StarGAN v2 Enhancement : Tackling Noise And Identity PreservationFacial Augmentation

Prof. Jeebanand Panda
Dept of Electronics
DTU

Aakash Gaur
Dept of Electronics
DTU

Abhinav Mishra
Dept of Electronics
DTU

Bhawna Gautam
Dept of Electronics
DTU

***Abstract*—Generative Adversarial Networks (GANs) have been extensively studied for transferring human facial expressions. However, the StarGAN v2[1] model faces issues with reference-based expression transfer when trained on noisy datasets, leading to identity loss and the morphing of faces from the original to the reference image. Noisy datasets in this context refer to the high variance in emotion depiction across different individuals, which is common in real scenarios. To address these challenges, we propose an improved StarGAN v2[1] model that includes an additional loss parameter and flow modifications to produce higher-quality images, allowing for arbitrary expression translation. Our approach was tested on a subset of the AffectNet dataset, and we compared the visual output with the baseline model and performed a quantitative analysis to show the effectiveness of our model.**

Keywords: Generative Adversarial Networks, Facial Emotion Transfer, StarGAN, Siamese Network, CNN

I. Introduction

The goal of multi-domain image-to-image transfer is to convert images (human faces) from one arbitrary domain (expression) to various target domains, such as changing a facial expression from happy to sad. Emotion augmentation has numerous applications in entertainment, education, and healthcare. However, a major challenge in face image synthesis is preserving the identity of the source face while transferring the desired attributes or emotions to the target face, especially when there are significant differences in pose, expression, illumination, or occlusion between the source and target faces.

Various methods have been proposed to address this issue using GANs, which consist of a generator that aims to produce realistic and diverse images and a discriminator that distinguishes between real and fake images. StarGAN v2[1], a recent successful method, can synthesize high-quality face images from multiple domains using a single generator and a single discriminator. It uses a style encoder to extract latent codes from reference images that capture domain-specific features like hair color, gender, or emotion, maps these codes to a shared latent space, and finally uses a generator to produce the target images conditioned on the source images and style codes.

Despite its success, StarGAN v2[1] has limitations that affect its performance, particularly in identity preservation between the source and target images, leading to unwanted changes in facial features or loss of identity information. This paper proposes an enhanced face image synthesis method that addresses these limitations and improves the quality of the generated images.

This paper proposes an enhanced method for face image synthesis that addresses these limitations and improves the quality of generated images. Our approach introduces an additional loss term that explicitly measures identity similarity and incorporates flow modifications to refine image generation, thereby enabling arbitrary expression translation while maintaining the identity of the source face.

By enhancing StarGAN v2[1] with these modifications, we aim to produce more accurate and identity-preserving facial augmentations, which are crucial for practical applications in various domains where identity integrity is paramount.

Main Contributions:

1) : Introduced an additional loss term in StarGAN v2's objective function to measure identity similarity between the source and target images using a pre-trained face recognition network. This encourages the generator to maintain the identity of the source face while transferring the desired attributes or emotions to the target face.

2) : Added facial identity embeddings to the generator as input, providing a specific target for the additional loss parameter. These identity embeddings were generated using Google XceptionNET[2] with additional fully connected layers.

We evaluate our proposed method on several face image datasets and compare it with StarGAN v2[1] and other state-of-the-art methods. We show that our method can generate more realistic and diverse face images that preserve the identity of the source face while transferring arbitrary attributes or emotions.

II. Related Works

Generative adversarial networks (GANs) are a powerful tool for image synthesis and manipulation, consisting of a generator that creates realistic and varied images and a discriminator that differentiates between genuine and synthetic images. Due to their capability to produce high-quality outputs, GANs have been extensively used in facial emotion transfer and other image-to-image translation tasks.

A prominent method for facial emotion transfer is StarGAN[6], which achieves one-to-many translations for multiple emotions using a single generator and discriminator. StarGAN[6] employs a conditional adversarial loss and a cycle-consistency loss to train the network, ensuring that the generated images are both realistic and consistent with the input images. Additionally, it uses an auxiliary classifier to control the generated images' emotions. However, the reliance on cycle-consistency loss can lead to noise and artifacts in the generated images, particularly when the input images contain noise.

Several enhancements to StarGAN[6] have been proposed to handle noisy data better. For example, LAUN improved StarGAN[8] by utilizing an L2 loss function instead of an L1 loss function for reconstruction loss. This adjustment helps to better magnify the discrepancies between real and generated images, leading to faster convergence. LAUN also introduced Contextual Loss, which measures image similarity based on context similarity using the cosine distance between feature map points extracted by VGG19. Despite these improvements, LAUN's method does not explicitly consider facial identity similarity, which is crucial for preserving the identity of the input face while transferring the desired emotion to the output face. This omission can result in outputs with distorted or unrecognizable faces.

StarGAN v2[1] enhances its predecessor by employing a single generator and discriminator to produce high-quality images across multiple domains. It includes a style encoder to extract latent codes from reference images, representing domain-specific features such as hair color, gender, or emotion. These latent codes are mapped to a shared latent space by a mapping network, which controls the style of the generated images. The generator then creates the target images based on the source images and style codes. However, StarGAN v2[1] does not explicitly enforce identity preservation between the input and output images, leading to unwanted changes in facial features or identity loss, particularly when dealing with noisy data.

To address these challenges, recent research has explored various strategies to enhance identity preservation in GAN-based facial emotion transfer. For instance, FA-GAN[4] employs identity-attribute disentanglement to create deformation-invariant face images. It introduces an attribute encoder that extracts attribute embeddings from face images and feeds them to the generator alongside identity embeddings, highlighting the importance of disentangling identity and attribute features for better identity preservation.

Moreover, the use of Siamese networks for face recognition has gained popularity. Siamese networks consist of two identical branches that share the same weights and parameters, taking two images as input and outputting a scalar value indicating their similarity. The triplet loss function, commonly used in Siamese networks, aims to minimize the distance between an anchor image and a positive image (same identity) while maximizing the distance between the anchor image and a negative image (different identity). Inspired by FaceNet[2], this approach enables the training of models that can learn discriminative features for identity preservation during facial transformations.

Furthermore, other works have integrated identity-preserving techniques into GAN frameworks. For example, integrating face recognition networks to assess identity similarity and incorporating additional loss functions to guide the generator in maintaining identity have shown promising results. These methods emphasize the significance of preserving identity while achieving high-quality emotion transfer, addressing the shortcomings of previous approaches.

In this paper, we propose a novel method that integrates an additional loss term similar to FA-GAN[4]'s identity-attribute disentanglement and incorporates ideas from Siamese networks with triplet loss. By combining these approaches with StarGAN v2[1]'s architecture, we aim to enhance identity preservation and improve the overall quality of generated images, especially in noisy datasets. Our method leverages the strengths of existing techniques while addressing their limitations, achieving more accurate and reliable facial emotion transfers.

By building on the foundational concepts of GANs, StarGAN, and recent advancements in identity preservation, our approach offers a comprehensive solution for high-fidelity facial emotion transfer. This integration of multiple strategies allows us to maintain the identity integrity of the source face while effectively transferring the desired attributes or emotions, providing a significant improvement over existing methods.

.

III. PROPOSED METHOD

A. Data Preparation

The AffectNet dataset is a large-scale facial expression recognition dataset that includes approximately 450,000 images of human faces, annotated with eight distinct emotion labels: neutral, happiness, sadness, surprise, contempt, fear, disgust, and anger. For our purposes, we used a subset of AffectNet, available on Kaggle, which contains around 42,000 images.

However, this dataset also contains many images that are mismatched with their labels, such as smiling faces labeled as sadness or neutral faces labeled as anger. This inconsistency can degrade the performance of models trained on such data. To improve the quality of the dataset, we applied a convolutional neural network (CNN) model to filter out these mismatched images based on their facial features and expressions. The CNN model we used is named Facial Emotion Recognition (FER) and is based on the VGG19 architecture, trained on the FER2013 dataset. The model achieved an accuracy of 70% on the validation set.

We then utilized this CNN model to predict the labels of the remaining images in AffectNet and removed those that had a low confidence score or a different label from the original one. This resulted in a refined AffectNet dataset with 19,628 images that are more consistent and reliable for facial expression recognition tasks. Although the dataset may seem small at first glance, the input is a combination of two images from different domains, leading to a total dataset size of 21,250,000 image pairs.

To further ensure the quality of the dataset, we conducted manual verification on a subset of images to confirm the accuracy of the labels assigned by the CNN model. This step helped us eliminate any remaining mislabeled images and provided a higher-quality dataset for training our model.Additionally, data augmentation techniques were applied to enhance the diversity of the dataset. These techniques included random rotations, scaling, horizontal flipping, and color adjustments, which helped to simulate various real-world scenarios and improve the robustness of our model. By augmenting the dataset, we aimed to increase the model's ability to generalize across different expressions and conditions.

B. *StarGAN version 2*

StarGAN v2 aims to train a single generator to produce diverse images of each domain that correspond to the input image. It consists of a Generator, Mapping Network, Style Encoder, and Discriminator. The losses used for training include Adversarial Loss, Style Reconstruction Loss, Diversification Loss, and Cycle Consistency Loss. The training objective minimizes these losses with hyperparameters to optimize the network and encoders.

Losses used: The following losses are used for training the Generator and Discriminator:

1) Adversarial loss:

$$\mathcal{L}_{adv} = E_{x,y}[logD_y(x)] + E_{x,\tilde{y},z}[1 - log(1D_{\tilde{y}}(G(x,\tilde{s})))]$$

where $D_y(x)$ is the D output corresponding to x. The mapping network F learns to supply the style code s̃ that is probably present in the target domain ye, and G learns to make use of es and produce an image $G(x,$ s̃) that is identical to actual pictures of the domain y.

2) Style reconstruction loss:

$$\mathcal{L}_{sty} = E_{x,\tilde{y},z}[||\tilde{s} - E_{\tilde{y}}(G(x,\tilde{s}))||_1]$$

They use many encoders to figure out how to translate a picture to its latent code. The key distinction is that we train a single encoder E to promote a variety of outputs across many domains.

3) Diversification Loss:

$$\mathcal{L}_{ds} = E_{x,\tilde{y},z1,z2}[||G(x,\tilde{s}_1) - G(x,\tilde{s}_2)||_1]$$

$s^{\tilde{}}_i = F_{y^{\tilde{}}}(z_i)$ for i $\in$ 1,2, where E produces the target style codes $s^{\tilde{}}_1$ and $s^{\tilde{}}_2$ conditioned on two random latent codes z_1 and z_2. In order to produce a variety of pictures, maximization of the regularization term requires G to sift through the image space and identify significant style traits. It should be noted that in the original version, the slight difference z_1 z_2 $_1$ in the denominator considerably raises the loss, causi||ng the training to be unstable owing to high gradients.

4) Cycle consistency loss:

$$\mathcal{L}_{cyc} = E_{x,y,\tilde{y},z}[||x - G(G(x,\tilde{s}),\hat{s})||_1]$$

where y is the original domain of x and $s = E_y(x)$ is the predicted style code of the input picture x. The generator G learns to maintain the original properties of the input picture x while faithfully modifying its style by being prompted to reconstruct the image with the estimated style codes.

Net equation for training:

$$\min_{(G,F,E)} \max_{D} [\mathcal{L}_{adv} + \lambda_{sty}\mathcal{L}_{sty} - \lambda_{ds}\mathcal{L}_{ds} + \lambda_{cyc}\mathcal{L}_{cyc}]$$

where λs being the hyperparameters and the Ls being the losses of above mentioned network and encoders.

C. *Identity Loss*

As discussed, one of the drawbacks of StarGAN v2[1] is that it fails to preserve the identity of the person whose face is being transformed by the generative adversarial network (GAN). This can be observed in Fig. 1, where we show some examples of the output images produced by the original StarGAN v2[1] on our dataset.

To overcome this limitation, we propose a novel method that incorporates an additional loss term, called Identity Loss and some modifications to the network flow. The main idea of this loss term is to ensure that the identity of the source image, i.e., the face that is being modified, is not changed or distorted by the GAN. To achieve this, we compare the generated image from GAN with the original image and penalize any differences that affect identity recognition.

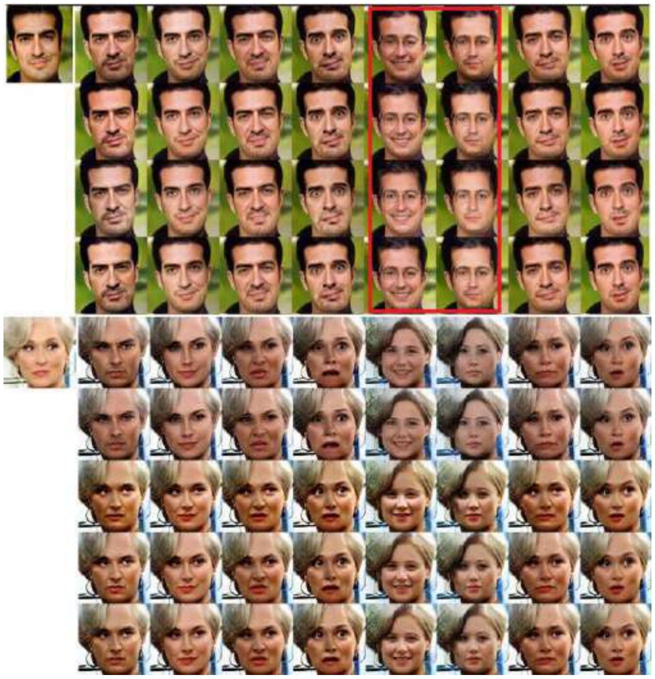

Fig. 1

Visible artifacts and loss of identity in StarGAN **v2[1]** results on noisy dataset

Our proposed method consists of two stages: feature extraction and similarity measurement. In the first stage, we use the Xception Network as the backbone to extract high-level features from the input images. The Xception Network is a deep convolutional neural network that consists of several modules of depthwise separable convolutions and residual connections.We add some fully connected layers on top of the Xception Network to obtain a fixed-length feature vector for each image.

In the second stage, we use the concept of Siamese network[5]

with the triplet loss function to measure the similarity between the feature vectors of different images. A Siamese network[5] is a network that has two identical branches that share the same weights and parameters. It takes two images as input and outputs a scalar value that indicates how similar or dissimilar they are. The triplet loss function is a loss function that aims to minimize the distance between an anchor image and a positive image (belonging to the same class or identity) and maximize the distance between the anchor image and a negative image (belonging to a different class or identity). The triplet loss function is defined as follows :

$$\mathcal{L} = \min \sum^{N} [||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha]_+$$

where *f* (*x*) denotes the output embeddings after an image passes through the network and the superscripts *a*,*p* and *n* are anchor, positive and negative respectively.*α* is a margin that is enforced between positive and negative pairs.

The triplet loss function is inspired by the FaceNet[2] paper, which proposed a method for face recognition and verification based on embedding faces into a low-dimensional space. By using the Siamese network[5] with the triplet loss function, we can train our sub-model to learn discriminative features that can preserve the identity of the faces while transforming their emotions or poses.

Therefore our net loss equation changes to:

$$\min_{(G,F,E)} \max_{D} [\mathcal{L}_{adv} + \lambda_{sty}\mathcal{L}_{sty} - \lambda_{ds}\mathcal{L}_{ds} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{\text{id}}\mathcal{L}_{\text{id}}]$$

where $\mathcal{L}_{id}$ is given by $\sum ||f(x_i^{\text{real}}) - f(x_i^{\text{generated}})||_2^2$

To accompany our loss we introduce flow changes to the StarGAN v2[1], which are shown in Fig. 2.

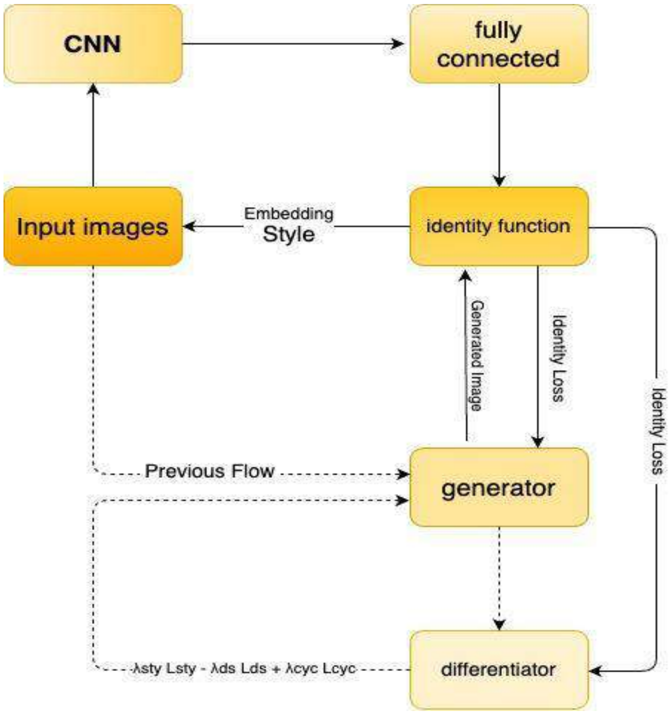

Fig. 2
Flow Changes, all the orange colored arrows depict the changes

The main changes are as follows-

1) Identity embeddings are added in the input image.
2) Identity loss function is calculated with the generated and the input image and given as feedback to both the discriminator and generator so as to preserve the identity of the source image.

I. TRAINING AND TESTING

A. Training Siamese Network

We trained a Siamese network with a triplet loss function[2] for face recognition. We added fully connected layers on top of the Xception model to reduce the dimensionality of the feature vectors. We used the LFW dataset, which contains more than 13,000 images of faces collected from the web. We followed the FaceNet[2] paper by Schroff et al., 2015, which proposed a method for face recognition and verification based on embedding faces into a low-dimensional space.

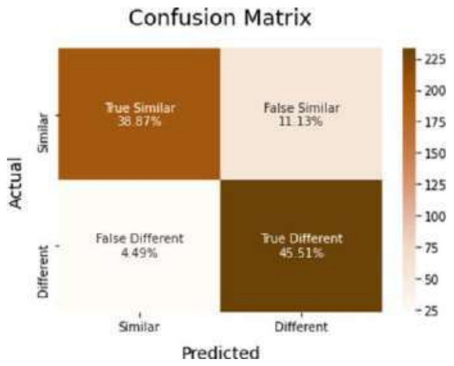

Fig. 3
Confusion matrix for testing of this Siamese Network

This training let us achieve 84.375% accuracy in identifying images of the same person which is more than enough for our purposes of introducing identity loss parameter in the StarGAN **v2[1]** network.

B. Training StarGANv2 + Identity Loss

The highlights of our training the StarGAN **v2[1]** are -

- Initially, the model generated random noisy images, hence the importance of identity doesn't have much significance. Therefore, we start the training with λ_{id} close to 0 so as to give more importance to the generation of valid human faces.
- As the model had been sufficiently trained to generate recognizable human faces with emotion translation, we increased our bias towards identity preservation.
- λ_{id} starts with the value of 0.1 till 4 with linear uniform steps after 25 thousand iterations.

TABLE I
Comparison of different methods on Similarity Distance

Method	Similarity Distance
FA-GAN	0.20428795
LAUN improved StarGAN	0.4089777
Ours	**0.22002675**

We also tested our model for emotion accuracy using the VGG16 and KDEF dataset[9], as mentioned in the LAUN improved StarGAN paper. We start by augmenting images of KDEF dataset using our method and then training VGG16 over it and comparing the accuracy of emotion detection, shown in Table 2.

TABLE II
Comparison of different methods on emotion detection accuracy

Method	Detection Accuracy
VGG16	93.78
VGG16 + StarGAN	94.00
VGG16 + LAUN improved StarGAN	95.97
Ours	**96.38**

Fig 4. shows some generated image results of different models along with ours for visual comparison.

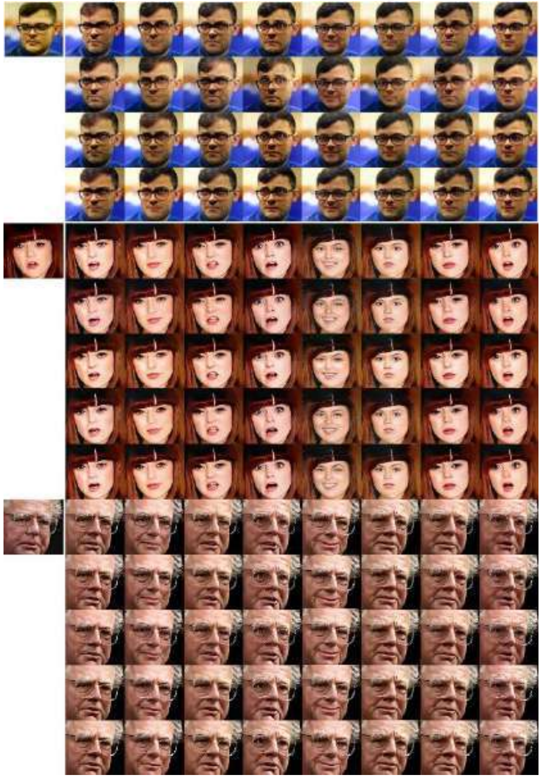

Fig. 4
Some generated results of our proposed model

II. Conclusion

This paper presents a novel approach to improving StarGAN **v2**[1] model for emotion transfer in noisy datasets,by incorporating additional loss term for improving identity preservation. The results show that our proposed model performs comparable to or better than other existing models, especially in the case of noisy datasets.

References

[1] Y. Choi et al., “StarGAN v2: Diverse Image Synthesis for Multiple Domains,” in IEEE Conference on Computer Vision and Pattern Reco nition (CVPR), 2020.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[3] Y. Li, X. Liu, H. Zhang, and Z. Guo, “Heterogeneous Face Recognition via Face Synthesis with Identity-Attribute Disentanglement,” IEEE Transactions on Image Processing, vol. 30, pp. 2066-2078, 2021.

[4] S. Wang, Y. Fu, H. Zhang, and X. Liu, “FA-GAN: Face Augmentation GAN for Deformation-Invariant Face Recognition,” IEEE Transactions on Image Processing, vol. 29, pp. 9070-9083, 2020.

[5] G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese Neural Networks for One-shot Image Recognition,” in ICML deep learning workshop, vol. 2, 2015.

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[7] A. Banerjee and D. Kollias, “Emotion Generation and Recognition: A StarGAN Approach,” arXiv preprint arXiv:1910.11090, 2019.

[8] X. Wang, J. Gong, M. Hu, Y. Gu and F. Ren, ”LAUN Improved StarGAN for Facial Emotion Recognition,” in IEEE Access, vol. 8, pp. 161509- 161518, 2020, doi: 10.1109/ACCESS.2020.3021531.

[9] Alshamsi, Humaid & Ke¨puska, Veton & Meng, Hongying. (2017). Real Time Automated Facial Expression Recognition App Development on Smart Phones. 10.1109/IEMCON.2017.8117150.