134 images are used for training and 348 images are

used for testing; when conducting experiments on Oulu-CASIA, 5886 images are used as the training dataset, and 654 images are used as the testing dataset. The class distribution of images in the CK+ and Oulu-CASIA datasets is listed in the Table 1. Some samples randomly selected from these two datasets are shown in Figure 5.

Table 1. Distribution of the test images in the CK+ and Oulu-CASIA.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
CK+	13.07%	20.07%	10.68%	24.35%	10.20%	21.63%
Oulu-CASIA	17.58%	15.18%	17.68%	17.60%	15.73%	16.22%

Performance Metric. The Accuracy (ACC) is adopted as the metric for evaluating the facial expression recognition performance:

$$ACC = [\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i == y_i)] \times 100\%, \quad (15)$$

where N is the number of samples in the testing dataset, and y and $\hat{y}$ are, respectively, the ground-truth label and the predicted facial expression class of a testing sample.

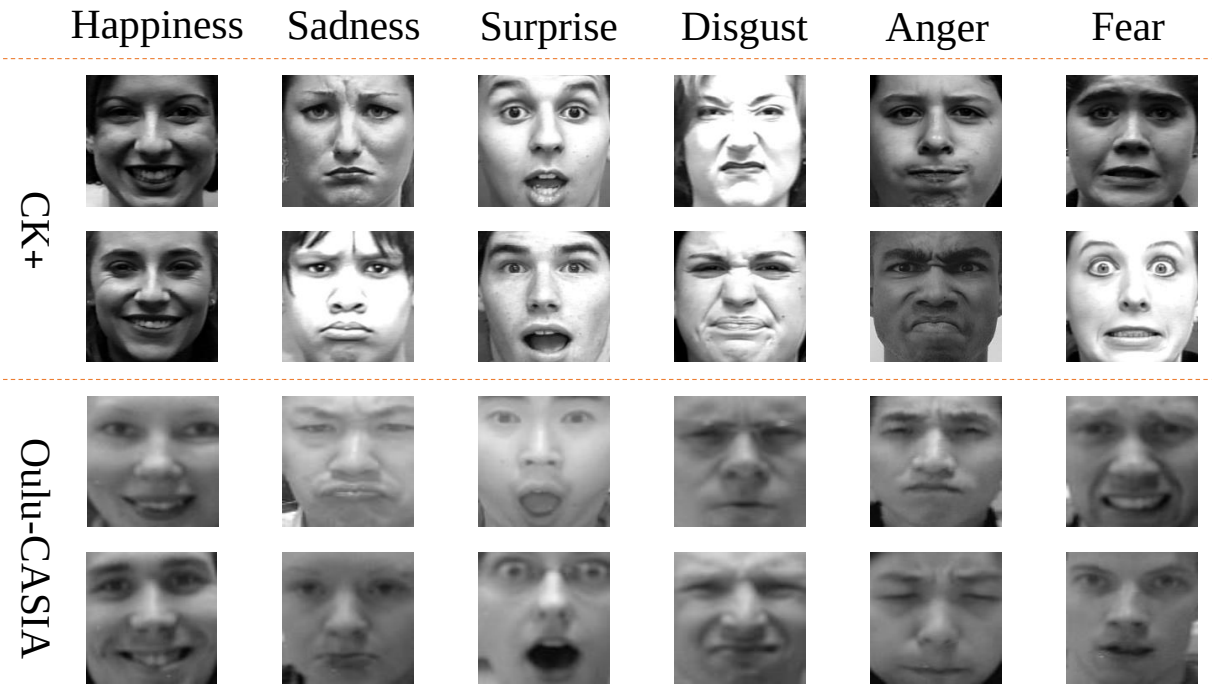

Figure 5. Example samples randomly selected from the CK+ [7] and Oulu-CASIA [14] datasets.

4.3. Experiments on CK+

Our SD-CNN method is compared with the recent state-of-the-art FER methods, including DeRL [12], FN2EN [18], FMPN [28], VGG-face [8] , MicroExpNet [34], GoogLeNet [17], MultiAttention [24], DSAE [26], GCNet [40], DynamicMTL [41], IA-gen [20], CompactCNN [30], DTAGN(Joint) [29], CPPN [27], DPND [10], PPDN [11], and FAN [31]. Table 2 reports our experimental results and shows the comparisons with these methods. In Table 2, the methods are divided into two groups according to the information used: static image-based FER and sequence-based FER. It is can be seen that our method achieves the best performance among the static image-based methods with an accuracy of 99.7%, which surpasses the current state-of-the-art image-based method (i.e., the DeRL [12]) by 0.4%. Even though much less information is used, our method achieves very competitive performance with respect to the state-of-the-art sequence-based method (i.e., the FAN [31]). For a subject,

our method only need a single image for facial expression recognition, but the FAN [31] needs to capture a sequence of images with facial expressions range from neural to peek. However, image sequence with different expression intensities for an individual subject is not always available in practice. Moreover, the high performance of sequence-based methods depends on the accuracy of expressive (peak) frames identification. As reported in Reference [27], when conducting experiments only on the images with weak expressions, the performance of sequence-based methods usually decreases sharply. For example, the accuracies of the PPDN [11] and the CPPN [27] on the weak expression images in CK+, respectively, decrease from 99.3% and 98.3% to 83.4% and 92.5%.

Table 2. Comparisons with state-of-the-art methods on the CK+ datasets.

Methods	ACC(%)	Image/Sequence
SD-CNN (Ours)	**99.7**	Image
DeRL [12]	99.3	Image
FN2EN [18]	98.6	Image
FMPN [28]	98.0	Image
VGG-face (fine-tuned) [8]	94.9	Image
GoogLeNet (fine-tuned) [17]	95.3	Image
MicroExpNet [34]	96.9	Image
MultiAttention [24]	96.4	Image
DSAE [26]	95.8	Image
GCNet [40]	97.3	Image
DynamicMTL [41]	99.1	Image
CompactCNN (frame-based) [30]	97.4	Image
IA-gen [20]	96.6	Image
FAN [31]	**99.7**	Sequence
FAN(w/o attention) [31]	99.1	Sequence
CompactCNN [30]	98.5	Sequence
DTAGN(Joint) [29]	97.3	Sequence
CPPN [27]	98.3	Sequence
DPND [10]	94.4	Sequence
PPDN [11]	99.3	Sequence

Figure 6 shows the confusion matrix of our SD-CNN on the CK+ dataset. It can be observed that happiness, sadness, fear and surprise are almost recognized perfectly, with accuracies of 99.88%, 99.72%, 99.46%, and 99.73%, respectively. Anger and disgust are relatively hard to recognize. For the testing samples with anger, 0.28% of them are misclassified as disgust, and 0.14% of them are misclassified as surprise. For the testing samples with disgust expression, 0.44% of them are misclassified as anger. Figure 7 illustrates some testing samples in CK+ that are misclassified by our method. Obviously, most of these misclassified samples are with very weak expressions, which are rather hard to be classified correctly by human beings.

In order to evaluate the effect of the facial expression generator', we use only 1 DiffNet to compare the test image with the reference (neutral expression) image and the performance on CK+ decreases from 99.7% to 92.4%.

To better understand why our method achieves good performance, we visualize the self-difference features extracted by our SD-CNN and deep features extracted by the fine-tuned VGG-face [8] using t-SNE [9]. As shown in Figure 8, the self-difference features output by our method are closely gathered according to their facial expression classes in the two-dimensional visualization space, while the deep features output by the VGG-face are entangled with each other, and there is obvious confusion among different classes.

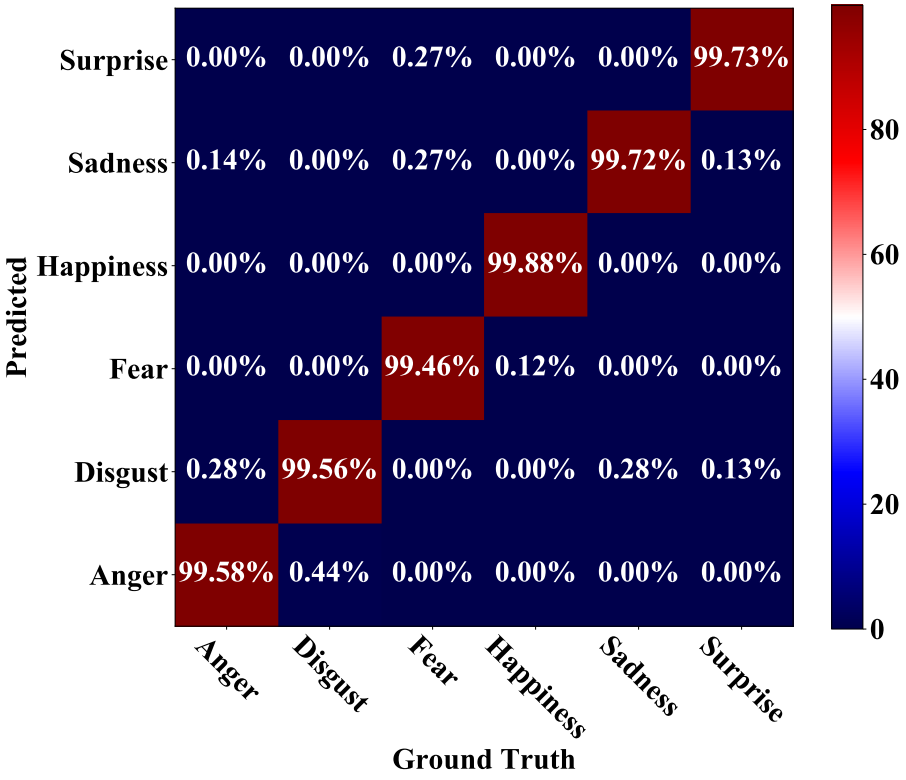

Figure 6. Confusion matrix for the CK+ dataset.

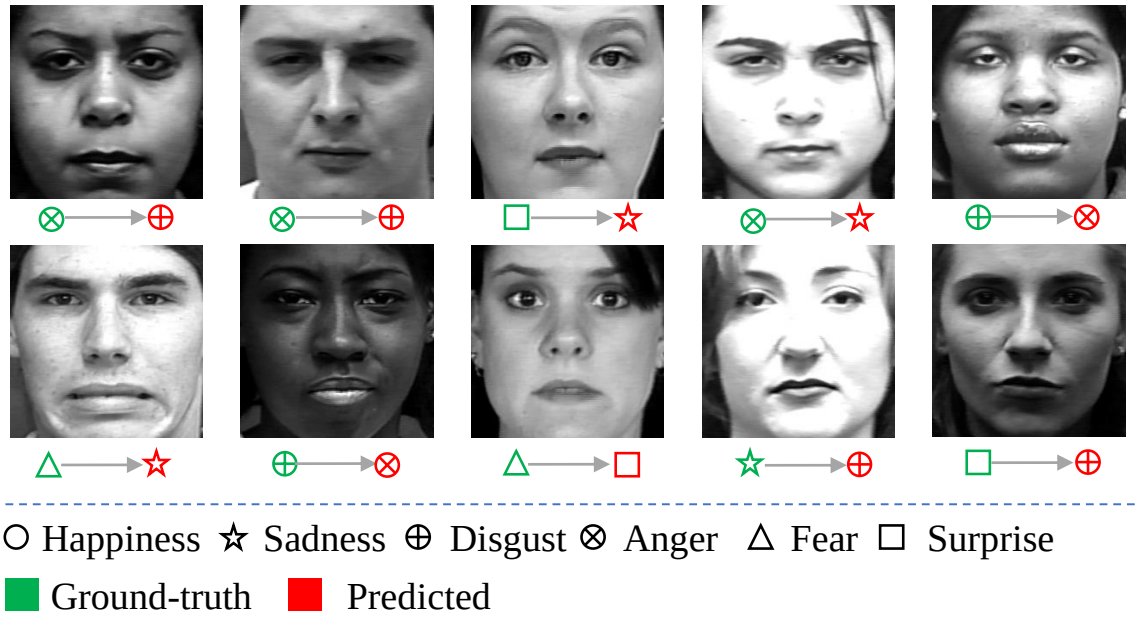

Figure 7. Testing samples in CK+ that are misclassified by our method.

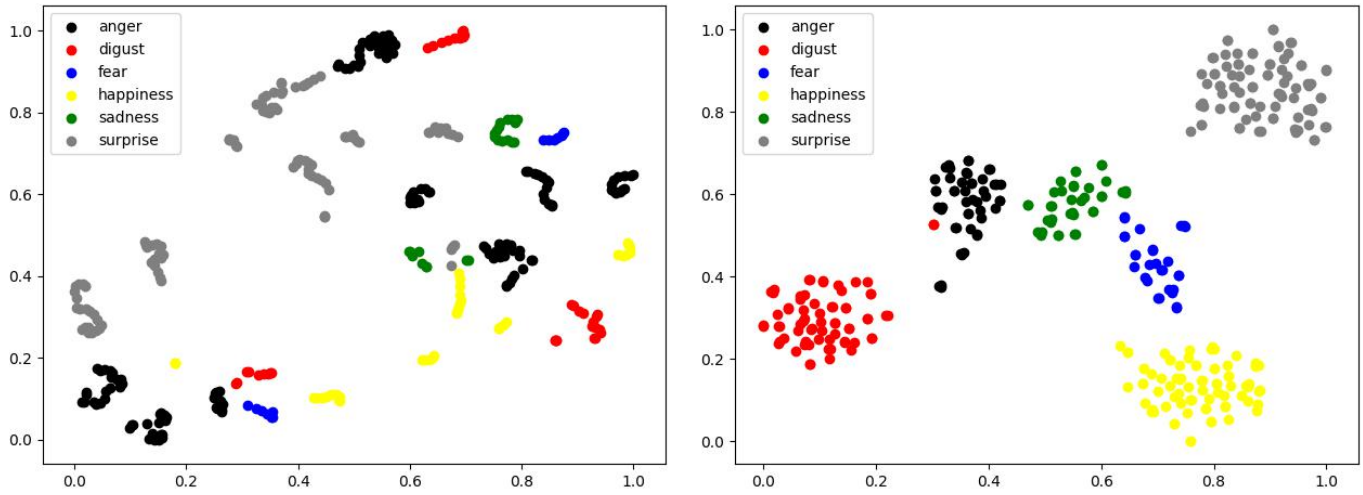

Figure 8. Visualization of deep features output by the VGG-face [8] (**left**) and our method (**right**) on the CK+ dataset. Each dot represents the deep feature extracted from a testing sample.

4.4. Experiments on Oulu-CASIA

Our SD-CNN is also compared with the recent state-of-the-art methods, including FN2EN [18], DeRL [12], GoogLeNet (fine-tuned) [17], VGG-face (fine-tuned) [8], GCNet [40], DynamicMTL [41], MicroExpNet [34], MultiAttention [24], IA-gen [20], PPDN [11], DPND [10], DTAGN(Joint) [29], and CompactCNN [30], on the Oulu-CASIA dataset. As shown in Table 3, our method achieves the highest accuracy of 91.3 %, which outperforms the state-of-the-art static image-based method (i.e., the DynamicMTL [41]) by 1.7% and also suppresses the state-of-the-art sequence-based method (i.e., the CompactCNN [30]) by 2.7%.

Table 3. Comparisons with state-of-the-art methods on the Oulu-CASIA datasets.

Methods	ACC(%)	Image/Sequence
SD-CNN (Ours)	**91.3**	Image
FN2EN [18]	87.7	Image
DeRL [12]	88.0	Image
GoogLeNet (fine-tuned) [17]	79.2	Image
VGG-face (fine-tuned) [8]	72.5	Image
GCNet [40]	86.4	Image
DynamicMTL [41]	89.6	Image
MicroExpNet [34]	85.8	Image
MultiAttention [24]	80.2	Image
IA-gen [20]	88.92	Image
PPDN [11]	84.6	Sequence
DPND [10]	75.3	Sequence
DTAGN(Joint) [29]	81.5	Sequence
LOMO [42]	82.1	Sequence
CompactCNN [30]	**88.6**	Sequence

Figure 9 shows the confusion matrix for the Oulu-CASIA dataset. Among the six facial expressions in this dataset, sadness, happiness, surprise, and fear are recognized with high accuracies of over 90%. Sadness is identified with the highest accuracy of 96.69%, while disgust is recognized with the lowest accuracy of 86.3%. Disgust and anger are two facial expressions that are easily misclassified from each other. For the testing samples with anger, 8.52% of them are misclassified as disgust. Meanwhile, 6.45% of testing samples with disgust are misclassified as anger. Figure 10 illustrates 12 randomly selected testing samples in the Oulu-CASIA dataset that are misclassified by our method. Similar

to the misclassified samples in CK+, most of these misclassified samples are with weak expressions, which are even quite difficult to be classified correctly by human beings.

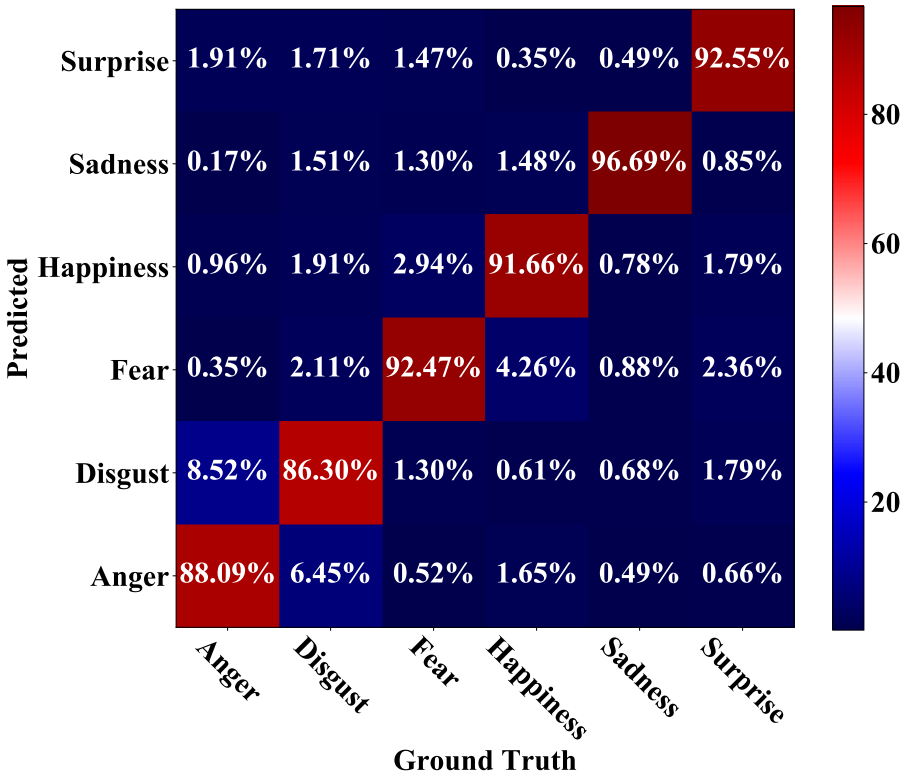

Figure 9. Confusion matrix for the Oulu-CASIA dataset.

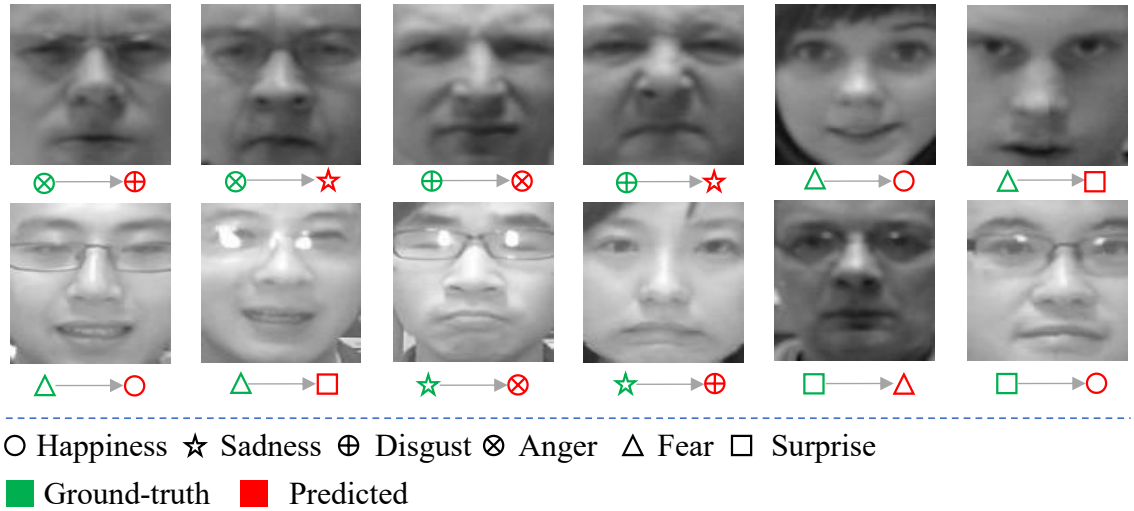

Figure 10. Testing samples in Oulu-CASIA that are misclassified by our method.

In order to evaluate the effect of the facial expression generator, we use only 1 DiffNet to compare the test image with the neutral reference image, and the performance on Oulu-CASIA decreases from 91.3% to 80.4%.

The deep features of all the 654 testing samples in Oulu-CASIA extracted by our method and the VGG-face [8] are visualized using t-SNE [9] in Figure 11. Similar to what happened on the CK+ dataset, the deep features of all the six facial expression classes output by the VGG-face are heavily entangled, while most of the deep features output by our method are clustered according to their facial expression classes.

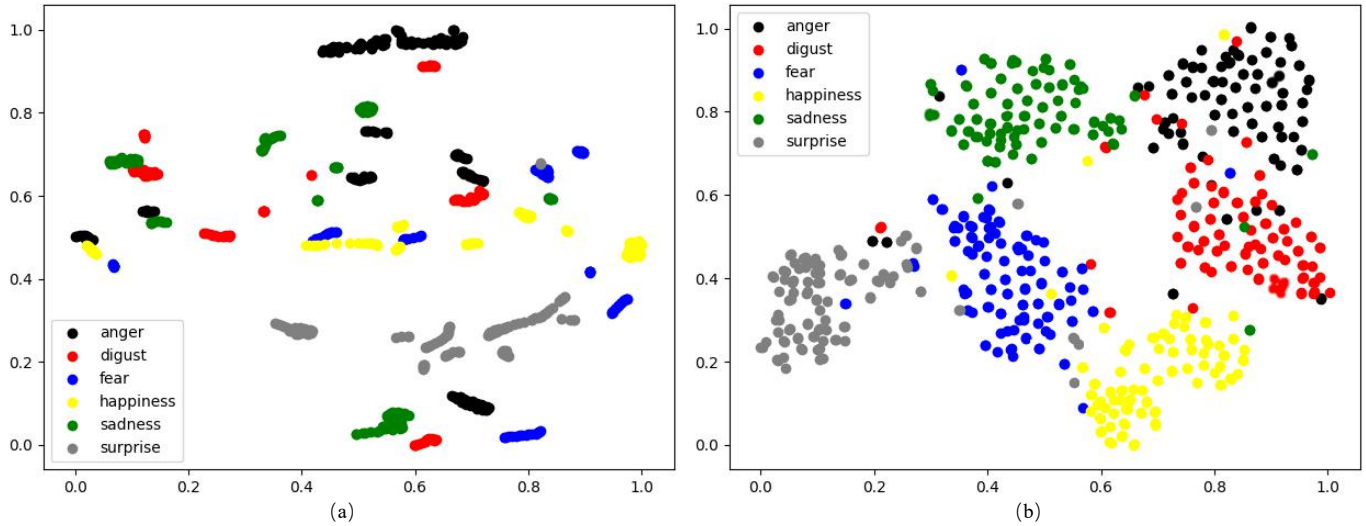

Figure 11. Visualization of deep features output by the VGG-face [8] (**a**) and our method (**b**) on the Oulu-CASIA dataset. Each dot represents the deep feature extracted from a testing sample.

5. Conclusions

In this paper, a self-difference convolutional network (SD-CNN) is proposed for facial expression recognition. First, the SD-CNN uses a conditional generative adversarial network to generate the six typical facial expressions for the same subject in the testing image. Second, six compact and light-weighted difference-based CNNs, called DiffNets, are designed for classifying facial expressions. Each DiffNet extracts a pair of deep features from the testing image and one of the six synthesized expression images and compares the difference between the deep feature pair. In this way, any potential facial expression in the testing image has an opportunity to be compared with the synthesized "Self". As most of the self-difference features of the images with the same facial expression will gather tightly in the feature space, the intra-class variation issue is significantly alleviated. Our SD-CNN has been extensively evaluated on two widely-used FER datasets (i.e., the CK+ and Oulu-CASIA datasets). Experimental results show that our SD-CNN achieves accuracies of 99.7% on CK+ and 91.3% on Oulu-CASIA. Without exploiting the spatial-temporal information, the SD-CNN outperforms the state-of-the-art static image-based methods and even most of the sequence-based methods on both the two datasets. Moreover, the model size of the online part of our SD-CNN is only 9.54 MB, which is much smaller than the recent deep-learning-based FER methods. In future work, we will upgrade our SD-CNN to jointly identify facial expression class and estimate facial expression intensity.

Author Contributions: Conceptualization, L.L. and J.C.; methodology, L.L. and R.J.; software, R.J. and J.H.; validation, R.J., L.L.; writing—original draft preparation, L.L. and R.J.; writing—review and editing, L.L. and J.C.; visualization, J.H. and R.J.; supervision, J.C.; project administration, J.C.; funding acquisition, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (61702208, 62077026, 61937001), and the Fundamental Research Funds for the Central Universities (CCNU19ZN004).

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

References

1. Paul, E.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
2. Chen, L.; Zhou, M.; Su, W.; Wu, M.; She, J.; Hirota, K. Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. *Inf. Sci.* **2018**, *428*, 49–61. [CrossRef]
3. Chen, J.; Luo, N.; Liu, Y.; Liu, L.; Zhang, K.; Kolodziej, J. A hybrid intelligence-aided approach to affect-sensitive e-learning. *Computing* **2016**, *98*, 215–233. [CrossRef]

4. Luo, Z.; Liu, L.; Chen, J.; Liu, Y.; Su, Z. Spontaneous smile recognition for interest detection. In Proceedings of the 2016 7th Chinese Conference on Pattern Recognition (CCPR), Chengdu, China, 3–7 November 2016; pp. 119–130. [CrossRef]
5. Chen, J.; Wang, G.S.; Zhang, K.; Wang, G.H.; Liu, L. A pilot study on evaluating children with autism spectrum disorder using computer games. *Comput. Hum. Behav.* **2019**, *90*, 204–214. [CrossRef]
6. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**. [CrossRef]
7. 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 2010 23th IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (CVPR-Workshops), San Francisco, CA, USA, 13–18 June 2010; pp. 94–101. [CrossRef]
8. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the 2015 26th British Machine Vision Conference (BMVC), Swansea, UK, 7–10 September 2015; pp. 1–12. [CrossRef]
9. Maaten, L.V.D.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605. [CrossRef]
10. Chen, J.; Xu, R.; Liu, L. Deep peak-neutral difference feature for facial expression recognition. *Multimed. Tools Appl.* **2018**, *77*, 29871–29887. [CrossRef]
11. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y. Peak-piloted deep network for facial expression recognition. In Proceedings of the 2016 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 425–442. [CrossRef]
12. Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In Proceedings of the 2018 31th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 19–21 June 2018; pp. 2168–2177. [CrossRef]
13. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the 2017 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July; pp. 1125–1134. [CrossRef]
14. 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]
15. Liu, L.; Gui, W.; Zhang, L.; Chen, J. Real-time pose invariant spontaneous smile detection using conditional random regression forests. *Optik* **2019**, *182*, 637–657. [CrossRef]
16. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the 2016 IEEE Winter Conference on Application of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10. [CrossRef]
17. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1–9. [CrossRef]
18. Ding, H.; Zhou, SK.; 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, Washington, DC, USA, 30 May–3 June 2017; pp. 118–126. [CrossRef]
19. Zhang, F.; Zhang, T.; Mao, Q.; Xu, C. Joint pose and expression modeling for facial expression recognition. In Proceedings of the 2018 31th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 19–21 June 2018; pp. 3359–3368. [CrossRef]
20. Yang, H.; Zhang, Z.; Yin, L. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 294–301. [CrossRef]
21. Fabiano, D.; Canavan, S. Deformable synthesis model for emotion recognition. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–5. [CrossRef]
22. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the 2020 33th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020; pp. 6897–6906. [CrossRef]
23. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef] [PubMed]
24. Gan. Y.; Chen, J.; Yang, Z.; Xu, L. Multiple Attention Network for Facial Expression Recognition. *IEEE Access* **2020**, *8*, 7383–7393. [CrossRef]
25. Zhu, J.; Luo, B.; Zhao, S.; Ying, S.; Zhao, X.; Gao, Y. IExpressNet: Facial Expression Recognition with Incremental Classes. In Proceedings of the 2020 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2899–2908. [CrossRef]
26. Zeng, N.; Zhang, H.; Song, B.; Liu, W.; Li, Y.; Dobaie, A.M. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing* **2018**, *273*, 643–649. [CrossRef]
27. Yu, Z.; Liu, Q.; Liu, G. Deeper cascaded peak-piloted network for weak expression recognition. *Vis. Comput.* **2018**, *34*, 1691–1699. [CrossRef]

28. Chen, Y.; Wang, J.; Chen, S.; Shi, Z.; Cai, J. Facial motion prior networks for facial expression recognition. In Proceedings of the 2019 33th IEEE Visual Communications and Image Processing (VCIP), Sydney, Australia, 1–4 December 2019; pp. 1–4. [CrossRef]
29. 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 15th IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 2983–2991. [CrossRef]
30. Kuo, CM.; Lai, SH.; Sarkis, M. A compact deep learning model for robust facial expression recognition. In Proceedings of the 2018 31th IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-Workshops), Salt Lake City, UT, USA, 19–21 June 2018; pp. 2121–2129. [CrossRef]
31. Meng, D.; Peng, X.; Wang, K.; Qiao, Y. Frame attention networks for facial expression recognition in videos. In Proceedings of the 2019 26th IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3866–3870. [CrossRef]
32. Pumarola, A.; Agudo, A.; Martinez, A.M.; Sanfeliu, A.; Moreno-Noguer, F. Ganimation: Anatomically-aware facial animation from a single image. In Proceedings of the 2018 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 818–833. [CrossRef]
33. Ekman, R. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: Oxford, UK, 1997. [CrossRef]
34. Cugu, I.; Sener, E.; Akbas, E. MicroExpNet: An Extremely Small and Fast Model For Expression Recognition From Face Images. In Proceedings of the 2019 9th International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 6–9 November 2019; pp. 1–6. [CrossRef]
35. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein gans. *arXiv* **2017**, arXiv:1704.00028.
36. Justin, J.; Alexandre, A.; Li, F. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the 2016 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 694–711. [CrossRef]
37. Hermans, A.; Beyer, L.; Leibe B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
38. Fabian Benitez-Quiroz, C.; Srinivasan, R.; Martinez, A.M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the 2016 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5562–5570. [CrossRef]
39. FiveCrop. Available online: https://pytorch.org/docs/stable/torchvision/transforms.html (accessed on 23 March 2021).
40. Kim, Y.; Yoo, B.; Kwak, Y.; Choi, C.; Kim, J. Deep generative-contrastive networks for facial expression recognition. *arXiv* **2017**, arXiv:1703.07140.
41. Ming, Z.; Xia, J.; Luqman, MM.; Burie, JC.; Zhao, K. Dynamic Multi-Task Learning for Face Recognition with Facial Expression. *arXiv* **2019**, arXiv:1911.03281.
42. Sikka, K.; Sharma, G.; Bartlett, M. Lomo: Latent ordinal model for facial analysis in videos. In Proceedings of the 2016 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5580–5589. [CrossRef]