electronics

MDPI

Article

Facial Pose and Expression Transfer Based on Classification Features

Zhiyi Cao [1], Lei Shi [2], Wei Wang [1,*] and Shaozhang Niu [3]

[1] College of Computer and Cyberspace Security, Hebei Normal University, Shijiazhuang 050025, China
[2] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
[3] Beijing Key Lab of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China
* Correspondence: wangwei2021@hebtu.edu.cn

Abstract: Transferring facial pose and expression features from one face to another is a challenging problem and an interesting topic in pattern recognition, but is one of great importance with many applications. However, existing models usually learn to transfer pose and expression features with classification labels, which cannot hold all the differences in shape and size between conditional faces and source faces. To solve this problem, we propose a generative adversarial network model based on classification features for facial pose and facial expression transfer. We constructed a two-stage classifier to capture the high-dimensional classification features for each face first. Then, the proposed generation model attempts to transfer pose and expression features with classification features. In addition, we successfully combined two cost functions with different convergence speeds to learn pose and expression features. Compared to state-of-the-art models, the proposed model achieved leading scores for facial pose and expression transfer on two datasets.

Keywords: facial pose and expression transfer; generation model; classification features; generative adversarial networks

check for updates

Citation: Cao, Z.; Shi, L.; Wang, W.; Niu, S. Facial Pose and Expression Transfer Based on Classification Features. *Electronics* **2023**, *12*, 1756. https://doi.org/10.3390/electronics12081756

Academic Editor: Savvas A. Chatzichristofis

Received: 11 March 2023
Revised: 31 March 2023
Accepted: 3 April 2023
Published: 7 April 2023

1. Introduction

Transferring facial pose and expression features from one conditional face to another source face is an interesting but challenging task in pattern recognition. It is widely used for the entertainment audience and video production industry [1]. Through this technology, it has become a reality that the face of a virtual digital human can be rotated to any angle and can make any facial expression [1]. However, the task has the following challenges: (1) it is difficult to learn from some conditional face mappings with low-dimensional classification labels; (2) it is difficult to control the transfer order for facial pose and expression transfer.

Recently, CR-GAN [1] built a two-path generative adversarial network to learn complete pose representations for most poses rather than incomplete pose representations for some poses. For facial expression transfer, previous solutions [2,3] were trained to learn some expression representations with data-driven generation. For facial pose and expression transfer, models such as [4] and Few-ShotFace [5] have been trained to learn most pose and expression representations through few-shot learning.

While existing models have achieved impressive results, there are two challenges that need to be addressed urgently. For the CR-GAN model and the Few-ShotFace model, they select classification labels as a condition for their generator networks to transfer pose and expression features. The first challenge is that low-dimensional classification labels cannot hold the differences in shape and size between conditional and source faces. First, it cannot hold the shape and size differences between conditional faces and source faces that are in the classification boundary. Second, it cannot hold the differences in shape and size between conditional and source faces, which account for the majority of the classification.

For example, if the conditional face is small and the source face is large, existing models will produce unreasonable outputs. The second challenge is that it is difficult to control the transfer order for facial pose and expression transfer. As we know, facial pose features are generally much harder to learn than facial expression features. If the order of transfer learning is not well controlled, it will affect the transfer effect and inevitably increase the convergence time of the model.

Zhang et al. [6] proposed a novel deep transfer neural network method based on multi-label learning (MNet) for facial attribute classification. Sankaran et al. [7] introduced a domain adaptive representation (DAR) learning method for facial action unit recognition. Bozorgtabar et al. [8] proposed a novel adversarial domain adaptation (ADA) for facial expression analysis. Tweaked Convolutional Neural Networks (TCNN) have shown that features extracted from deeper layers capture rough landmark locations. For the first challenge, inspired by TCNN, MNet, DAR, and ADA, we propose using high-dimensional classification features rather than classification labels as a condition for the proposed generator network. A novel two-stage classifier is presented here to capture normal classification features from deeper layers. Compared to classification labels, classification features have a higher dimension and contain more normal facial features. With this improvement, virtual digital humans in the entertainment audience and video production industry can appear smoother and produce more realistic visualizations of humans and their faces. Subsequently, the average features class conditional face, the average classification features of the source face, and the average classification characteristics of the source face are used to estimate the classification characteristics of the target face for the proposed generator network to generate the target face. Through this technology, after virtual digital humans change conditions, the results generated by the model can still be regarded as the same visually reasonable digital human (shown in Figure 1b). Such research has not yet been widely carried out, so this has inspired our research, which is described in the next paragraph.

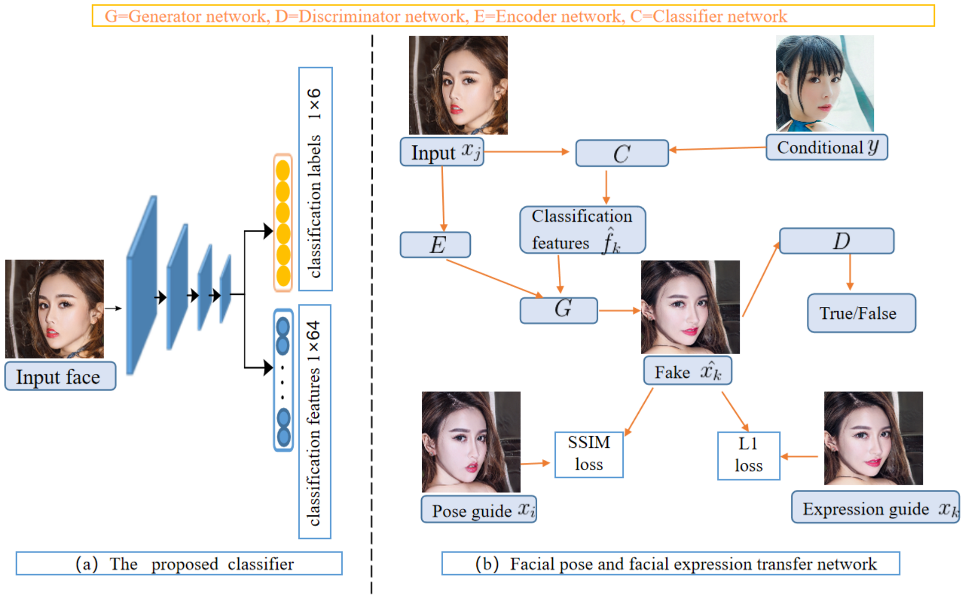
G=Generator network, D=Discriminator network, E=Encoder network, C=Classifier network
Input face
classification labels 1×6
classification features 1×64
(a) The proposed classifier
Input x_j
C
Conditional y
E
Classification features $\hat{f}_k$
D
G
True/False
Fake $\hat{x}_k$
SSIM loss
L1 loss
Pose guide x_i
Expression guide x_k
(b) Facial pose and facial expression transfer network

Figure 1. The proposed network architecture for facial pose and expression transfer. (**a**) The proposed classifier network *C*. (**b**) The proposed global network architecture.

To summarize, this paper presents a generative adversarial network model based on classification features (short for CF-GAN) for facial pose and facial expression transfer. Our research fills a gap in the development of more accurate virtual digital human face synthesis. The main contributions of the proposed method are as follows. First, we present a novel two-stage classifier to capture classification features for each face. In addition, it is the first time that classification features can hold the differences in shape and size between conditional and source faces. Finally, we successfully combined two cost functions with different convergence speeds to learn pose and expression features.

2. Related Work

In recent years, facial pose and expression transfer tasks have been extensively studied. This paper focuses on generative adversarial networks (GANs) [9]-related research. According to the transfer difficulty level, we divided the existing models into two classifications, facial pose transfer, facial pose, and expression transfer.

2.1. Facial Pose Transfer

For facial pose transfer, it aims to transfer a source face from one viewpoint to another that is equal to the angle of one given conditional face. As we know, StarGAN [10] is the first model for one-to-many transfers. StarGAN introduces an auxiliary classifier for a single discriminator network (short for D) to control one-to-many transfers. BiGAN [11] presents a generator network (short for G) with an encoder (short for E) for post-transfer. Then other models such as [12–14] introduce a single-path model for pose transfer: an encoder network and a generator network are followed by a discriminator network. DR-GAN [12] proposes an identity-preserved representation for pose transfer. TP-GAN [15] presents a novel two-path model to capture global features and local details for post-transmission. For the above models that have achieved unreasonable results for some poses with incomplete representations, CR-GAN [1] introduces a two-path architecture to learn complete representations for all poses from the dataset. For the CR-GAN model, if the angle of the source face is different from most faces in the existing dataset, it will generate an unreasonable output. Ramamoorthi et al. [16] presented a method to analytically construct th principal components for images of a Lambertian object from a single viewpoint. Lee et al. [17] showed that configurations of single light source directions exist that are effective for face recognition. Kent et al. [18] used principal component analysis to find the most representative spatial daylight distribution patterns. Savelonas et al. [19] provided an overview of computational methods aiding geoscientists in the analysis of 2D or 3D imaging data. Inspired by these models, we introduced a classifier to classify similar facial pose features as a classification. We aimed to learn limited and effective mapping rather than endless mapping.

2.2. Facial Pose and Expression Transfer

For facial pose and expression transfer, it aims to transfer facial pose and some expressions such as ‘happy’, and ‘angry’ from one conditional face to another. In previous pioneer works, Wang et al. [20] extended a GANs framework to interactive visual manipulation with two additional features. Hosoi et al. [21] proposed a status score to transfer both head posture and facial expression. Pumarola et al. [22] introduced a novel GAN conditioning scheme based on Action Units (AU) annotations. Zhu et al. [23] presented a GANs model to translate an image from a source domain to a target domain in the absence of paired examples. Chang et al. [2] introduced a generative adversarial network for automatic cycle-consistent photo editing. Chen et al. [3] proposed a model for digital face manipulation based on an end-to-end convolutional neural network. Wu et al. [24] presented a novel learning-based framework with a latent boundary space for face reenactment. Wang et al. [25]’s model separated the constraints for intrinsic subject-specific characteristics and age-specific facial changes with respect to elapsed time. These models are trained to learn expression representations for expression transfer and typically introduce an encoder network to obtain expression representations and a generator network to generate expression transfer results. For facial pose and expression transfer, X2face [26] introduces a self-supervised network architecture that allows the pose and expression of a given face to be controlled by another face. Following that, other models such as OneShotFace [4], and Funit [27] were trained with meta-learning architecture. The most representative model is Few-ShotFace [5]. This model uses an embedded network to obtain pose and expression representations and a generator network to generate transfer results from facial landmarks. In [28], they present a novel framework, Generative Priorguided UNsupervised Image-to-image Translation (GPUNIT), to improve overall quality and applicability. Similarly, MSPC [29] proposes a universal regularization technique called maximum spatial perturbation consistency, which enforces a spatial perturbation function. Despite the impressive results of these models, there are still some challenges [30]. For the Few-ShotFace model, if the source face and conditional face have larger differences in shape and size, they will produce unreasonable output.

3. Proposed Method

In this section, we first introduced the proposed classifier. Following that, we introduced the proposed generator. In addition, we introduced network architecture. Then, the CF-GAN model objective was provided. Finally, we introduce the proposed algorithm in the form of pseudocode.

3.1. Proposed Formula for Classifier

Given a source face x, previous researchers typically built a classifier C to output the classification labels c_o by: $C(x) \longrightarrow c_o$. However, this is just an ideal state. If the classifier C is introduced into the CR-GAN model, the correct transfer result will not be obtained because the low-dimensional classification labels cannot hold the shape and size differences between conditional and source faces.

The problem of classification labels can be analyzed from two aspects. For example, a face with an angle of 30°, which is the boundary angle of two classifications, may be classified into either of the two classifications. Following that, the wrong classification labels will lead to the wrong viewpoint guide face and the expression guide face as shown in Figure 1b. Finally, existing models will produce the wrong facial pose and expression transfer result. If most faces with strange shapes in the classification have larger or smaller sizes, even if the correct classification labels are obtained, they only learned the wrong mapping between them rather than the correct mapping between faces with normal shapes and sizes.

For the classification label problem, this paper trained a two-stage C classifier with K classifications (K = 18 in our experiments). In the first stage, we selected J (J = 10 in our experiments) positive faces that have normal shapes and sizes and are at the center of each classification to train the model. Since the selected faces do not have the above problems, the model at this stage is easy to train, but it is not suitable for the facial pose and expression transfer with all the angles. Given a source face x_f, the classifier C_f of the first stage will output the classification features f_f and the classification labels c_f.

$$f_f, c_f = C_f(x_f) \tag{1}$$

where c_f is a 6-dimensional vector and f_f is a 64-dimensional vector. For classification features $f_{k_1}, \ldots, f_{k_n}$ from the classification k, the average classification features $\overline{f_{f_k}}$ can be expressed as:

$$\overline{f_{f_k}} = \frac{\sum\limits_{i=1}^{n} f_{ki}}{n} \tag{2}$$

Using the average of K classifiers, we can obtain the general features of all faces in the classification. Note that we saved all the average classification features for each classification. Then, we denoted $CE(\cdot,\cdot)$ as the cross-entropy loss. In the first stage, the cross-entropy loss was used to calculate the difference between the output classification labels c_f and the real classification labels c:

$$\mathcal{L}_f = CE\left(c_f, c\right) \tag{3}$$

In the second stage, we aimed to model from all angles. Given a source face x, the classifier C of the second stage will output the classification features f, and the classification labels c_o.

$$f, c_o = C(x) \tag{4}$$

The cross-entropy loss was employed to calculate the difference between the output classification labels c_o and the real classification labels c firstly. Here, we denote $CS(\cdot,\cdot)$ as the cosine similarity loss. Following that, $CS(\cdot,\cdot)$ was used to calculate the difference between the output classification features f and the average classification features $\overline{f_{f_k}}$. The overall cost of the second stage is as follows:

$$\mathcal{L}_s = CE(c_o, c) + CS\left(f, \overline{f_{f_k}}\right) \tag{5}$$

In order to eliminate the wrong mapping caused by faces with strange shapes or normal sizes, we control their classification characteristics to train the second-stage classifier. In detail, for faces whose similarity between the classification features and the average classification features in any classification is less than 0.7, we replaced their classification features with the average classification features from the first stage. This process can be expressed as follows:

$$f_{k_i} = \begin{cases} f, CS\left(f, \overline{f_{f_k}}\right) >= 0.7 \\ \overline{f_{f_k}}, CS\left(f, \overline{f_{f_k}}\right) < 0.7 \end{cases} \tag{6}$$

Following that, the average classification features for the K classification of the second stage can be expressed as:

$$\overline{f_k} = \frac{\sum\limits_{i=1}^{n} f_{ki}}{n} \tag{7}$$

where n represents the number of faces in each classification.

In this way, the second stage classifier can avoid the problems caused by classification labels to build the right mapping for facial pose and expression transfer. The average classification features of the first stage can be used as a powerful condition to improve the classification ability of the second-stage model. For boundary faces, even if the classification labels are wrong, the correct classification features can be obtained with the help of Equation (6). For faces with larger or smaller shapes and sizes, Equation (6) can provide normal classification features. Accordingly, the proposed classification features can hold the differences in shape and size between conditional and source faces in most conditions. Here, we save all the average classification features for each classification. Note that when training the global CF-GAN model, only the second stage classifier model is used to obtain classification features and classification labels.

3.2. Proposed Formula for Generator

The StarGAN model introduces the classification label c for facial expression transfer: $G(x, c) \longrightarrow y$. Next, the CR-GAN model introduces an encoder E to optimize the source faces: $G(E(x), c) \longrightarrow y$. Although it shows some progress in facial pose transfer, the training process of the two-path model is very time-consuming. Since there is no clear relationship between E and c, the CR-GAN model is difficult to converge.

In contrast to the CR-GAN model, we propose usimg the classification features f for facial expression and facial pose transfer: $G(E(x), f) \longrightarrow y$. In this paper, we focus on condition-based facial expression and facial pose transfer: $G(E(x_j), f_k, y) \longrightarrow x_k$. For instance, a source face x_{j_1}, the classification features f_{k_2} from the target face x_{k_2} and the condition face y are used to generate the target face x_{k_2}. Note that x_{k_2} should look like the same person as x_{j_1} and have the same pose and expression as y.

When training, we produced the classification features f_{k_2} from target face x_{k_2}, but when we tested the model f_{k_2} and x_{k_2} are unknown. Therefore, it is necessary to estimate the classification features f_{k_2} based on condition face y and source face x_{j_1}.

Based on Equation (4), we can get the classification features f_{j_1} and their classification labels. Following that, we achieved the average classification features $\overline{f_j}$ by its classification labels from the saved average classification features. Similarly, the average classification features $\overline{f_k}$ for condition face y can be obtained. Depending on the condition face y to reason about the target faces x_{k_2}, it is clear that they come from the same classification and have the same average classification features $\overline{f_k}$.

Based on Equation (5), the following conclusions can be drawn: the classification features f_{j_1} are close to the average classification features $\overline{f_j}$: $f_{j_1} \approx \overline{f_j}$ and the classification features f_{k_2} are close to the average classification features $\overline{f_k}$: $f_{k_2} \approx \overline{f_k}$. For unknown classification features f_{k_2}, the nearest value is evaluated by the following formula:

$$f_{k_2} \approx f_{j_1} - (\overline{f_j} - \overline{f_k}) \tag{8}$$

Obviously, the obtained classification features f_{k_2} are only related to the classification of the condition face y, but not related to its shape and size. In addition, the classification features f_{k_2} are based on source face x_{j_1}, so it has the deep features of source face x_{j_1}. In this way, the results generated by the model and the source face x_{j_1} have the same characteristics of the same person. y is used to obtain approximate classification features f_k. At last, $G(E(x_j), f_k, y) \longrightarrow x_k$ can be rewritten as:

$$G(E(x_j), f_{k_2}) \longrightarrow x_k \tag{9}$$

3.3. Proposed Network Architecture

As shown in Figure 1b, CF-GAN consists of four modules: the discriminator network (D), the classifier network (C), the generator network (G), and the encoder network (E).

As shown in Figure 1a, the proposed classifier network comprises six convolutional layers. Among them, the fifth convolutional layer outputs classification labels, and the sixth convolutional layer outputs classification features.

The discriminator network (D) learns to make a distinction between the result of the generator network and the real face first and outputs classification features for the input faces. The discriminator network applies PatchGANs [31], which consists of six convolutional layers for downsampling. In addition, the last two convolutional layers are used to obtain the distinction (0 or 1) and classification features.

The encoder network (E) learns to capture the style feature of each face. When we transfer x_j to x_k, it is should keep that they can be recognized as the same person. This is achieved by: $E(x_j) \approx E(x_k)$. The encoder network consists of three convolutional layers, followed by three residual blocks to downsample the source faces. In addition, all convolutional layers are followed by Instance Normalization and ReLU units.

The proposed generator network (G) learns to obtain facial expression and facial pose transfer results. For G, we only enter $E(x_j)$ and f_k. Following that, we used a full convolutional layer to combine them. Next, we used three residual blocks and two deconvolution layers to upsample the features. The last convolutional layer was used to hold the obtained features that are the same size as the source faces. In addition, all convolutional layers were followed by Instance Normalization and ReLU units.

3.4. Overall Objective

We trained the proposed CF-GAN model by solving a minimax optimization problem for the generator network G and the discriminator network D:

$$\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda_T \mathcal{L}_T(G) + \lambda_R \mathcal{L}_R(G) + \lambda_C \mathcal{L}_C(G) \tag{10}$$

where $\mathcal{L}_{GAN}$, $\mathcal{L}_T$, $\mathcal{L}_R$, and $\mathcal{L}_{CR}$ are the GAN cost, the face transfer cost, the face reconstruction cost, and the classification features cost. The GAN cost can be expressed as:

$$\mathcal{L}_{GAN}(G,D) = \mathbb{E}_{x_j \sim p_{data}(x_j)}\left[\log D_{X_j}(x_j)\right] + \mathbb{E}_{x_k \sim p_{data}(x_k)}\left[\log(1 - D_{X_k}(G)\right] \tag{11}$$

The GAN cost ensures that the discriminator D is not fooled by the generated result of the generator G. To control facial expression and facial pose transfer, transfer cost was introduced. It will be easier to converge by shifting the face angle first and transferring the facial expression later.

In general, it is more difficult to learn facial pose features than to learn facial expression features. For example, when the angle of the conditional face is greater than 60° or less than $-60°$, the transfer result usually looks unnatural because the correct facial pose features cannot be obtained. If the order of transfer learning is not well controlled, it will affect the transfer effect.

In previous papers [32], it was proved that SSIM (the structural similarities) loss converges faster than L1 (Mean Absolute Error) loss for face generation tasks. Inspired by it, we introduced an improved version of SSIM loss (short for s) to control the learning of the face angle and use the L1 loss to control the learning of facial expressions. Specifically, if x_j tries to learn x_k's facial expression and facial pose, we introduced x_i to achieve it. What needs to be guaranteed is that a pose difference will exist between x_j and x_i and a facial expression difference between x_j and x_k.

In order to show that the improved version of SSIM loss converges faster than L1 loss, we devoted the mathematical difference between x_j and x_i as q. We first defined the formula of L1 loss:

$$\mathrm{y}^{\mathrm{L1}} = \mathrm{q} \tag{12}$$

Then we can get the formula of improved SSIM loss:

$$\mathrm{y}^{\mathrm{SSIM}} = 2\mathrm{q} - \mathrm{q}^2 \tag{13}$$

Obviously, it can be seen from the above formula that when the value of q changes from 0 and 1, the value of SSIM loss under any condition is greater than the value of L1 loss. This means that SSIM losses can discover more differences in the pose so that the model can converge faster. On the change curve, it can also be seen that the improved SSIM loss converges faster than the L1 loss shown in Figure 2.

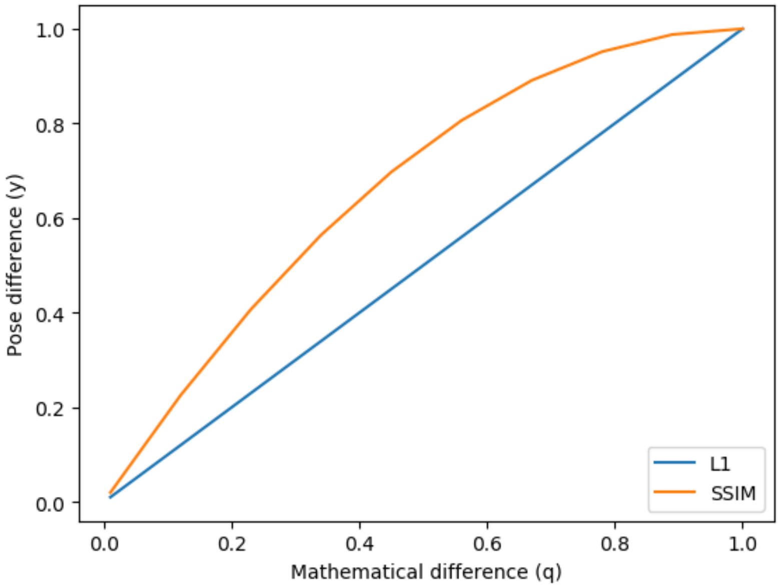

Figure 2. Under the same mathematical difference condition, comparing the pose difference of L1 loss, SSIM loss.

To control the order of transfer learning, the transfer loss can be expressed as:

$$\mathcal{L}_T(G) = ffE_{x_j\sim p_{data}(x_j)}[||G - x_i||_{ssim}] + (1-ff)E_{x_j\sim p_{data}(x_j)}[||G - x_k||_{l1}] \tag{14}$$

where ff gradually changes from infinitely close to 1 to infinitely close to 0 during training. For transfer loss and GAN loss, our G uses the false classification features f_{k_2} from Equation (3) to generate the correct faces. In addition, we need to make sure that G can use the true classification features f to generate the right faces. To achieve this, we used face reconstruction:

$$\mathcal{L}_R(G) =\mathbb{E}_{x_k\sim p_{data}(x_k)}[||G - x_k\|_1] \tag{15}$$

Although we can already generate faces $\hat{x_k}$ that are closer to the target faces x_k, we still need to ensure that the classification features of the generated faces are close to the classification features of the target faces. Our classification features of the generated faces implement from the discriminator D: $\hat{f} = D(\hat{x_k})$. Our classification features of the target faces achieve from the classifier C: $f = C(x_k)$. Finally, the loss of classification features can be expressed as:

$$\mathcal{L}_C(G) =\mathbb{E}_{x_j\sim p_{data}(x_j)}\left[\left\|f - \hat{f}\right\|_1\right] \tag{16}$$

3.5. Our Algorithm

In this section, we will introduce our algorithm in the form of pseudocode. The whole process is shown in Algorithm 1.

Algorithm 1 Facial pose and happy expression transfer algorithm.

```
Require:
    The set of source faces and conditional faces $y$ (128 × 128) for current batch, $x_j$ (128 × 128);
    The trained classifier network, $C$;
    The fixed average classification features from different classifications, $F$;
Ensure:
 1: We input conditional faces $y$ for the classifier network $C$ to get classification features $f_y$ and classification labels $c_y$;
 2: We use the classification labels $c_y$ from conditional faces $y$ to get the pose guide faces $x_i$ and the expression guide faces $x_k$ for current batch;
 3: We input source faces $x_j$ for the classifier network $C$ to get classification features $f_j$ and classification labels $c_x$;
 4: The classification labels $c_y$ and classification labels $c_x$ is used to get the average classification features $f_a$;
 5: The classification features $f_j$ and the average classification features $f_a$ is used to get the unknown classification features $\hat{f}_k$ followed Equation (8);
 6: The style feature $E(x_j)$ and unknown classification features $\hat{f}_k$ are used to get fake faces $\hat{x_k}$ from the generator network $G$;
 7: The discriminator network $D$ is used to distinguish between real and fake faces and classify the real faces to its corresponding classification;
 8: Equation (11) is used to calculate the GAN loss for the generator network $G$ and the discriminator network $D$;
 9: Equation (12) is used to calculate the transfer loss for for the generator network $G$;
10: Equation (13) is used to calculate the face reconstruction loss for the generator network $G$;
11: Equation (14) is used to calculate the classification features loss for the generator network $G$;
12: Update the gradient values of the discriminator network, the generator network, and the encoder network.
13: Repeat 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 about 20,000 times;
14: return the trained generator network $G$;
```

4. Experiments

To explore the generality of the CF-GAN model, we trained and tested the model on a variety of tasks. Experimental comparisons with the most advanced models were conducted to accomplish the same tasks.

4.1. Datasets and Preprocessing

We utilized two datasets for testing and comparison. For Multi-PIE [33] dataset, we first divided it into nine classifications according to angles from -60° to 60°, and then according to whether the mouths of different angles are open or not, we expanded it into 18 classifications (The opening mouth is used to express angry expression). The face angle between each of the first nine classifications and the last nine classifications is the same, but their expressions are different. In order to test our model at more angles, we built a Chinese female multi-angle dataset. We used Stylegan [34] to generate 5000 Chinese female face faces. Following that, we used Stylegan [34] (https://github.com/a312863063/seeprettyface-face_editor (accessed on 10 March 2023)) to generate some missed faces from -60° to 60° with happiness or not for 5000 female faces. We used the face detection algorithm [35] to verify the reliability of the angle and expression of the generated face. Similar to the above, we also divided the faces into 18 classifications. Before starting the experiment, we resized the faces to 128 × 128. For CF-GAN, we used Adam with $\beta1$ = 0.5 and $\beta2$ = 0.999. Our batch size was set to 6 and we used a learning rate of 0.0001. The parameter values we used are $\lambda_T = 10$, $\lambda_R = 10$, $\lambda_C = 20$. The training takes two days on a single GTX1080Ti GPU for about 100 epochs. We first trained the classifier network (C), then trained the discriminator network (D) ten times before training the encoder network (E) and the generator network (G) once.

4.2. Baselines

To compare the performance of CF-GAN model, we adopted the TP-GAN [15] model, CR-GAN [1] model, X2face [26] model, and the Few-ShotFace [5] model as our baseline models.

- **TP-GAN** uses two pathways to capture global features and local features for pose transfer. Since the TP-GAN model cannot generate faces based on the condition face y, in order to compare fairly with it, we obtained classification labels when testing.
- **CR-GAN** introduces a two-path architecture to learn complete representations for all poses. For the CR-GAN model, we also obtained classification labels when testing.
- **X2face** introduces a self-supervised network architecture that allows the pose and expression of a given face to be controlled by another face.
- **Few-ShotFace** (short for FSface) introduces a meta-learning architecture, which involves an embedded network and a generator network. For a fair comparison, we only compared the results according to a single source face with the FSface model.

4.3. Evaluation Index

Using the same evaluation metrics, we compared our method against several baselines qualitatively and quantitatively.

- **AMT**. For these tasks, we ran “real vs fake” perceptual studies on Amazon Mechanical Turk (AMT) to assess the realism of our outputs. We followed the same perceptual study protocol from Isola et al. [31], and we gathered data from 50 participants per algorithm we tested. Participants were shown a sequence of pairs of faces, one real face and one fake (generated by our algorithm or a baseline), and were asked to click on the real face to be considered.
- **Classification (Cf for short)**. We trained Xception Version 3 [36]-based binary classifiers for each face dataset. The baseline is classification accuracy in real faces. Higher classification accuracy means that the transferred faces may be easier to distinguish.
- **Consistency (Cs for short)**. We compared domain consistency between real faces and transferred faces by calculating average distance in feature space. Cosine similarity was used to evaluate the perceptual distance in the feature space of the VGG-16

network [37] that pre-trained in faceNet [38]. We calculated the average difference of the five convolution layers preceding the pool layers. The larger average value will lead to a smaller cosine similarity value, meaning the few similarities between the two faces. In the test stage, we randomly sampled the real face and the transferred face from the same person to make up data pairs and compute the average cosine similarity distance between each pair.

- **Convergence time (TIME for short)**. For these tasks, we recorded the time required to reach the final state of convergence under the same data set and GPU conditions for different models for comparison. The recorded results can reflect the time-consuming model training. The unit of TIME is an hour.

4.4. Experimental Results

The experimental results of the CF-GAN model on MultiPIE [33] dataset and the Chinese female multi-angle dataset are shown in Figure 3. For facial pose and anger expression transfer (the opening mouth is used to express angry expression), the experiment results are shown in the first six columns. In Figure 3, it is observed that CF-GAN is able to generate faces followed by conditional faces. For instance, generated faces match the pose classification and the opening mouth classification of conditional faces. For facial pose and happy expression transfer, the experiment results are shown in the last six columns. It is also shown that the generated faces match the pose classification of conditional faces and the happy classification of conditional faces.

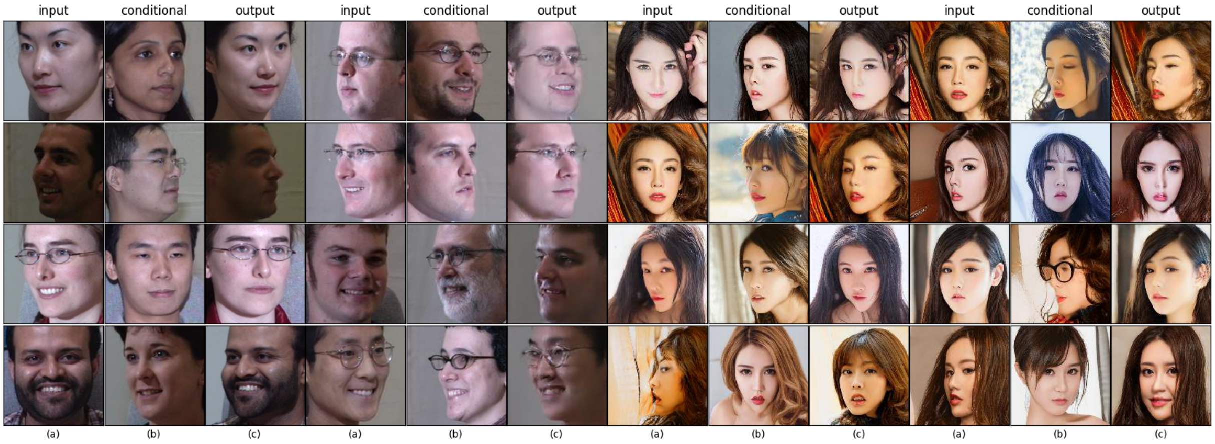

Figure 3. Experimental results from two datasets under source face and conditional face have larger differences in shape and size. (**a**) Source faces (128 × 128 pixels). (**b**) Conditional faces. (**c**) Transfer results by CF-GAN.

4.5. Base Model Comparison

Here, we evaluated the performance of four different models. In Figure 4, it is shown that the proposed model generates more reasonable faces than other models, and the generated facial pose and happy expression are close to conditional faces. The TP-GAN model and the X2face model can handle some face pose and expression transfer tasks when the source faces are similar to conditional faces. The CR-GAN model and the FSface model are able to handle most facial pose and expression transfer tasks when the source faces have smaller differences in shape and size with conditional faces. The CF-GAN model enables us to complete most facial pose and expression transfer tasks when the source faces have a larger shape and size differences under conditional faces.

Experimental results also show that for facial pose and anger expression transfer tasks, all models produce natural faces similar to the source face and have the same facial pose and expression as conditional faces. However, for facial pose and happy expression transfer tasks, the CF-GAN model produces more natural faces shown in the right of Figure 4. After extensive analysis, we found that for any classification in Multi-PIE [33] dataset, the angles of most faces are fixed. For instance, the angles under a certain classification only have 45° faces.

For facial pose and anger expression transfer, all models only learn $K(K-1)$ mappings for K classifications (K = 18 in our experiments). It should be noted that for any classification in the Chinese female multi-angle dataset, the angles of most faces are not fixed. For instance, angles under a certain classification can fluctuate from 45 to 60°. For facial pose and happy expression transfer, although CF-GAN only learns $K(K-1)$ mappings, other models need to learn $J(J-1) \times K(K-1)$ (J means angles and $J \geq 120$) mappings. These demonstrate that CF-GAN only learns a limited number of maps, while other models learn more maps. If the differences in shape and size between conditional and source faces are taken into account, there are endless maps that need to be established by them.

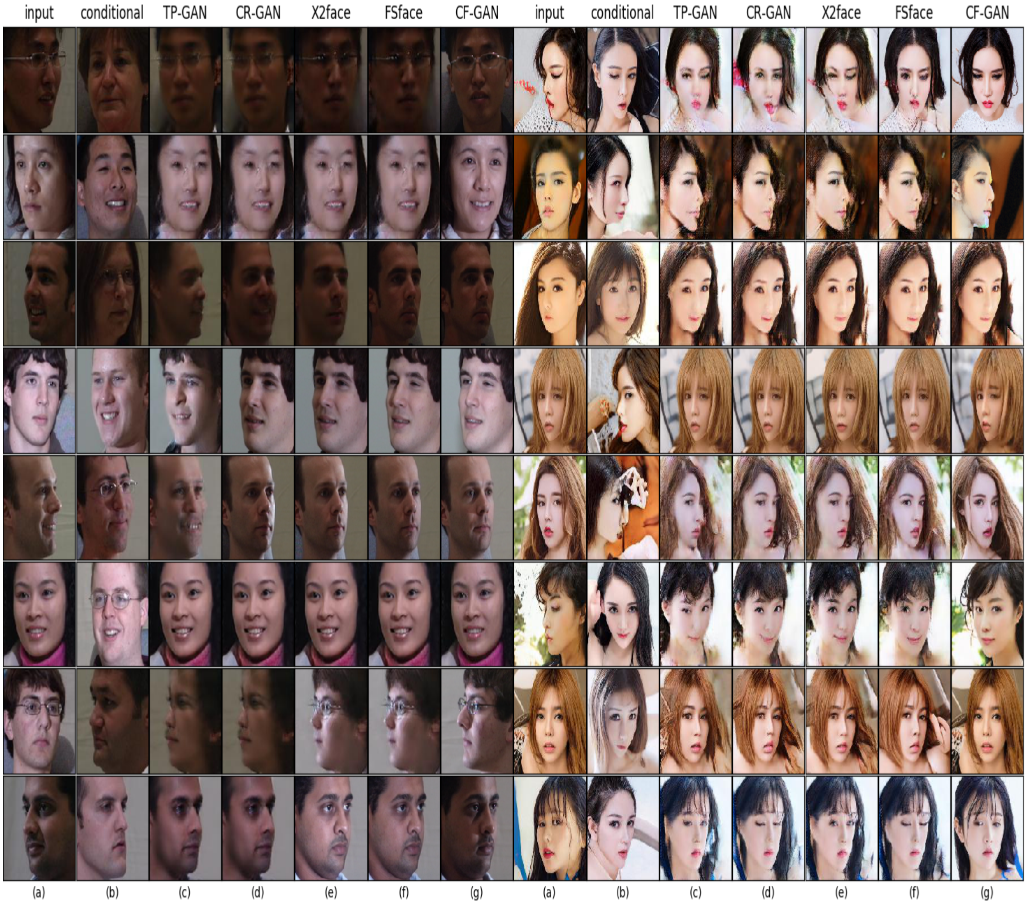

Figure 4. Experimental comparison results from two datasets under source face and conditional face have larger differences in shape and size. (**a**) Source faces. (**b**) Conditional faces. (**c**) Transfer results by TP-GAN. (**d**) Transfer results by CR-GAN. (**e**) Transfer results by X2face. (**f**) Transfer results by FSface. (**g**) Transfer results by CF-GAN.

For facial pose and anger expression transfer tasks, although all models learn the same mapping, the generated faces from CF-GAN exhibit the highest conformity to the conditional faces. The face of the Chinese female multi-angle dataset contains any angle from −90 to 90°, which poses two challenges to existing models that use classification labels as generation conditions. The first challenge is that it cannot hold the differences in shape and size between conditional and source faces, which are within the classification boundary. Boundary faces are easily misclassified. The second challenge is that it cannot hold the differences in shape and size between conditional faces and source faces, which are responsible for the majority of the classification. For these two challenges, the existing model cannot generate a reasonable face according to the conditions facing any angle.

Unlike these models, we used classification features rather than classification labels in the generator network. In addition, the average classification features of the conditional face, the average classification features of the source face, and the average classification features of the source face were used to estimate the classification features of the target face for the proposed generator network to generate the target face. The proposed two-stage classifier only cares about the normal classification features of the conditional face and not the difference from the source face. In this way, the mappings that we built can ignore the shape and size differences between conditional faces and source faces.

4.6. Quantitative Evaluations

The comparison effect for two tasks on three evaluation indicators is shown in Table 1.

Table 1. Quantitative evaluations in terms of facial pose and anger expression transfer (A for short), facial pose, and happy expression transfer (B for short) tasks from different models.

Tasks	Baselines	AMT	Cf	Cs	TIME
A	TP-GAN	24% ± 1.0%	0.71	0.69	36
	CR-GAN	26% ± 1.0%	0.79	0.75	36
	X2face	24% ± 1.0%	0.71	0.69	48
	FSface	26% ± 1.0%	0.80	0.76	48
B	CF-GAN	39% ± 1.0%	0.92	0.80	22
	TP-GAN	23% ± 1.0%	0.71	0.67	48
	CR-GAN	26% ± 1.0%	0.76	0.70	48
	X2face	23% ± 1.0%	0.71	0.67	61
	FSface	26% ± 1.0%	0.78	0.71	61
	CF-GAN	28% ± 1.0%	0.82	0.77	28

As can be seen from Table 1, the proposed model achieved leading numerical results compared to other models. This means that it not only generates more reasonable faces than other models but also successfully captures the latent facial pose and expression features.

4.7. Limitation

The proposed CF-GAN model shows that designing a reasonable classification features algorithm will obtain the generated facial pose and expression of experimental results that are close to conditional faces. With this improvement, virtual digital humans in the entertainment audience and video production industry can appear smoother and produce more realistic visualizations of humans and their faces. There is a jitter phenomenon when we synthesize these continuous face images into a video. For this limitation, stability features need to be learned from real-life videos.

Although the CF-GAN model has achieved leading results for facial pose and expression transfer, we noted that about 3% of the test results were unreasonable. For example, the unreasonable result and the input face do not appear to be the same person. Randomly selected unreasonable results will raise readers' doubts about the reliability of the experiment. The best transfer algorithm should eliminate all unreasonable results. The main reason for unreasonable results is classifier error. For this limitation, a better classifier is very necessary for the future.

5. Conclusions

This paper proposes a novel Generative Adversarial Networks model for facial pose and expression transfer. We used classification features rather than classification labels for the generator network. The mappings we built can ignore the difference in shape and size between conditional faces and source faces. In addition, we proposed combining two cost functions with different convergence speeds to learn pose and expression features. Compared to state-of-the-art models, the proposed CF-GAN model achieved leading scores for facial pose and expression transfer on two datasets. Our research provides an approach for more accurate virtual digital human synthesis.

Author Contributions: Conceptualization, Z.C.; methodology, Z.C. and L.S.; software, L.S.; validation, L.S. and W.W.; formal analysis, W.W. and S.N.; investigation, S.N. and W.W.; resources, L.S. and S.N.; data curation, Z.C. and S.N.; writing—original draft preparation, W.W.; writing—review and editing, S.N. and Z.C.; visualization, W.W. and L.S.; supervision, Z.C.; project administration, Z.C.; funding acquisition, W.W. 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 (No. 61702158), the Central Guidance on Local Science and Technology Development Fund of Hebei Province (No. 226Z1808G), the Science Foundation of Hebei Normal University (No. L2022B22), the Science Foundation of Hebei Normal University (No. L2021B38) and the Humanities and the Social Science Research Project of Hebei Education Department (No. SZ2023161).

Data Availability Statement: Multi-PIE [33] is a public dataset. Chinese female multi-angle dataset is generated by Stylegan [34] (https://github.com/a312863063/seeprettyface-face_editor, accessed on 10 March 2023).

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Tian, Y.; Peng, X.; Zhao, L.; Zhang, S.; Metaxas, D.N. CR-GAN: Learning complete representations for multi-view generation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; AAAI Press: Palo Alto, CA, USA, 2018; pp. 942–948.
2. Chang, H.; Lu, J.; Yu, F.; Finkelstein, A. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 40–48.
3. Chen, Y.C.; Lin, H.; Shu, M.; Li, R.; Tao, X.; Shen, X.; Ye, Y.; Jia, J. Facelet-bank for fast portrait manipulation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3541–3549.
4. Zhang, Y.; Zhang, S.; He, Y.; Li, C.; Loy, C.C.; Liu, Z. One-shot Face Reenactment. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019.
5. Zakharov, E.; Shysheya, A.; Burkov, E.; Lempitsky, V. Few-shot adversarial learning of realistic neural talking head models. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9459–9468.
6. Zhuang, N.; Yan, Y.; Chen, S.; Wang, H.; Shen, C. Multi-label learning based deep transfer neural network for facial attribute classification. *Pattern Recognit.* **2018**, *80*, 225–240. [CrossRef]
7. Sankaran, N.; Mohan, D.D.; Lakshminarayana, N.N.; Setlur, S.; Govindaraju, V. Domain adaptive representation learning for facial action unit recognition. *Pattern Recognit.* **2020**, *102*, 107–127. [CrossRef]
8. Bozorgtabar, B.; Mahapatra, D.; Thiran, J.P. ExprADA: Adversarial domain adaptation for facial expression analysis. *Pattern Recognit.* **2020**, *100*, 107–111. [CrossRef]
9. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 8–13 December 2014; pp. 2672–2680.
10. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8789–8797.
11. Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; Courville, A. Adversarially learned inference. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
12. Tran, L.; Yin, X.; Liu, X. Disentangled representation learning gan for pose-invariant face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1415–1424.
13. Zhao, B.; Wu, X.; Cheng, Z.Q.; Liu, H.; Jie, Z.; Feng, J. Multi-view image generation from a single-view. In Proceedings of the 2018 ACM Multimedia Conference on Multimedia Conference, Seoul, Republic of Korea, 22–26 October 2018; ACM: New York, NY, USA, 2018; pp. 383–391.
14. Hannane, R.; Elboushaki, A.; Afdel, K. A Divide-and-Conquer Strategy for Facial Landmark Detection using Dual-task CNN Architecture. *Pattern Recognit.* **2020**, *107*, 107504. [CrossRef]

15. Huang, R.; Zhang, S.; Li, T.; He, R. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2439–2448.
16. Ramamoorthi, R. Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1322–1333. [CrossRef]
17. Lee, K.C.; Ho, J.; Kriegman, D.J. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 684–698. [PubMed]
18. Kent, M.G.; Schiavon, S.; Jakubiec, J.A. A dimensionality reduction method to select the most representative daylight illuminance distributions. *J. Build. Perform. Simul.* **2020**, *13*, 122–135. [CrossRef]
19. Savelonas, M.A.; Veinidis, C.N.; Bartsokas, T.K. Computer Vision and Pattern Recognition for the Analysis of 2D/3D Remote Sensing Data in Geoscience: A Survey. *Remote Sens.* **2022**, *14*, 6017. [CrossRef]
20. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8798–8807.
21. Hosoi, T. Head Pose and Expression Transfer Using Facial Status Score. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; IEEE: New York, NY, USA, 2017; pp. 573–580.
22. 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 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 818–833.
23. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
24. Wu, W.; Zhang, Y.; Li, C.; Qian, C.; Change Loy, C. Reenactgan: Learning to reenact faces via boundary transfer. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 603–619.
25. Yang, H.; Huang, D.; Wang, Y.; Jain, A.K. Learning face age progression: A pyramid architecture of gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 31–39.
26. Wiles, O.; Sophia Koepke, A.; Zisserman, A. X2face: A network for controlling face generation using images, audio, and pose codes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 670–686.
27. Liu, M.Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; Kautz, J. Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10551–10560.
28. Yang, S.; Jiang, L.; Liu, Z.; Loy, C.C. Unsupervised image-to-image translation with generative prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18332–18341.
29. Xu, Y.; Xie, S.; Wu, W.; Zhang, K.; Gong, M.; Batmanghelich, K. Maximum Spatial Perturbation Consistency for Unpaired Image-to-Image Translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18311–18320.
30. Pugliese, R.; Regondi, S.; Marini, R. Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Sci. Manag.* **2021**, *4*, 19–29. [CrossRef]
31. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 5967–5976.
32. Cao, Z.; Niu, S.; Zhang, J.; Wang, X. Generative adversarial networks model for visible watermark removal. *IET Image Process.* **2019**, *13*, 1783–1789. [CrossRef]
33. Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S. Multi-pie. *Image Vis. Comput.* **2010**, *28*, 807–813. [CrossRef] [PubMed]
34. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410.
35. Valdenegro-Toro, M.; Arriaga, O.; Plöger, P. Real-time Convolutional Neural Networks for emotion and gender classification. In Proceedings of the ESANN, Bruges, Belgium, 27–29 April 2019.
36. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
38. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, 60, 1097–1105. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.