symmetry

MDPI

Article

The Facial Expression Data Enhancement Method Induced by Improved StarGAN V2

Baojin Han [1,2,*] and Min Hu [1,2]

[1] Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230601, China
[2] Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, School of Computer and Information, Hefei University of Technology, Hefei 230601, China
* Correspondence: hanbaojin@mail.hfut.edu.cn

Abstract: Due to the small data and unbalanced sample distribution in the existing facial emotion datasets, the effect of facial expression recognition is not ideal. Traditional data augmentation methods include image angle modification, image shearing, and image scrambling. The above approaches cannot solve the problem that is the high similarity of the generated images. StarGAN V2 can generate different styles of images across multiple domains. Nevertheless, there are some defects in gener-ating these facial expression images, such as crooked mouths and fuzzy facial expression images. To service such problems, we improved StarGAN V2 by solving the drawbacks of creating pictures that apply an SENet to the generator of StarGAN V2. The generator's SENet can concentrate at-tention on the important regions of the facial expression images. Thus, this makes the generated symmetrical expression image more obvious and easier to distinguish. Meanwhile, to further im-prove the quality of the generated pictures, we customized the hinge loss function to reconstruct the loss functions that increase the boundary of real and fake images. The created facial expression pictures testified that our improved model could solve the defects in the images created by the original StarGAN V2. The experiments were conducted on the CK+ and MMI datasets. The correct recognition rate of the facial expressions on the CK+ was 99.2031%, which is a 1.4186% higher accuracy than that of StarGAN V2. The correct recognition rate of the facial expressions on the MMI displays was 98.1378%, which is 5.059% higher than that of the StarGAN V2 method. Furthermore, contrast test outcomes proved that the improved StarGAN V2 performed better than most state-of-the-art methods.

Keywords: face expression recognition; data enhancement; StarGAN V2; hinge loss; SENet; symmetry and asymmetry; GAN; deep learning

Citation: Han, B.; Hu, M. The Facial Expression Data Enhancement Method Induced by Improved StarGAN V2. *Symmetry* **2023**, *15*, 956. https://doi.org/10.3390/sym15040956

Academic Editor: Hsien-Chung Wu

Received: 6 March 2023
Revised: 27 March 2023
Accepted: 29 March 2023
Published: 21 April 2023

1. Introduction

Facial expression recognition (FER) has a vital role in emotional communication. As technology advances, facial expression recognition has been widely merged into our lives [1–4]. Human–computer symmetrical interaction and e-learning, etc., have become the current research hotspots for machine learning and artificial intelligence [5–8]. Traditional expression recognition consists mainly of three branches: image preprocessing, feature extraction, and classification. The most important branch in the whole process is feature extraction. This process is related to the correct recognition rate of facial expressions. In feature extraction, there are mainly active appearance models (AAM), which are based on the localization of facial feature points, and local feature extraction/algorithms, such as Gabor wavelets, local binary patterns (LBP) [9], and multi-feature fusion [10,11], etc. Traditional feature extraction approaches in facial expression recognition applications can also be detected in [12,13]. Nevertheless, these feature extractions require sufficient artificial experience to design and they have limitations, such as a weak robustness to image sizes, illumination effects, and image angles.

Compared to traditional feature extraction methods, deep learning can autonomously learn the features in facial expression images and obtain a better recognition rate. Santosh et al. [14] proposed an effective facial expression recognition method for identifying six basic facial expression images. Wang et al. [15] proposed a multi-task depth framework that used key features to recognize facial expressions. Ruan et al. [16] proposed a novel deep disturbance-disentangled learning (DDL) method for FER. To improve the recognition rate of these facial expressions, CNN models continuously add depth and breadth to the network [17–22]. However, with an increase in the depth and width of the network, the number of parameters will also increase rapidly. In this case, if we continue to use the small dataset, the over-fitting problem occurs. With an increase in the parameters, over-fitting is more likely to happen. However, it is impossible to offer extensive samples in some training phases. Some facial expression datasets have small sample sizes and unbalanced sample distributions. The numbers of happy and disgusted images in the CK+ dataset are 69 and 59, respectively, while the numbers of fear and contempt images are 25 and 18, respectively. As discussed above, this makes it hard for deeper CNN models to get good outcomes. Therefore, solving the issues of insufficient sample sizes and unbalanced sample distributions has become the key problem.

Data enhancement is one of the effective approaches to solving insufficient data and unbalanced sample distributions. Traditional image enhancement methods are generally geometric transformations. Zhou et al. [23] constructed a new face sample by passing the symmetry of an image through a classifier that combined conventional cooperative expression with inverse linear regression. Li et al. [24] used a horizontal mirror transformation for their data enhancement. Tripathi et al. [25] proposed an adaptive geometric filter approach for gray and color image enhancement. Different from simple geometric transformations or increased illumination for data enhancement, generative adversarial networks (GAN) [26] can effectively solve the high similarity of generated images. GANs can generate the same facial expressions, and different models need to be retrained when getting different facial expressions. Thus, this has caused training to be redundant. Choi et al. [27] proposed StarGAN V2, which generates images of diverse styles over multiple domains. Therefore, various facial expression images can be generated in a model, reducing the redundant models. As shown in Figure 1, the generator is inputted by different features to generate the target images. Important features will have a huge effect on the generated images. To generate more vivid images, we introduce an SENet [28] to select the important features. Hinge loss [29] is used to find the maximum margin between the real and fake images to enhance the realism of the created images.

To solve the problems of insufficient sample sizes and unbalanced sample distributions in these facial expression images, we introduced StarGAN V2 to enhance the facial expression datasets. To further improve the vividness of the created images and reduce the redundant features, an SENet was added to the generator in StarGAN V2. The SENet mainly extracted the vital facial expression features. Our network introduced the idea of relative discrimination [30]. We replaced the standard discriminator with a relative discriminator. Additionally, we increased the ratio of the fake samples in the initial training to achieve a better training state. Our network introduced hinge loss to improve the authenticity of the created images.

The main contributions of this work are as follows:

(1) To solve the problems of insufficient data and unbalanced sample distributions in facial expression datasets, we used an improved StarGAN V2 model to generate facial images with different emotions. StarGAN V2 is an efficient method for enhancing facial expression images. StarGAN V2 generates different expression images in a model.
(2) To improve the vividness of the created images, an SENet was added to the generator in StarGAN V2 to extract the important facial expression features.
(3) We introduced hinge loss and relative discrimination. Hinge loss was used to find the maximum margins among the different sample distributions and improve the

authenticity of the generated images. Relative discrimination made our network achieve a better training state.

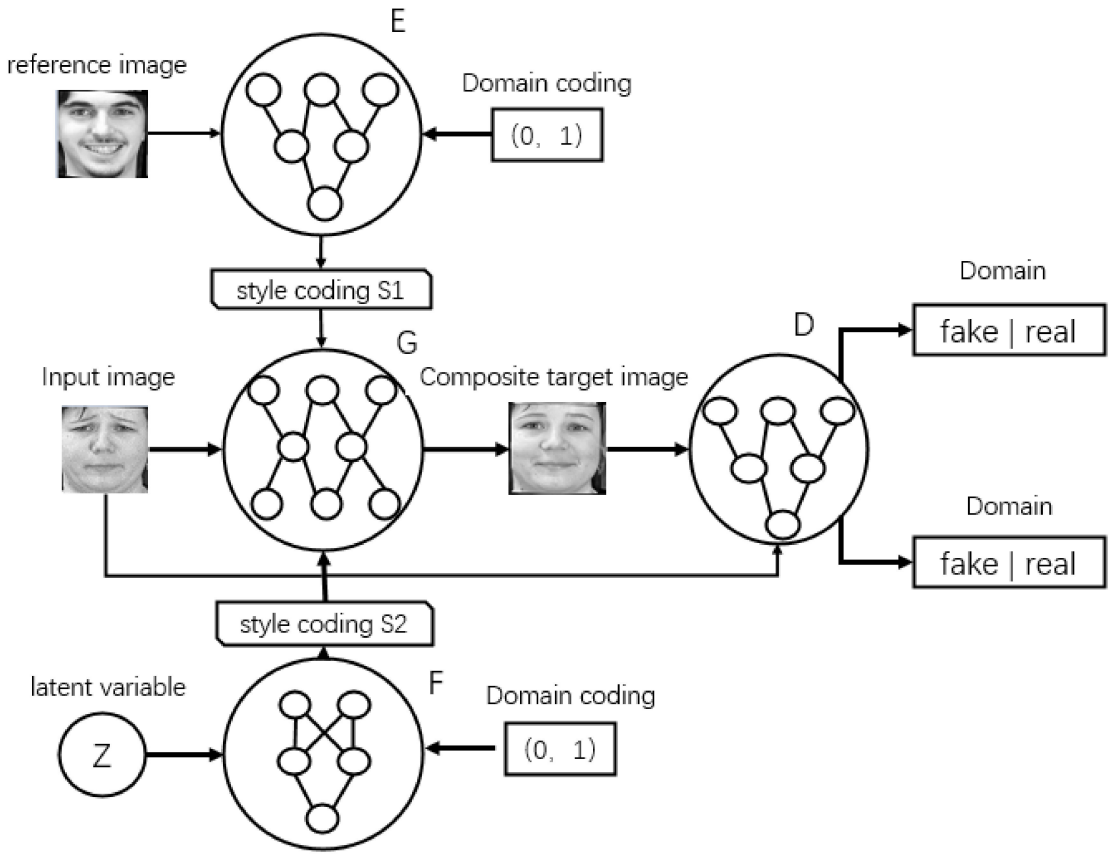

Figure 1. The structure of StarGAN V2.

The remainder of the paper is organized as follows. In Section 2, we review some related work, including StarGAN V2 and the SENet. Our proposed network is described in detail in Section 3. In Section 4, we introduce some public databases and ablation experiments and display our experimental results, while the conclusion is placed in Section 5.

2. Related Work

2.1. Facial Expression Recognition

FER has been universally utilized within various fields, such as human–computer interaction, medical assistance, and digital entertainment. Traditional expression recognition can be split into three main procedures: image preprocessing, feature extraction, and classification. The most important of these steps is feature extraction, which directly affects the correct recognition rate of facial expressions. Traditional expression feature extraction relies on the various statistics of the pixels' values, including the facial images. Examples of this include principal component analyses (PCA) [31], LBPs [32], and Gabor transforms [33]. Aroram et al. [34] applied the hybridization of feature extraction, which achieved good results. Islam et al. [35] extracted features from the segmented parts using a fusion of the histogram of oriented gradients (HOG) and LBPs. The dimension of the feature vector was reduced by using a PCA. Bisogni et al. [36] proposed a multi-input hybrid FER mode, due to the various limitations of traditional feature extraction methods. These limitations are their manual design, less characteristic information, and other problems. Traditional feature extraction methods have various drawbacks. Therefore, traditional feature extraction methods find it difficult to achieve good results for facial expression recognition.

Over the last few years, facial expression recognition based on deep neural networks has advanced a lot. It is as follows that the current progress of feature extraction uses

deep learning. Deep learning has challenged the traditional feature extraction methods in facial expression recognition and it can execute feature extraction and classification. Feature extraction executed by deep learning uses continuous optimization with loss functions. Most of the applications of deep learning in facial expression recognition are based on a CNN structure. The CNN structure includes AlexNet [17], VGGNet [18], Googlenet [19], ResNet [20], MobileNet [21], and DensNet [22]. Naim et al. [37] merged a CNN and SVM to produce a new model that improved the recognition rate of facial expressions. Sadeghi et al. [38] put forward deep histogram metric learning in a CNN for facial expression recognition. However, some of the existing emotional datasets have some problems. These problems are insufficient data, unbalanced sample distributions, and high-similarity samples. These problems lead to unsatisfactory recognition results of the network. The data enhancement method can effectively solve the above problems.

As cited, [37,38] include the the latest technologies. In [37], the CNN and SVM were merged to create a new model. Triple loss is discussed in [38]. They all use the CNN structure and demonstrate the effectiveness of this structure for facial expression recognition. The technologies learn facial expression images from various aspects and are effective in improving the recognition rate of these facial expression images.

2.2. Data Enhancement

With the advancement of multimedia technology, the facial expression recognition technique is universally used in social life. Due to various reasons, some existing facial expression datasets have insufficient data and unbalanced sample distributions. With an increase in the network's scale, the network also increases its number of parameters. The over-fitting phenomenon would appear in the network because the facial expression datasets are limited. Therefore, data enhancement is necessary. Some examples of traditional image augmentation methods are geometric transformation [39] and color space transformation [40]. Xin et al. [41] used a single sample face image for facial reconstruction. The reconstructed facial image and original counterpart were treated as a new training sample set. The effectiveness of the algorithm was proven by a commonly used database. Ramasubramanian et al. [42] presented an automatically generated 3D facial model. Majid et al. [43] proposed a triple dynamic clipped histogram equalization (TDCHE) method. Traditional image enhancement approaches have good outcomes for some aspects. However, these approaches are mainly dependent on manual design and do not have the ability for autonomous learning. In addition, the images generated by traditional image enhancement methods have a high similarity.

With the advancement of deep learning in computer vision, GANs have many advantages and can generate different types of images. These generated image types include style transfer, attribute transfer, improved image vividness, and so on. Adversarial training makes it easier to generate more realistic images.

A GAN contains two parts: a generator and discriminator. The generator is input into a random vector and generates the images. The discriminator distinguishes between the real and fake pictures. After successive iterations, GANs can produce the target images. With the development of deep learning, GAN models have also been improved. Mirza et al. [44] introduced sufficient semantic guidance and a penalty mechanism. Fu et al. [45] proposed a conditional generative adversarial network (cGAN) to establish the relationship among the emotion-related EEG data, coarse markers, and facial expression images. Their experimental results proved the rationality of the method. Zhu et al. [46] translated an image from a source domain to a target domain in the absence of paired examples. Dou et al. [47] proposed an asymmetric cycle–GAN model to use the asymmetric need in NIR–RGB translations.

Choi et al. [48] proposed StarGAN to perform the image-to-image translations for multiple domains using a model. The model was proposed based on GANs, deep convolutional generative adversarial networks (DCGAN), and conditional generative adversarial networks (CGAN). Therefore, StarGAN can flexibly translate an input picture into a different target style via a model. The generator in StarGAN can generate various facial

expression images. The generated images keep the original identity information of the input pictures. Experiments have proved effective in face attribute synthesis and facial expression transformation.

StarGAN V2 was proposed based on the StarGAN network and has achieved a good performance in generating facial expression images [27]. The generated images retain the original identity information of the input images. A mapping network was added to StarGAN V2 that could convert the images of one domain into multiple images of the target domain. Therefore, we chose StarGAN V2 to generate our facial expression pictures. However, we found some defects in the images generated by StarGAN V2. The generated images have crooked mouths, fuzzy images, and so on. To further promote the quality of the generated images, we replaced the standard loss function with the hinge loss function in StarGAN V2. We added an SENet to StarGAN V2's generator to promote the vividness of the generated images.

The structure of StarGAN V2 is shown in Figure 1. StarGAN V2 is composed of a generator G, discriminator D, mapping network F, and style encoder E. The style encoder E generates the style code $S1$. The mapping network F generates the style code $S2$. $S1$, $S2$, and the original images are input into the generator G. The generator G can generate different styles of the target images. The discriminator D is a multi-task discriminator that consists of multiple output branches. Each branch is a binary classification to determine if an image is real or fake. The mapping network F consists of an MLP to provide style codes for all the available domains. Additionally, the mapping network F contains two inputs: one is a potential encoding converted into a multiple domain style encoding and the other is generated by random noise. The style encoder E can produce diverse style codes from diverse reference pictures.

YANG et al. [49] proposed an Image-to-Image (I2I) Translation based on StarGAN V2. Experiments have proven the effectiveness of this method.

In [47,49], the latest technologies are included. Paired samples were not used for training in [47,49]. This could solve samples being newly generated due to paired samples. In [47,49], various loss functions were employed to ensure the quality of the generated images. Both models used generative adversarial networks. There was a trend towards unpaired samples in the image generations. This solved the problem of not being able to generate new images due to paired image samples. Multiple loss functions were used to ensure the quality of the generated images. The networks learnt the original image from different perspectives and improved the quality of the generated images.

From the above theory derivation and application, StarGAN V2 is indeed effective in generating images. Therefore, we used StarGAN V2 to generate facial expression images.

2.3. Squeeze-and-Excitation Networks

An SENet (Squeeze-and-Excitation network) [28] is an attention mechanism. The model can be inserted into CNN models with a low overhead. Its key features are strengthened to improve the quality of the generated images. Figure 2 is the structure of an SENet, which mainly consists of two parts: Squeeze (Figure 2 F_{sq}) and Excitation (Figure 2 F_{ex}). These two parts complete the adaptive scale of the feature channel.

The feature ($X \in R^{w \times h \times c}$) is input into the SENet, and the output becomes the feature ($\widetilde{X} \in R^{w \times h \times c}$). The SENet contains a squeeze operator (Figure 2 F_{sq}) and excitation operator (Figure 2 F_{ex}). The squeeze operator embeds information from the global receptive field into a channel in each layer. The squeeze operator produces a sequence S in $1 \times 1 \times c$, which represents the correlations of each layer. The squeeze operator is (1):

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j), \quad (1)$$

z_c is the c-th element of the squeezed channels and F_{sq} is the squeeze function. u_c represents the c-th channel of the input features. The height and width of the input images are H and W, respectively.

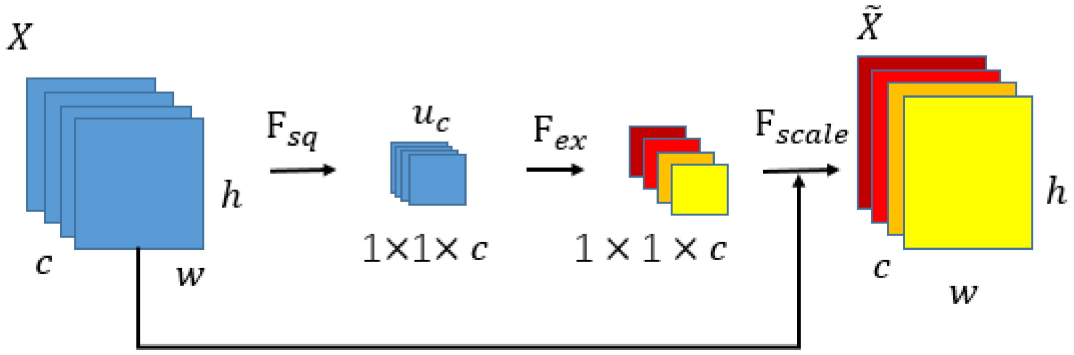

Figure 2. SENet network structure diagram.

The excitation operator is later used to execute a feature recalibration. The excitation operator is shown as follows (2):

$$s_c = F_{ex}(z, W) = \sigma(W_u \delta(W_d z_c)) \tag{2}$$

F_{ex} is the excitation function and z_c is the input-squeezed signal from the last layer. δ represents the ReLU activation function. $W_d \in \mathbb{R}^{c \times \frac{c}{r}}$ is the channel using the 1×1 convolution and the dimensionality reduction ratio r is reduced. The final output of the block $\widetilde{x_c}$ is readapted to make use of the channel s_c, and is as follows (3):

$$\widetilde{x_c} = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{3}$$

Chen et al. [50] proposed a three-stream 3D CNN, which is called an SE three-stream fusion network (SETFNet) for near-infrared facial expression recognition. By using an SE block, the model automatically, adaptively learns the weights of various local features to further promote an accurate rate of recognition.

Nguyen et al. [51] put forward a SqueezeNet–SE model, which combined CNNs with an SE block. An SE block was used to indicate the importance of the feature maps in each module.

As cited, [50,51] include the latest technologies that have been proposed. Both models contain SENets, which suggests that SENets can identify some important regions for network learning, with a focus on the regions of interest.

The above theory derivation and application examples suggest that, under the same conditions, an SENet can achieve better vital features. Therefore, this paper employed an SENet, which could effectively improve the vividness of the generated images.

3. Proposed Method

With a continuously increasing network structure, the demand magnitude of the required training data increases. However, many existing facial expression datasets affect the performance of deep neural networks. Those datasets contain insufficient sample sizes and unbalanced sample distributions. Increasing these sample sizes and balancing the sample distributions become necessary. Therefore, data enhancement tasks apply GANs, but this method needs to train different models repeatedly, which takes up a lot of resources and a long training time. In the meantime, StarGAN V2 provides a better result for the above problems. StarGAN V2 generates different expressions in facial images when inputting one facial image and retains the identity information of the input images. Its generator can achieve various styles when learning the different original images. Nevertheless, the importance of each feature is different, which produces different effects on the image generation. Concerning the generation of more vivid images, our model adopted an SENet to pick up the important features, while ignoring the redundant ones. Our generator is the attention mechanism + generator (*AttG*) in Figure 3. To improve the quality of the

image generation, we improved the reconstruction loss by introducing the hinge loss. Our network structure is shown in Figure 4.

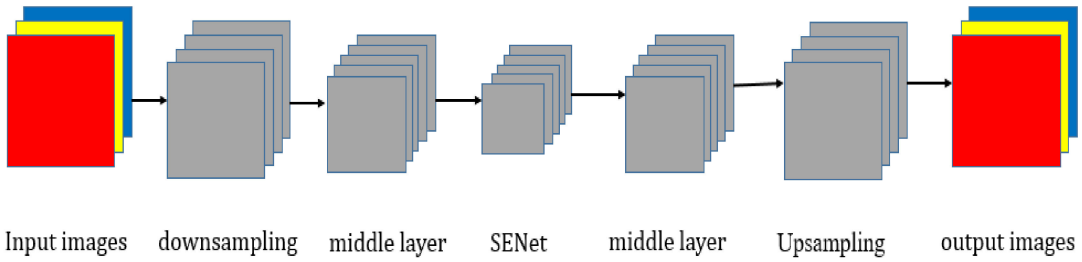

Figure 3. The Flowchart of AttG.

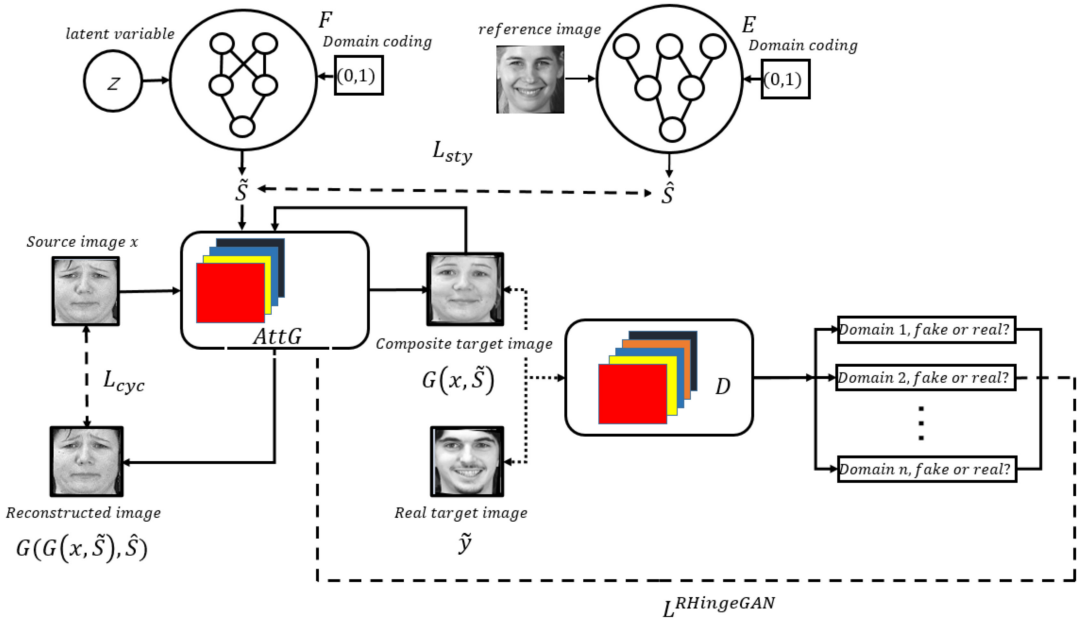

Figure 4. The Flowchart of proposed network architecture.

3.1. The Proposed Network

As shown in Figure 4, our network structure is composed of $AttG$, a discriminator D, a mapping network F (mapping network F forms the style code $\widetilde{S}$), and a style encoder E (style encoder E produces the style code $\hat{S}$). F learns the style coding of the target domain $\widetilde{S}, \widetilde{S} = F_{\tilde{y}}(z)$. E learns the style coding of the source domain $\hat{S}$, $\hat{S} = E_{\tilde{y}}(x)$, while the discriminator D distinguishes between the real and fake images. The network is optimized by an adversarial loss $L^{RHingeGAN}$.

The generator $AttG$ can produce different styles of emoticon images. The discriminator D is a multi-task discriminator that is composed of multiple output branches. Each branch is a binary classification that is used to determine whether an image is real or fake. The mapping network F is composed of an MLP with multiple output branches to provide style codes for all the available domains. Additionally, the mapping network F contains two inputs: one is a potential encoding converted into a multiple domain style encoding, and the other is generated by random noise. The style encoder E can produce diverse style codes from the various reference emoticon images.

When training $AttG$, an original picture x and target domain style code $\widetilde{S}$ are imported into $AttG$. $AttG$ outputs the synthesized target image $G\left(x, \widetilde{S}\right)$. The synthesized target image $G\left(x, \widetilde{S}\right)$ and the source image style code $\hat{S}$ are inputted into $AttG$ again. Finally,

AttG outputs a reconstructed image $G\left(G\left(x,\widetilde{S}\right)\right)$. Preserving the similarity between x and $G\left(G\left(x,\widetilde{S}\right)\right)$ is maintained through the cyclic consistency loss L_{cyc}. Through continuous adversarial training, *AttG* generates realistic pictures that can be classified as the target domains.

3.1.1. Attention Generator *G*

The original generator *G* structure of StarGAN V2 is shown in Table 1. The generator *G* finds the sample distribution law and generates facial expression images with similar distribution laws. When a source image x is inputted, a corresponding composite image is generated. The content information provided by the original picture x and the style code provided by the target domain picture $\widetilde{S}$ are inputted into the generator *G*. The generator *G* generates the image $G\left(x,\widetilde{S}\right)$. $\widetilde{S}$ is injected into *G* using adaptive instance normalization (*AdaIN*). We observe that $\widetilde{S}$ is designed to represent the style of a particular domain of the source image x. This eliminates the need to supply a source image to *G* and allows *G* to synthesize an image of the target domain. The generator converts the input image into an output image that reflects the style code that is specific to the domain. To better extract the important features, an SENet is added to *G*. The specific structure is shown in Table 2. This makes the generated image more vivid.

Table 1. Generator architecture.

Layer	Resample	Norm	Output Shape
Image x	-	-	256 × 256 × 3
Conv1 × 1	-	-	256 × 256 × 64
ResBlock	AvgPool	IN	128 × 128 × 128
ResBlock	AvgPool	IN	64 × 64 × 256
ResBlock	AvgPool	IN	32 × 32 × 512
ResBlock	AvgPool	IN	16 × 16 × 512
ResBlock	-	IN	16 × 16 × 512
ResBlock	-	IN	16 × 16 × 512
ResBlock	-	AdaIN	16 × 16 × 512
ResBlock	-	AdaIN	16 × 16 × 512
ResBlock	Upsample	AdaIN	32 × 32 × 512
ResBlock	Upsample	AdaIN	64 × 64 × 256
ResBlock	Upsample	AdaIN	128 × 128 × 128
ResBlock	Upsample	AdaIN	256 × 256 × 64
Conv1 × 1	-	-	256 × 256 × 3

Table 2. Attention mechanism + Generator architecture.

Layer	Resample	Norm	Output Shape
Image x	-	-	256 × 256 × 3
Conv1 × 1	-	-	256 × 256 × 64
ResBlock	AvgPool	IN	128 × 128 × 128
ResBlock	AvgPool	IN	64 × 64 × 256
ResBlock	AvgPool	IN	32 × 32 × 512
ResBlock	AvgPool	IN	16 × 16 × 512
ResBlock	-	IN	16 × 16 × 512
ResBlock	-	IN	16 × 16 × 512
SENet	-	-	16 × 16 × 512
ResBlock	-	AdaIN	16 × 16 × 512
ResBlock	-	AdaIN	16 × 16 × 512
ResBlock	Upsample	AdaIN	32 × 32 × 512
ResBlock	Upsample	AdaIN	64 × 64 × 256
ResBlock	Upsample	AdaIN	128 × 128 × 128
ResBlock	Upsample	AdaIN	256 × 256 × 64
Conv1 × 1	-	-	256 × 256 × 3

In the generator, Instance Normalization (*IN*) normalizes the feature statistics of a single sample to keep the contents of the images. Thus, the content of the images before and after translation is kept unchanged, while the style of the images is changed. The AdaIN layer aligns the mean and variance of the content features with those of the style features and realizes the style transfer. Hence, we use the SENet to enhance some detailed features to make the generated images more vivid. We place the SENet in the middle layer (the middle layer is shown in Figure 3).

3.1.2. Multi-Task Discriminator *D*

The discriminator *D* tries to distinguish between the fake and real pictures. The discriminator in our network is a multi-task discriminator, which contains multiple linear output branches. Each branch employs a binary classification to determine the real domain $\widetilde{y}$ of an image y or the synthesized target image $G\left(x, \widetilde{S}\right)$ generated by *AttG*. Multiple classifiers are avoided by making general judgments about whether the resulting image is authentic or not. This is because we want the generated images to be true in a particular domain, rather than the entire image to be real. The generated images make the optimization more specific. The specific structure of the multi-task discriminator *D* is shown in Table 3.

Table 3. Discriminator architectures and style encoder.

Layer	Resample	Norm	Output Shape
Image x	-	-	$256 \times 256 \times 3$
Conv1 $\times$ 1	-	-	$256 \times 256 \times 64$
ResBlock	AvgPool	-	$128 \times 128 \times 128$
ResBlock	AvgPool	-	$64 \times 64 \times 256$
ResBlock	AvgPool	-	$32 \times 32 \times 512$
ResBlock	AvgPool	-	$16 \times 16 \times 512$
ResBlock	AvgPool	-	$8 \times 8 \times 512$
ResBlock	AvgPool	-	$4 \times 4 \times 512$
LReLU	-	-	$4 \times 4 \times 512$
Conv4 $\times$ 4	-	-	$1 \times 1 \times 512$
LReLU	-	-	$1 \times 1 \times 512$
Reshape	-	-	512
Linear $\times$ K	-	-	D $\times$ K

3.1.3. Style Coder *E*

The style encoder *E* generates different style codes using various reference pictures. An input picture is symbolized as x and its corresponding target domain is marked as y. *E* can extract the stylistic encoding $\hat{S}, \hat{S} = E_{\widetilde{y}}(x)$ from the source image x. The style encoder is the same as the multi-branch discriminator structure setting. The style encoder *E* can generate diversified style codes using different reference pictures. It allows the generator *G* to synthesize an output image that reflects the style $\hat{S}$ of the reference images. The style encoder *E* is used to extract the different image style features from the different reference images. Therefore, the network can provide a variety of style features with the use of different reference images for training. The specific structure of the style encoder *E* is shown in Table 3.

3.1.4. Mapping Network

The mapping network *F* accepts the latent code from the standard Gaussian distribution, and subsequently, the generator can get rid of the label constraint to generate the target images. Given a domain y, with a latent encoding z as its input, a network encoding $\widetilde{S} = F_{\widetilde{y}}(z)$ is generated by the mapping network *F*. *F* is composed of an MLP with multiple output branches to offer style codes for all the available domains. The specific structure of the mapping network *F* is displayed in Table 4.

Table 4. Mapping network.

Type	Layer	Actvation	Output Shape
Shared	Latent z	-	16
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
Shared	Linear	ReLU	512
Unshared	Linear	ReLU	512
Unshared	Linear	ReLU	512
Unshared	Linear	ReLU	512
Unshared	Linear	-	64

3.2. Improved Reconstruction Loss Function

The choice of a loss function is a significant factor affecting the network's performance. The essence of a GAN's adversarial loss is to find the Nash equilibrium solution in the zero-sum game. During the image's translation, the generator will generate samples that appear to match the distribution of the source dataset. $P_{data}(x)$ represents the sample distribution of the image's domain x, $P_{data}(z)$ represents the sample distribution of the image's domain z, and the loss function is defined in the original GAN as follows:

$$min\ L_{GAN}(D_x) = -\left(E_{x\sim P_{data(x)}}[logD_x(x)] + E_{z\sim P_{data(z)}}[log(1 - D_x(G(z))]\right) \tag{4}$$

$$min\ L_{GAN}(G) = E_{x\sim P_{data(x)}}[logD_x(x)] + E_{z\sim P_{data(z)}}[log(1 - D_x(G(z))] \tag{5}$$

In Formula (4) and Formula (5), the discriminator D measures the optimization of G under the optimal D as equal to the optimization of the JS divergence (Jensen–Shannon divergence) [26] between $P_{data}(x)$ and $P_{data}(z)$. There is a minimum value for the JS divergence, but the discriminator does not know the a priori knowledge that half of the input data is true and half is fake. There may be an actual situation in that all the inputs of x have $D(x) \approx 1$. This makes it difficult for the discriminator to rely on both the real and created data. Eventually, the probability of the real data and created data finds it difficult to reach 0.5 in its ideal state, and it is hard to find the real Nash equilibrium settlement.

Considering the problem that standard GAN adversarial loss fails to make complete use of the prior knowledge of the input data being half real and half fake, our network introduces the idea of relative discrimination. This means replacing the standard discriminator with a relative discriminator to increase the ratio of the fake samples in the initial training, in order to achieve a better training state. In facial expression generation, increasing the spacing boundary between true and false images can improve the authenticity of the created pictures. Therefore, using a combination of relative discrimination and hinge loss, the loss functions of the discriminator and generator are shown in Formula (6) and Formula (7), respectively:

$$L_{D_{src}}^{RHingeGAN} = E_{x\sim P_{data(x)}}[\max(0, 1 - D_{src}(x) + D_{src}(G(x, l_{target}))) + \max(0, 1 + D_{src}(G(x, l_{target})) - D_{src}(x))] \tag{6}$$

$$L_G^{RHingeGAN} = E_{x\sim P_{data(x)}}[(D_{src}(G(x, l_{target}), x)))] \tag{7}$$

In Formulas (6) and (7), P_{data} represents the real data distribution. D_{src} represents the real and fake discriminative structure in the discriminator D. l_{target} represents the target label.

The style reconstruction loss forces $AttG$ to use the style encoding $\widetilde{s}$ when generating the image $G(x, \widetilde{s})$. This is shown in the following Formula (8):

$$L_{sty} = E_{x,\widetilde{y},z}\left[\|\widetilde{s} - E_{\widetilde{y}}(G(x, \widetilde{s}))\|_1\right] \tag{8}$$

To further make *AttG* produce different images, the diversity sensitivity loss is used to regularize *AttG*, as in Formula (9),

$$L_{ds} = E_{x,\widetilde{y},z_1,z_2}[\|G(x,\widetilde{s_1}) - G(x,\widetilde{s_2})\|_1] \tag{9}$$

The target style codes $\widetilde{s_1}$ and $\widetilde{s_2}$ are generated by the mapping style F, according to the two random latent codes z_1 and z_2, namely $\widetilde{s_i} = F_{\widetilde{y}}(z_i)$ for $i \in \{1,2\}$. Maximizing the regular term can force *AttG* to explore the picture space and find meaningful style features to create various pictures. However, it is not guaranteed that the created picture will only change the content related to the input picture domain and retain the other contents of the input picture. Therefore, the cyclic consistency loss is used in the generator, as in Formula (10)

$$L_{cyc} = E_{x,y,\widetilde{y},z}[\|x - G(G(x,\widetilde{s}),\hat{s})\|_1] \tag{10}$$

where $\hat{s}$ is the style code of the input picture x. y is the target domain of x. The synthesized image $G(x,\widetilde{s})$ and $\hat{s}$ are input into *AttG*, in an attempt to reconstruct the source picture x. Additionally, the reconstructed image $G(G(x,\widetilde{s}),\hat{s})$ and x calculate the difference of the $L1$ norm. *AttG* learns to change its style while retaining the original characteristics of x. Finally, the complete objective function of optimizing *AttG* and D is as in Formula (11)

$$L_{F,G,E} = L^{RHingeGAN} + \lambda_{sty}L_{sty} - \lambda_{ds}L_{ds} + \lambda_{cyc}L_{cyc} \tag{11}$$

where λ_{sty}, λ_{ds}, and λ_{cyc} denote the style reconstruction loss hyperparameters, diversity sensitive loss hyperparameters, and cycle consistency loss hyperparameters, and the correlation coefficients are set as 1, 2, 1. $L^{RHingeGAN}$ represents the adversarial loss between the generator and discriminator.

3.3. The Algorithm in the Paper

The generator *AttG* and multi-task discriminator D conduct adversarial training in alternating ways. First, we fix *AttG* to train D, followed by fixing D to train *AttG*, and then continue with the cyclic training. The abilities of *AttG* and D are enhanced. Eventually, the images generated by *AttG* can be seen as real. For example, in facial expression generation, when one facial expression image generates different facial expression images, the specific training process is as follows:

(1) *AttG* is fixed, training D with $\widetilde{y}$ and training n epochs (once for each sample in the training sample set).
(2) $G(x,\widetilde{s})$ is generated from *AttG*, training D with $G(x,\widetilde{s})$ and training n epochs.
(3) D is fixed, using the output of D as the image's label, calculating the loss function, and continuing with training *AttG* and the n epochs.
(4) Steps (1)–(3) are repeated until the images generated by *AttG* can be seen as real.
(5) The test image set is selected to assess the performance of the final network and the quality of the generated image.

4. Experimental Results and Analysis

The experiments were implemented on Pytorch 1.6.0, Python 3.6.10, Tensorflow 1.14.0, and 18.04.1-Ubuntu operating systems. The improved StarGAN V2 in the experiment ran on the Intel(R)Xeon(R)CPU E5-2620V3@2.40 GHz in the CPU and NVIDIA GEFORCE GTX TITAN X graphics card in the GPU. In the experiment, the GPU was used to speed up the model's computation and decrease the training time.

All the models in our network were trained using the Adam optimization algorithm, the initial learning rate of the network was set to 1×10^{-4}, the batch size was 128, and each network was trained for 100 K iterations.

4.1. Experimental Dataset

To evaluate our network, this section will cover the experiments conducted on two public facial expression datasets, which were the extended Cohn-Kanade library (CK+) [52] dataset and the MMI dataset [53]. CK+ and MMI were captured in controlled lab environments. Since the experiment in our network was for static images, the three peaks of the expression changes in the video sequence ((CK+) expression library and MMI expression library) were taken as image samples, and all the images were scaled to 512 × 512. Among them, the entire number of CK+ images was 981, and the entire number of MMI pictures was 609. Tables 5 and 6 display the number of various emotion pictures in the CK+ and MMI datasets.

Table 5. The number of different emotion pictures in CK+ training set and test set.

CK+	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Train	90	36	118	50	138	56	166
Test	45	18	59	25	69	28	83

Table 6. The number of different emotion pictures in MMI training set and test set.

MMI	Anger	Disgust	Fear	Happy	Sadness	Surprise
Train	64	56	56	84	64	82
Test	32	28	28	42	32	41

4.2. Using Different GANs to Generate Different Samples on the CK+ Dataset

Figure 5 shows the original image (an angry expression in the CK+ data). Figure 6 shows the different facial expression images generated from the angry expressions in the CK+ data in each of the different networks. The first column (a) represents the sample generated by our network. The second column (b) represents the samples generated by *Att-StarGAN V2*. The third column (c) represents the samples created by *Hinge-StarGAN V2* and (d) represents the samples created by StarGAN V2. The *ATT-StarGAN V2* represents a combination of StarGAN V2 and SENet. The *Hinge-StarGAN V2* denotes a combination of StarGAN V2 and the hinge loss.

As shown in Figure 6, the images generated by our network still had some advantages over the images generated by the other three networks. The details of the expressions produced by our network were better performed. For example, when the facial expression was in the fear state, the images generated by our network were more realistic and vivid, and the expression details were processed appropriately. The images generated by our network appeared to be more vivid than the images generated by *ATT-StarGAN V2*.

The facial expression image details generated by *ATT-StarGAN V2* were suited. It can be seen that the detail characteristics of the generated images in column (b) are more perfect compared to column (d). The expression details of the images generated by *ATT-StarGAN V2* were better than the expression details of the images generated by *Hinge-StarGAN V2*. The images generated by our network had advantages in their realism and expression details over the images generated by StarGAN V2.

The facial expression images generated by *Hinge-StarGAN V2* were more realistic. The created pictures in column (c) are more realistic than those in column (d). Another example is that when the facial expression was surprised, it could better reflect the advantages of our network.

Figure 5. Angry expression in CK+ data.

Figure 6. CK+ dataset generated sample comparison chart ((**a**) represents the sample generated by our network, column (**b**) represents the samples generated by Att-StarGAN V2, (**c**) represents the samples created by Hinge-StarGAN V2, and (**d**) represents the samples created by StarGAN V2).

4.3. Using Different GANs to Generate Different Samples on the MMI Dataset

Figure 7 shows the original image (an angry expression in the MMI data). Figure 8 shows the different types of expressions generated from the angry expressions in the MMI data in each of the different networks. The first column (a) represents the samples generated by our network, the second column (b) represents the samples generated by *ATT-StarGAN V2*, column (c) represents the samples generated by *Hinge-StarGAN V2*, and column (d) represents the samples generated by StarGAN V2. In the MMI dataset, the facial expression amplitude changed greatly, so it was also possible to generate a poor image quality, such as sad expressions, as seen in Figure 8. Our network could generate better-quality facial expression images, but the facial expression images generated by the other three networks were deformed. In column (b) and column (d), the left eyes of the pictures in the sad row, created by *ATT-StarGAN V2* and StarGAN V2, respectively, are incomplete. The emotional details of the facial expression pictures were processed without processing the authenticity of the facial expression image, which proved that it was difficult to achieve the expected effect. Hinge loss could make the generated facial expression images more realistic. However, by adding the SENet, it was possible for the performance of the expression details in the generated facial expression images to be more complete and

vivid. For example, in the fear row, which shows the facial expression images generated by StarGAN V2, the eyebrows are deformed, and the quality of the images from the other three networks displays some progress.

Figure 7. Angry expression in MMI data.

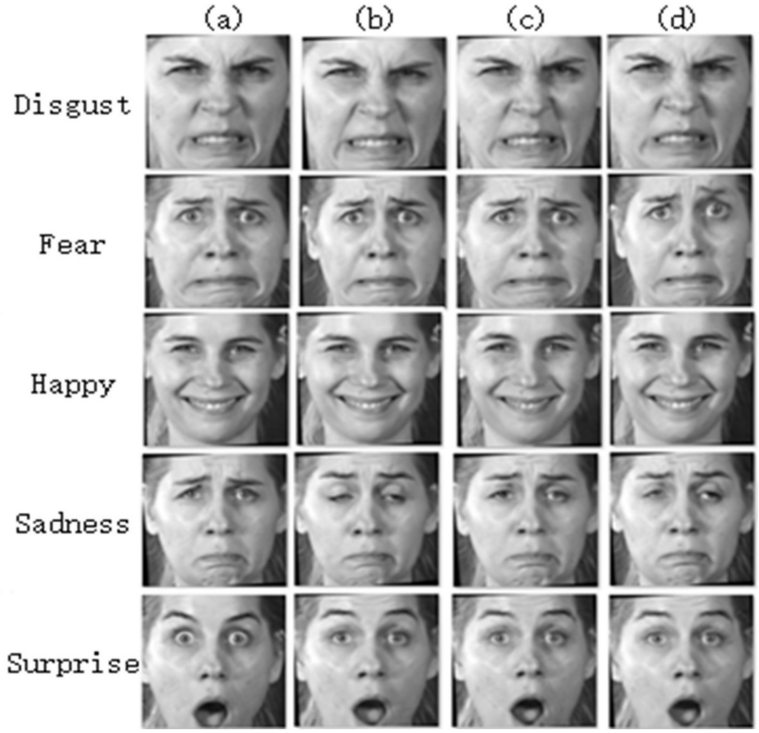

Figure 8. MMI-dataset-generated sample comparison chart ((**a**) represents the samples generated by our network, (**b**) represents the samples generated by ATT-StarGAN V2, column (**c**) represents the samples generated by Hinge-StarGAN V2, and column (**d**) represents the samples generated by StarGAN V2).

To verify the availability of our network, the generated image data and original data were trained and tested. The numbers of the various emotional pictures in the training set and test set are displayed in Tables 7 and 8.

Table 7. The number of different emotion pictures in CK+ training set and test set.

CK+	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Train	2070	2202	2002	2168	1953	2153	1885
Test	900	970	880	950	850	920	804

Table 8. The number of different emotion pictures in MMI training set and test set.

MMI	Anger	Disgust	Fear	Happy	Sadness	Surprise
Train	1260	1294	1285	1217	1266	1221
Test	540	552	549	519	540	522

In the experiment, the images were rescaled to 48 × 48 in the training and testing phases. We employed VGG19 for the training and testing. The training and testing phases were repeated 60 times, and we selected the best recognition rate as the final result.

4.4. Ablation Experiment

The ablation experiment outcomes are displayed in Table 9, and the recognition accuracy (%) is served as the performance evaluation.

Table 9. Ablation studies for key modules of our network on the CK+ and MMI databases.

Method	CK+	MMI
VGG19	96.8085	74.576
VGG19+StarGAN V2	97.7845	93.0788
VGG19+Hinge-StarGAN V2	98.6930	95.4687
VGG19+Middle layer + SENet + StarGAN V2	98.4699	93.9168
VGG19+SENet+Middle layer + StarGAN V2	98.9002	95.0652
VGG19+Att-StarGAN V2	99.173	95.2514
VGG19+ our network	99.2031	98.1378

As demonstrated in Table 9, our network could improve the facial expression recognition rate. Our network had a certain improvement effect compared to the other six networks.

In the original CK+ and MMI datasets, the accuracy of VGG19 was 96.8085% and 74.576%, respectively. The accuracy of VGG19 + StarGAN V2 was 97.7845% and 93.0788%, respectively. Compared to the original CK+ and MMI datasets, the accurate rates of recognition showed increases of 0.976% and 18.5028%, respectively, by using StarGAN V2. After adding the hinge loss into StarGAN V2, the accurate rates of recognition were 98.6930% and 95.4687% for the CK+ and MMI datasets, respectively. Compared to the StarGAN V2-enhanced CK+ and MMI datasets, the accurate rates of recognition showed increases of 0.9085% and 2.3899%, respectively, by using the *Hinge-StarGAN V2*-enhanced CK+ and MMI datasets.

After adding the SENet to StarGAN V2, the accurate rates of recognition were 99.173% and 95.2514% for the CK+ and MMI datasets, respectively. Compared to the StarGAN V2-enhanced CK+ and MMI datasets, the accurate rates of recognition showed increases of 1.3885% and 2.1726% by using the *ATT-StarGAN V2*-enhanced CK+ and MMI datasets. For the CK+ and MMI datasets that were enhanced by our network, the accurate rates of recognition were 99.2031% and 98.1378%, respectively. Compared to the StarGAN V2 enhanced-CK+ and MMI datasets, the accuracy showed increases of 1.4186% and 5.059% by using our enhanced network. Compared to the CK+ and MMI datasets with no enhancement, the recognition rates of our network improved by 2.3946% and 23.5618%, respectively.

Table 9 shows that the dataset was enhanced and that the expression recognition rates of the CK+ and MMI datasets were improved. Compared to the quality of the images generated by StarGAN V2, our network's generated image quality was also improved.

After the data enhancement, the recognition rate of the CK+ dataset was significantly improved. The main causes for this were that the images in the CK+ dataset consisted primarily of young men and women and that the characteristics were obvious. Therefore, compared to the original CK+ dataset, the recognition rate of the images generated by our network increased by 2.3946%. Compared to the original MMI dataset, the recognition rate of the images generated by our network increased by 23.5618%. The recognition rate of the MMI dataset was greatly improved by our enhanced network. There were two main reasons for this. On the one hand, it was due to the small number of samples in the original MMI dataset; on the other hand, it was due to age and facial occlusion. Thus, after the data enhancement, we solved the problems of insufficient data and unbalanced sample distributions, which was more conducive to network learning.

To further prove that the location of the SENet selection in the network was conducive to the quality of the network improvement, we placed the SENet at the front and back

of the middle layer (the middle layer is shown in Figure 3). The middle layer + SENet denotes that the middle layer was placed in front of the SENet. The SENet + middle layer denotes that the SENet was placed behind the middle layer. Table 9 shows the gap among the accuracy rates. The improvement effect of the SENet + middle layer was better than that of the middle layer + SENet, as it had a certain effect on improving the accuracy rate. It could focus on the original person's identity characteristics, so our network was based on the SENet + middle layer.

Our proposed model could effectively enhance facial expression images and generate different types of these facial expression images. The comparison results of the ablation experiment show that our proposed model could effectively improve the accuracy of facial expression recognition. Therefore, our proposed model could effectively enhance facial expression images, which is more conducive to improving the correct recognition rate for facial expression images.

4.5. The Score FID Different Models on CK+ and MMI Database

To further imply the quality of the generated pictures, we utilized Fréchet inception distance (FID) [54] (lower is better) as the evaluation indicator to measure the visual quality. FID is a common metric for evaluating pictures created by GANs. It conveys the quality and diversity of the generated images by comparing the feature vectors among the different images. The results of the paper's comparison are displayed in Table 10. The *ATT-StarGAN V2* represents the combination of StarGAN V2 and the SENet. The *Hinge-StarGAN V2* denotes the combination of StarGAN V2 and the hinge loss. The results show that the SENet- and hinge-loss-improved StarGAN V2 achieved the best outcomes. Compared to StarGAN V2, our method effectively improved the quality of the created pictures.

Table 10. The score of FID by different models on CK+ and MMI database.

Method	CK+	MMI
our network	31.6228	26.2882
Att-StarGAN V2	36.9607	27.9682
Hinge-StarGAN V2	76.1028	29.4659
StarGAN V2	83.0125	31.54335

4.6. Comparison with Other Works

The comparison results with other works are shown in Table 11:

Table 11. Recognition accuracy (%) of the proposed network and comparison with previous work.

Method	CK+	MMI
ELBP+PHOG+SVM [11]	95.33	-
HOG+VGGFace [12]	98.12	-
Ref. [13]	-	94.98
HOG, LBP+SVM [14]	99.18	-
Ref. [15]	98.60	78.44
Ref. [16]	99.16	83.67
Ref. [36]	97.83	-
VGG19 + SVM [37]	98.76	-
HistNet [38]	98.47	83.41
FMN [55]	98.61	81.39
DLP-CNN [56]	95.78	78.46
IPA2LT [57]	92.45	65.61
Ref. [58]	96.12	-
PAU-Net (type I) [59]	-	85.42
PAU-Net (type II) [59]	-	85.89
DSN-DF [60]	98.9	79.33
VGG19+ our network	99.2031	98.1378

As shown in Table 11, compared to other methods, our network had noticeable, obvious advantages with regard to the MMI dataset. Since there were many different characters in the MMI dataset and the expression amplitudes were quite different, it was difficult to accurately identify the facial expressions. Our network could generate more samples so that multiple sample expressions could be learned and the facial expression recognition rate could be better improved. After the data enhancement, the recognition rate of the MMI dataset was greatly promoted. Compared to the traditional feature extraction method, Refs. [11–14] used the hybrid feature extraction method, which significantly promoted the recognition accuracy. Due to the limitations of traditional methods, the extracted feature information was not comprehensive enough and it was necessary to add auxiliary measures to improve the recognition accuracy. In [15,36–38,55,58–60], deep learning methods were used to conduct global learning, local learning, or a combination of global and local learning. However, due to the small number of samples and the unbalanced sample distributions in the MMI and CK+ datasets, redundant learning and repetitive learning occurred from time to time. The accuracy of the recognition rate was not as accurate as that of our network. Our network could generate many samples, which effectively improved the recognition rate. In [16,56,57], the problem of facial expression image recognition across datasets was solved, but due to the fusion of multiple datasets, the difficulty of the recognition increased.

5. Conclusions

Facial expression datasets often contain insufficient data and unbalanced sample distributions. This article constructed, implemented, and demonstrated an improved StarGAN V2 model for a facial expression data enhancement. Firstly, we used StarGAN V2 to generate different facial expression images and enrich the expression dataset. Secondly, the SENet paid more attention to the vital regions of the images and improved the vividness of the created pictures. The model was integrated into the generator in StarGAN V2 to improve the quality of the generated images. Lastly, we introduced the hinge loss to StarGAN V2 to distinguish between the real or fake samples and improve the authenticity of the generated images. The outcomes of the two public CK+ and MMI datasets showed the effectiveness of our method. The improved StarGAN V2 was conducive to improving the accuracy of the network recognition, which could retain the identity information and transform the different styles.

Our proposed network was compared to previous studies from the LR. Compared to [61], our network was based on StarGAN V2. The network generated diverse images within a model and retained the original identity information. Compared to pix2pix [62], our network did not require the training of image pairs. Our network products had more freedom in generating different image styles.

The advantages of our proposed model are as follows:

(1) We propose a new facial expression image generation model. The model can generate different facial expression images by applying a simple structure.
(2) Our network is an effective model, avoiding the training of redundant models and saving a lot of time and resources.
(3) The generated facial expression images maintain the identity information of the original input image. Additionally, our model improves the recognition rate of the facial expression images.

The weaknesses of our proposed model are shown below:

(1) There is a lot of text and speech information in the expression dataset. We only enhance the images.
(2) In addition to inputting the original image, the generator also inputs the corresponding stylistic features. The amount of input into the generator should be minimized while maintaining the quality of the generated image. This reduces the computational effort of the model.

(3) We only enhance the facial expression images, without corresponding transformations for hairstyles and clothing, etc.

With the emergence of multimedia technology, facial expression recognition is not only limited to images, but also involves many other aspects, such as sound and text. Methods of integrating expression recognition into multimedia technology, in order to make it more conducive to research on expression recognition, will be the focus of our future research.

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

Funding: This work was supported in part by National Natural Science Foundation of China under Grant 62176084, and Grant 62176083, and in part by the Fundamental Research Funds for the Central Universities of China under Grant PA2022GDSK0066 and PA2022GDSK0068.

Data Availability Statement: Not applicable.

Acknowledgments: We acknowledge the use of the equipment provided by the Hefei University of Technology.

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

References

1. Yang, J.Q.; Chen, C.H.; Li, J.Y.; Liu, D.; Li, T.; Zhan, Z.H. Compressed-encoding particle swarm optimization with fuzzy learning for large-scale feature selection. *Symmetry* **2022**, *14*, 1142. [CrossRef]
2. Tang, Y.; Pedrycz, W. Oscillation-bound estimation of perturbations under Bandler-Kohout subproduct. *IEEE Trans. Cybern.* **2022**, *52*, 6269–6282. [CrossRef] [PubMed]
3. Tang, Y.; Pedrycz, W.; Ren, F. Granular symmetric implicational method. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, *6*, 710–723. [CrossRef]
4. Poux, D.; Allaert, B.; Ihaddadene, N.; Bilasco, I.M.; Djeraba, C.; Bennamoun, M. Dynamic facial expression recognition under partial occlusion with optical flow reconstruction. *IEEE Trans. Image Process.* **2022**, *31*, 446–457. [CrossRef] [PubMed]
5. Tang, Y.; Ren, F.; Pedrycz, W. Fuzzy c-means clustering through SSIM and patch for image segmentation. *Appl. Soft Comput.* **2020**, *87*, 105928. [CrossRef]
6. Han, B.; Hu, M.; Wang, X.; Ren, F. A Triple-Structure Network Model Based upon MobileNet V1 and Multi-Loss Function for Facial Expression Recognition. *Symmetry* **2022**, *14*, 2055. [CrossRef]
7. Tang, Y.; Pan, Z.; Pedrycz, W.; Ren, F.; Song, X. Viewpoint-based kernel fuzzy clustering with weight information granules. *IEEE Trans. Emerg. Top. Comput. Intell.* **2023**, *7*, 342–356. [CrossRef]
8. Tang, Y.; Zhang, L.; Bao, G.; Ren, F.J.; Pedrycz, W. Symmetric implicational algorithm derived from intuitionistic fuzzy entropy. *Iran. J. Fuzzy Syst.* **2022**, *19*, 27–44.
9. Sujanaa, J.; Palanivel, S.; Balasubramanian, M. Emotion recognition using support vector machine and one-dimensional convolutional neural network. *Multimed. Tools Appl.* **2021**, *80*, 27171–27185. [CrossRef]
10. Liu, Y.; Fu, G. Emotion recognition by deeply learned multi-channel textual and EEG features. *Future Gener. Comput. Syst.* **2021**, *119*, 1–6. [CrossRef]
11. Harifnejad, M.; Shahbahrami, A.; Akoushideh, A.; Hassanpour, R.Z. Facial expression recognition using a combination of enhanced local binary pattern and pyramid histogram of oriented gradients features extraction. Image Processing. *IET* **2020**, *15*, 468–478.
12. Ahadit, A.B.; Jatoth, R.K. A novel multi-feature fusion deep neural network using HOG and VGG-Face for facial expression classification. *Mach. Vis. Appl.* **2022**, *33*, 55. [CrossRef]
13. Shanthi, P.; Nickolas, S. Facial landmark detection and geometric feature-based emotion recognition. *Int. J. Biom.* **2022**, *14*, 138–154. [CrossRef]
14. Santosh, M.; Sharma, A. Fusion of multi representation and multi descriptors for facial expression recognition. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1057*, 012093. [CrossRef]
15. Wang, S.M.; Shuai, H.; Liu, Q.S. Facial expression recognition based on deep facial landmark features. *J. Image Graph.* **2020**, *25*, 813–823.
16. Ruan, D.; Yan, Y.; Chen, S.; Xue, J.; Wang, H. Deep disturbance-disentangled learning for facial expression recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2833–2841.
17. Sekaran, S.A.P.R.; Lee, C.P.; Lim, K.M. Facial emotion recognition using transfer learning of AlexNet. In Proceedings of the 2021 9th International Conference on Information and Communication Technology (ICoICT), Virtual, 3–5 August 2021; pp. 170–174.

18. Kansizoglou, I.; Bampis, L.; Gasteratos, A. An active learning paradigm for online audio-visual emotion recognition. *IEEE Trans. Affect. Comput.* **2019**, *13*, 756–768. [CrossRef]
19. 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 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
20. Li, B.; Lima, D. Facial expression recognition via ResNet-50. *Int. J. Cogn. Comput. Eng.* **2021**, *2*, 57–64. [CrossRef]
21. Kansizoglou, I.; Bampis, L.; Gasteratos, A. Deep feature space: A geometrical perspective. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6823–6838. [CrossRef]
22. Sang, D.V.; Ha, P.T. Discriminative deep feature learning for facial emotion recognition. In Proceedings of the 2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR), Ho Chi Minh City, Vietnam, 5–6 April 2018; pp. 1–6.
23. Zhou, X.; Zhou, J.; Xu, R. New algorithm for face recognition based on the combination of multi-sample conventional collaborative and inverse linear regression. *J. Electron. Meas. Instrum.* **2018**, *32*, 96–101.
24. Li, W.; Li, M.; Su, Z.; Zhu, Z. A deep-learning approach to facial expression recognition with candid images. In Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 18–22 May 2015; pp. 279–282.
25. Tripathi, R.K. Adaptive geometric filtering based on average brightness of the image and discrete cosine transform coefficient adjustment for gray and color image enhancement. *Arab. J. Sci. Eng.* **2020**, *45*, 1655–1668.
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *arXiv* **2014**.
27. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.W. StarGAN V2: Diverse Image Synthesis for Multiple Domains. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8188–8197.
28. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
29. Liu, D.; Ouyang, X.; Xu, S.; Zhou, P.; He, K.; Wen, S. SAANet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing* **2020**, *413*, 145–157. [CrossRef]
30. Cheng, J.; Liang, R.; Liang, Z.; Zhao, L.; Huang, C.; Schuller, B. A deep adaptation network for speech enhancement: Combining a relativistic discriminator with multi-kernel maximum mean discrepancy. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *29*, 41–53. [CrossRef]
31. Saurav, S.; Singh, S.; Saini, R.; Yadav, M. Facial expression recognition using improved adaptive local ternary pattern. In *Proceedings of the 3rd International Conference on Computer Vision and Image Processing, Macau, China, 23–25 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 39–52.
32. Niu, B.; Gao, Z.; Guo, B. Facial expression recognition with LBP and ORB features. *Comput. Intell. Neurosci.* **2021**, *2021*, 8828245. [CrossRef] [PubMed]
33. Lu, F.; Zhang, L.; Tian, G. User Emotion Recognition Method Based on Facial Expression and Speech Signal Fusion. In Proceedings of the 2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA), Chengdu, China, 1–4 August 2021; pp. 1121–1126.
34. Arora, M.; Kumar, M. AutoFER: PCA and PSO based automatic facial emotion recognition. *Multimed. Tools Appl.* **2021**, *80*, 3039–3049. [CrossRef]
35. Islam, B.; Mahmud, F.; Hossain, A.; Goala, P.B.; Mia, S. A facial region segmentation-based approach to recognize human emotion using fusion of HOG & LBP features and artificial neural network. In Proceedings of the 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, 13–15 September 2018; pp. 642–646.
36. Bisogni, C.; Castiglione, A.; Hossain, S.; Narducci, F.; Umer, S. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Trans. Ind. Inform.* **2022**, *18*, 5619–5627. [CrossRef]
37. Naim, S.; Chaibi, H.; Abdessamad, E.R.; Saadane, R.; Chehri, A. A Hybrid Automatic Facial Expression Recognition Based on Convolutional Neuronal Networks and Support Vector Machines Techniques. In *Human Centred Intelligent Systems*; Smart Innovation, Systems and Technologies; Zimmermann, A., Howlett, R.J., Jain, L.C., Eds.; Springer: Singapore, 2022; p. 310.
38. Sadeghi, H.; Raie, A.A. HistNet: Histogram-based convolutional neural network with Chi-squared deep metric learning for facial expression recognition. *Inf. Sci.* **2022**, *608*, 472–488. [CrossRef]
39. Sarkar, K.; Halder, T.K.; Mandal, A. Adaptive power-law and cdf based geometric transformation for low contrast image enhancement. *Multimed. Tools Appl.* **2021**, *80*, 6329–6353. [CrossRef]
40. Aquino-Morínigo, P.B.; Lugo-Solís, F.R.; Pinto-Roa, D.P.; Legal-Ayala, H.A.; Noguera, J.L. Bi-histogram equalization using two plateau limits. *Signal Image Video Process.* **2017**, *11*, 857–864. [CrossRef]
41. Xin, M.; Zhou, Y.; Yan, J. Single Sample Face Recognition using LGBP and Locality Preserving Discriminant Analysis. *Appl. Math. Inf. Sci.* **2015**, *9*, 353–360. [CrossRef]
42. Ramasubramanian, M.; Rangaswamy, M.; Prabha, H.; Dilipan, A. 3D Facial Model Construction and Expressions from a Single Face Image. *Artif. Intell. Syst. Mach. Learn.* **2014**, *6*, 274–277.
43. Zarie, M.; Hajghassem, H.; Eslami Majd, A. Contrast enhancement using triple dynamic clipped histogram equalization based on mean or median. *Optik* **2018**, *175*, 126–137. [CrossRef]
44. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.

45. Fu, B.; Li, F.; Niu, Y.; Wu, H.; Li, Y.; Shi, G. Conditional generative adversarial network for EEG-based emotion fine-grained estimation and visualization. *J. Vis. Commun. Image Represent.* **2021**, *74*, 102982. [CrossRef]
46. Zhu, J.; 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; pp. 2223–2232.
47. Dou, H.; Chen, C.; Hu, X.; Peng, S. Asymmetric Cyclegan for Unpaired NIR-to-RGB Face Image Translation. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 1757–1761.
48. Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; Choo, J. StarGAN: Unifed 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.
49. Yang, F.; Wang, Y.; Herranz, L.; Cheng, Y.; Mikhail, M.G. A novel framework for image-to-image translation and image compression. *Neurocomputing* **2022**, *508*, 58–70. [CrossRef]
50. Chen, Y.; Zhang, Z.; Zhong, L.; Chen, T.; Chen, J.; Yu, Y. Three-Stream Convolutional Neural Network with Squeeze-and-Excitation Block for Near-Infrared Facial Expression Recognition. *Electronics* **2019**, *8*, 385. [CrossRef]
51. Nguyen, T.T.; Le, T.H. Fusion of Attentional and Traditional Convolutional Networks for Facial Expression Recognition. *EAI Endorsed Trans. Pervasive Health Technol.* **2021**, *7*, e2. [CrossRef]
52. 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 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
53. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands, 6 July 2005; p. 5.
54. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Ochreiter, S. Gans trained by a two time-scale update rule converge to a local Nash equilibrium. *Proc. Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6626–6637.
55. Haghpanah, M.A.; Saeedizade, E.; Masouleh, M.T.; Kalhor, A. Real-Time Facial Expression Recognition using Facial Landmarks and Neural Networks. In Proceedings of the 2022 International Conference on Machine Vision and Image Processing (MVIP), Ahvaz, Iran, 22–24 February 2022; pp. 1–7.
56. Zhou, J.; Zhang, X.; Lin, Y.; Liu, Y. Facial expression recognition using frequency multiplication network with uniform rectangular features. *J. Vis. Commun. Image Represent.* **2021**, *75*, 103018. [CrossRef]
57. Shan, L.; Deng, W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. Image Process.* **2018**, *28*, 356–370.
58. Zeng, J.; Shan, S.; Chen, X. Facial expression recognition with inconsistently annotated datasets. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 222–237.
59. Wang, X.; Zhang, T.; Chen, C.P. PAU-Net: Privileged Action Unit Network for Facial Expression Recognition. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 8. [CrossRef]
60. Gan, C.; Yao, J.; Ma, S.; Zhang, Z.; Zhu, L. The deep spatiotemporal network with dual-flow fusion for video-oriented facial expression recognition. *Digit. Commun. Netw.* **2022**, *in press*. [CrossRef]
61. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
62. Isola, P.; Zhu, J.Y.; Zhou, T.H.; Efros, A.A. Image- to- image translation with conditional adversarial networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; IEEE Computer Society: Washington, DC, USA, 2017; pp. 5967–5976.