Enhancing Facial Expression Synthesis through GAN with Multi-Scale Dilated Feature Extraction and Edge-Enhanced Facial Features*

Nimitha. U[1], P Gunasagar[2], VVSS Durgaprasad[3] , Abhijith KS[4], RVV Manikantha Sai[5] and Dr. Ameer PM[6]

***Abstract*— Affective computing aims to facilitate effective communication between humans and machines. Many affective computing systems use machine learning models trained on labeled data, like images and videos, to recognize emotions. Among these, the Generative Adversarial Network (GAN)-based expression GAN (ExprGAN) stands out as it can generate faces displaying various expressions. However, the generated faces often lack clarity in crucial facial features, such as the eyes and lips, which are essential for defining facial expressions. To address this issue, a novel feature extraction block is proposed. This module incorporates two parallel channels with multi-scale dilated convolution to mimic the human visual system and extract multi-scale facial features from the facial images. Additionally, an unsharp masking filter is integrated into the pre-processing stage to enhance the quality of facial expression features, making them sharper and clearer. The proposed model can generate faces with six distinct expressions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. To evaluate, a Convolutional Neural Network (CNN)-based expression classifier, a Principal Component Analysis (PCA)-based face classifier, and the FID (Frechet Inception Distance) score are used for comparison. The results demonstrate that the proposed model outperforms the existing ExprGAN, providing better quality and more expressive faces, which can be beneficial for improving human-machine interaction.**

I. INTRODUCTION

Human communication involves various forms such as speech, written words, body language, and facial expressions. Facial expressions play a significant role in non-verbal communication, providing additional information that enhances the effectiveness of verbal messages. In fact, studies suggest that 55% of a listener's emotions are conveyed through facial expressions [1]. Facial expression analysis encompasses tasks like facial expression recognition, classification, and facial expression synthesis. Among these, expression synthesis is particularly challenging and finds applications in diverse fields such as computer games [2], data augmentation [3], and facial expression transfer [4].

The underlying mechanics of facial expressions, such as anger, disgust, fear, happiness, sadness, and surprise, are modeled by facial expression synthesis models. Expression synthesis transforms the input face into the target expression. The numerous muscles that make up the human face determine the shape of the face and the placement of the eyes, mouth, nose, and jaw lines. The facial expression is determined by the shape and placement of the eyes, lips, nose, jaw lines, and face outline and they differ depending on the expression. In literature there exist different techniques for generating facial expressions. This work proposes an Expression GAN (ExprGAN) based model [5] with modified feature extraction layers for effectively extracting features from input faces. As a pre-processing step, we utilize an unsharp masking filter to enhance facial features such as eyes, mouth, and nose. The unsharp masking filter works by enhancing edges through the subtraction of a blurred version of the image from the original image.

The structure of the paper is as follows: Section 2 provides an overview of existing facial expression generation techniques. In Section 3, deals with the architecture of ExprGAN. The proposed model is detailed in Section 4. Results obtained from applying the approach on the OULU dataset are presented in Section 5, followed by the concluding remarks in Section 6.

II. LITERATURE SURVEY

Facial expression generation (FEG) involves utilizing image morphing techniques to generate expressions. Local warping methods, such as piece-wise polynomial transformation and radial basis function transformation, are effective in dealing with local deformations and approximating actual facial features [6]. Another technique, the geometry-driven approach, automatically generates expression images with photo-realistic and natural-looking details based on the geometry - feature point placements - of a facial expression [7]. An example-based approach estimates feature points from a subset of tracked points on the face. Deep belief networks (DBN) with multiple feature layers are used for FEG, transforming high-level facial characteristics into representations of actual faces, enabling modeling of unseen faces and remembering faces seen after training [8].

Facial Expression Editing with Controllable Expression Intensity (ExprGAN) [5] generates expressions at various intensities using an encoder-decoder network and an expression controller module. Expression for dyadic interactions [9] employs conditional Long Short Term Memory (LSTM) and the affect-shape category to generate facial expressions with head movement and other facial features. The Generative architecture (FExGAN) [10] generates facial expressions for Meta-Humans images by combining character identities and basic expressions in latent space to produce more complex ones. Fine-tuning can further enhance the outcomes.

One limitation in deep learning models is the fixed sizes of neuronal receptive fields in each layer, which hinders their ability to gather spatial information at different scales and affects the reconstruction process [11]. To address this, multi-scale convolutional networks like U-Finger [12] have been introduced, merging information from different scales to effectively denoise images. Additionally, a second-order multi-scale super-resolution network [13] concatenates second-order multi-scale features, while MSRNet (Multi-scale Residual Reconstruction Network) [14] incorporates dilated convolution for compressive sensing image reconstruction.

With the above-cited problems and solutions, we adopt multi-scale dilated convolution techniques for generating

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

facial expressions too. The encoder and decoder in the ExprGAN also use fixed kernel sizes to extract facial features and fail to capture the facial features at different scales. By adopting the solutions cited, we propose a modified encoder-decoder for feature extraction. The contributions of this work are:

Contributions of the work are summarized as follows :

- **Two channel multi-scale dilated convolution feature extraction block:** By employing two parallel convolutional channels and incorporating multi-scale dilated convolutions, the proposed block facilitates efficient extraction of multi-scale features. As a result, the network becomes capable of capturing and leveraging information from different scales within the input data.
- **Unsharp mask filter in the pre-processing stage to enhance the facial features:** The facial features are well represented by the facial parts like eyes, lips, mouth, etc. To enhance these facial features, an unsharp mask filter along with the rescaling block is added to the pre-processing stage.

III. Expression GAN (ExprGAN)

In 2014, Goodfellow et al. introduced the Generative Adversarial Network (GAN) [15] as an unsupervised machine learning technique. To enhance GAN's capabilities, Conditional GAN [16] was proposed, introducing labels to enable learning of multi−modal data. Autoencoder [17] compresses the input into the latent vector and the output is regenerated by this latent vector having a lower dimension. The ExprGAN model, as shown in Figure. 1, utilizes two neural network architectures: Conditional Generative Adversarial Network (CGAN) and Adversarial Autoencoder (AAE) [18]. For facial expression synthesis, ExprGAN takes an image and an expression label as input, allowing it to edit the face and generate different expressions. The model employs an encoder network to project the input image into a latent representation, preserving the facial identity. The expression label is converted into a more expressive code using an expression controller module. Besides the expression controller module and regularizer network, ExprGAN utilizes the generator and discriminator framework.

A. Network Architecture

ExprGAN incorporates an encoder, denoted as G_{enc}, which transforms the input image x into a latent representation named g(x), preserving the image's identity. Subsequently, a decoder, G_{dec}, reconstructs the image $\widehat{x}$ based on the latent representation g(x). To further enhance the synthetic image x with photo-realistic textures, a discriminator D_{img} is applied to the decoder G_{dec}. Additionally, to ensure that the learned identity representation is comprehensive and free of gaps, a discriminator D_z is employed on the encoder G_{enc} [19]. This process helps to refine the identity representation and ensure its completeness.

1) Generator Network G: The generator network $G = (G_{enc}, G_{dec})$ employs an autoencoder architecture. In the proposed network, the encoder (G_{enc}) plays a crucial role in converting the input image x into a latent representation while retaining as much identity information as possible. Once the identity representation g(x) and the expression code

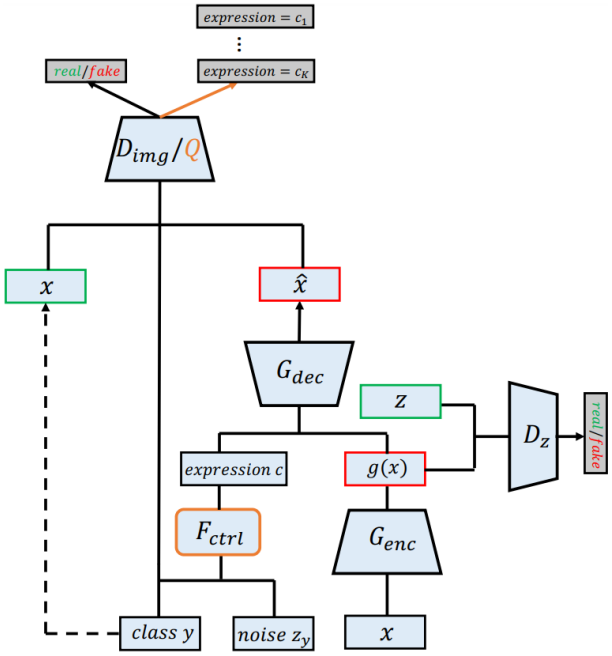


Fig. 1. ExprGAN Architecture [5]

c are derived, the decoder (G_{dec}) utilizes them to generate an image $\widehat{x} = G_{dec}(G_{enc}(x), c)$ that closely resembles the original input x. A pixel-wise reconstruction loss for the generator is the L1 loss. To retain similarity in the feature space, a pre-trained discriminative deep face model has been used, further preserving the face identity between x and $\widehat{x}$:

$$\min_{G_{enc},G_{dec}} L_{id} = \sum_{l} \beta_l L_1 \left(\phi_l \left(G_{dec}\left(G_{enc}(x), c\right)\right), \phi_l(x)\right) \tag{1}$$

In the proposed architecture, the term ϕ_1 refers to the feature maps of the l^{th} layer obtained from a pre-trained discriminative deep face model known as VGG Face [20]. These feature maps are essential for measuring similarity in the feature space. The weight β_l corresponds to the importance or significance attributed to the l^{th} layer during the similarity measurement process.

2) Discriminator on Identity: To ensure a comprehensive and hole-free representation in the latent space, $g(x)$ is modeled by applying a uniform distribution. This ensures that the generated face images from G_{dec} remain consistent with the face manifold. By incorporating this approach, the following objectives within an adversarial learning strategy are achieved.

$$\begin{aligned} \min_{G_{enc}} \max_{D_z} L^z_{adv} &= E_{z\sim P_z(z)}\left[\log D_z(z)\right] \\ &+ E_{x\sim P_{\text{data}}(x)}\left[\log\left(1 - D_z\left(G_{enc}(x)\right)\right)\right]. \end{aligned} \tag{2}$$

3) Discriminator on Image: To enhance the quality of the generated images further, a discriminator D_{img} is employed on the decoder G_{dec}. This discriminator refines the synthesized image, giving it photo-realistic textures. The loss function for the discriminator D_{img} is calculated using the

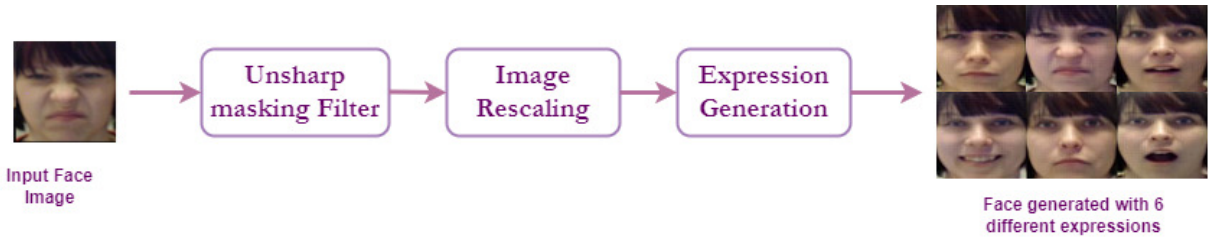


Fig. 2. Overview of Proposed Expression Generating Model

following function:

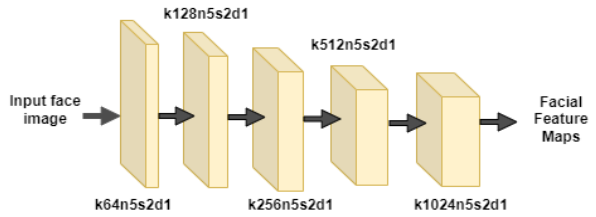
$$\min_{G_{\text{enc}},G_{\text{dec}}} \max_{D_img} L_{adv}^{img} = E_{x,y\sim P_{\text{data}}(x,y)} \left[\log D_{\text{img}}(x,y)\right] + E_{x,y\sim P_{\text{data}}(x,y),z_y\sim P_{zy}(z_y)}(x) \sim \left[\log\left(1 - D_{\text{img}}\left(G_{\text{dec}}\left(G_{\text{enc}}(x), F_{ctrl}(z_y,y)\right),y\right)\right]\right. \quad (3)$$

IV. Proposed Architecture

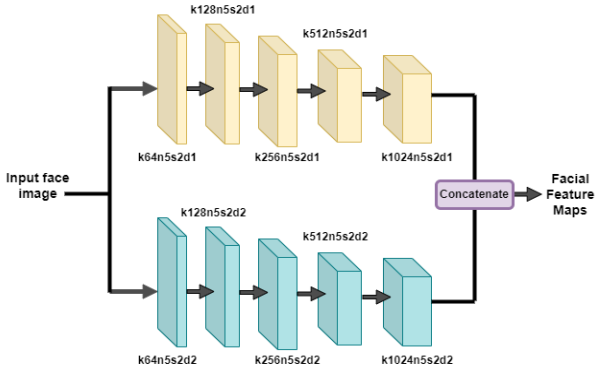
The facial features - eyes and lips - determine the emotions on the human face. Facial emotions convey the intention of a person. The expression of an emotion, such as happiness, sadness, or anger [21], conveys the mental state of the person. One-third of human communication happens verbally, while the other two-thirds happens non-verbally. Additionally, facial expressions play a significant role in interpersonal communication. The processing of blurred face images is generally limited by the texture extraction module of face recognition [22]. Sharpness enhancement techniques that affect the face region are introduced to strengthen edges and textures in order to address this issue. Unsharp mask method can improve visual clarity significantly. So as a pre-processing step, the unsharp masking filter can be used to improve the face textures. In the proposed model, the face images are filtered initially using an unsharp mask filter and then the images are re-scaled between -1 to 1. These two steps constitute the pre-processing steps for the proposed model. The proposed system for face expression generation is shown in Figure. 2.

The ExprGAN encoder in Figure. 3.a compresses into a latent space by the encoder, while the decoder aims to generate the input images from this latent space. In other words, the encoder performs the opposite task of the decoder, and vice versa. To extract features, the encoder consists of five down-sampling blocks and a fully connected layer. The down-sampling process, reducing the feature size by a factor of 2, is achieved using strided convolution with a kernel size of 5x5 and stride of 2. Each down-sampling block includes a convolutional layer and a ReLU layer. As a result of each down-sampling block, the feature map size is reduced by a factor of 2, while the number of feature maps is doubled.

The suggested feature extraction in Figure. 3.b involves two concurrent convolutional channels that employ diverse dilated factors to generate dilated convolutional kernels of varying sizes, each corresponding to unique receptive fields.By performing convolutional operations, the proposed feature extraction block acquires diverse scale feature information for each channel. This multi-scale feature information, obtained, encompasses structural details and image-specific information, enabling effective reconstruction of the textures, edges, and finer aspects of the facial images. The proposed model is composed of ten down-sampling blocks, distributed across two parallel channels, with each channel containing five blocks. In the first channel, the dilated factor is 1, and in the second channel dilated factor is 2 with both channels having a kernel size of 5x5.



(a) ExprGAN



(b) Proposed

Fig. 3. Feature Extraction Module

A. Overall Objective Function

The weighted sum of all the losses specified above is the final training loss function:

$$\min_{G_{enc},G_{\text{dec}}} \max_{D_img,D_z} L_{\text{ExprGAN}} = L_{\text{pixel}} + \lambda_1 L_{id} + \lambda_2 L_Q + \lambda_3 L_{adv}^{img} + \lambda_4 L_{adv}^{z} + \lambda_5 L_{tv} \quad (4)$$

To remove spike artifacts, we apply a total variation regularisation L_{tv} to the rebuilt image.

B. Training

The Oulu-CASIA dataset [23] consists of images from 80 subjects displaying six different expressions: Angry, Disgust, Fear, Happy, Sad, and Surprise. The images were captured under three different illumination conditions: Dark, Strong, and Weak. This dataset was utilized for evaluating the

model's performance. The dataset was split into 90% training data(1296 images) and 10% testing data(144 images). The network is trained using incremental training algorithm with Adam optimizer having a learning rate of 0.0001, β_1=0.5, and β_2=0.999. The model is trained in two stages; the image reconstruction stage (Stage-1) and the image refinement stage (Stage-2). x is initially set to random noise in stage 1. To make sure the network does not lose what it has previously learned, $Dimg$ is again trained. L_{pixel}, L_{id}, and L_{img} are the three parts of the loss function. Finally, by adding D_{img} and D_z, entire network is trained to modify the image (stage-2) to be more photo-realistic. Incremental training is used to generated images for smaller dataset.

C. Model Evaluation

The proposed model is evaluated against the existing ExprGAN using three different techniques - facial expression classifier using CNN model, Principal Component Analysis (PCA) based face recognition model, and The Fréchet inception distance (FID) score. The proposed model has an expression classification accuracy of 97% while the ExprGAN has a classification accuracy of 91% when evaluated using the CNN classifier. The proposed face expression classifier is shown in Figure. 4. The model consists of convolutional layers, Max pooling layers, and dense layers. A dropout layer is added in the network to overcome the problem of overfitting. The model uses softmax as the final layer with 6 classes - anger, disgust, fear, happy, sadness and surprise. The classifier uses Adam optimizer for optimizing the network with a learning rate of .0005.

PCA-based face recognition model uses eigenvectors named eigen faces [24] to extract features from face images. For testing the proposed model face images of 6 persons with different expressions were used, from which the model generated a total 36 face images - 6 face images for each person with 6 different expressions. Using the model, features are extracted and a minimum distance classifier is designed to classify the images into 6 different classes. The PCA-based model could classify the generated face images with an accuracy of 100% with ten eigenfaces. On the other hand, with ten eigenfaces, the ExprGAN could achieve an accuracy of 96% only.

The Frechet Inception Distance, known as FID, serves as a measurement for evaluating the quality of generated images and was specifically devised to assess the efficacy of generative adversarial networks. Features from the pre-trained Inception V3 model on the Imagenet dataset is used to measure FID score. The FID score for the proposed model is 79.10 and for the ExprGAN it is 91.32. A score of 0.0 signifies that the two images are indistinguishable, i.e, smaller values indicate a higher degree of similarity. The comparison details are summarized in table Table. 1. The proposed model outperforms the existing ExprGAN when compared with different comparison methods.

V. Results

The faces generated by the Expression GAN and the proposed model are presented in Figure. 5, Figure. 6, and Figure. 7. As mentioned earlier, an incremental training algorithm is used in training the model. The training outputs for the epochs 200 are shown in Figure. 5, the faces generated by the proposed model retain the facial features better compared to the ExprGAN. The edge details in the proposed methods are more clear and more visible. Each row in the output images represents a specific expression. The order of expression in each row is Angry, Disgust, Fear, Happiness, Sadness, and Surprise. Figure. 6(a) shows the test image, Figure. 6(b) is the faces generated by the EXprGAN and Figure. 6(c) is the faces generated by the proposed model. Comparing the test results, the expressions are more clearer with sharp edge details of eyes, lips, teeth, and mouth in the face expressions generated by the proposed model. Upon examining the happy faces generated by both the models in Figure. 6(b) and Figure. 6(c), teeth are more clearer and visible in the latter. The features of the nose are generated very well in the "Surprise" expression of the proposed model in Figure. 6(c), compared to the ExprGAN output in Figure. 6(b). Figure. 7, shows a set of test images and different facial expressions generated on the test images by ExprGAN and the proposed model. The loss curve for the models is shown in Figure. 8; for 300 epochs loss curve for the ExprGAn model in Figure. 8.a indicates that the model is still undergoing the learning phase and has not yet reached a stable point. However, the loss curve for the proposed network in Figure. 8.b demonstrates a faster learning rate and convergence compared to the ExprGAN network. This observation suggests that the proposed model exhibits improved stability and is capable of effectively learning and capturing features.

VI. Conclusion

In this work, we introduce a robust approach for generating facial expressions using a generative adversarial network (GAN) architecture, enhanced by a modified feature extraction block. To produce multi-scale receptive fields, the network contains two parallel convolutional channels and leverages dilated convolutions. In addition, an unsharp masking filter is utilized to enhance facial characteristics, which results in more enhanced and detailed features in the generated expressions. A CNN-based expression classifier, a PCA facial recognition classifier, and FID score are used to evaluate the effectiveness of the proposed method. The experimental results demonstrate that the proposed network outperforms existing state-of-the-art methods, achieving significant performance gains. These findings indicate that our architecture is reliable for generating, categorizing, and modifying facial expressions. As part of future work, we aim to generate complex facial expressions. Furthermore, the modeling of additional parallel channels with different dilation factors and feature fusion can be done for better facial image generation.

Acknowledgement

The authors thank the Centre for Computational Modelling and Simulation (CCMS) and Central Computer Centre (CCC) at NITC for the NVIDIA DGX station facility. The authors would also like to thank the Department of Science & Technology, Government of India for supporting this work

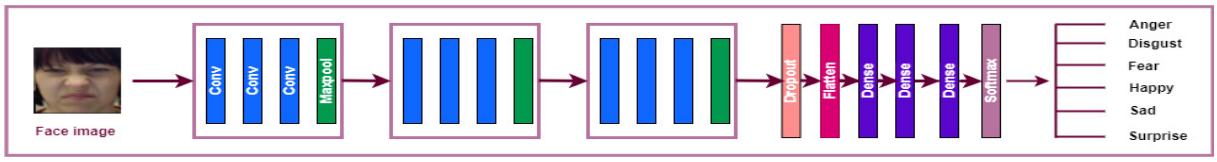
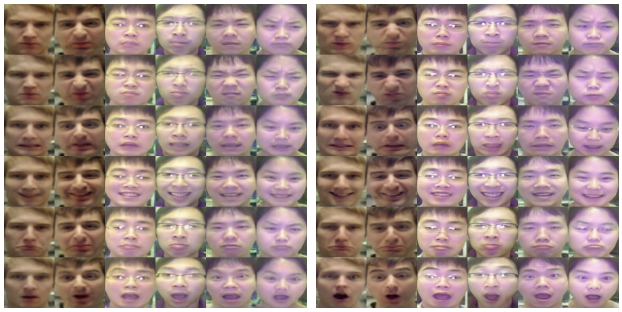


Fig. 4. CNN-based Expression classifier

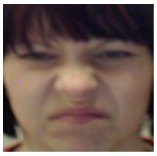
TABLE I
COMPARISON OF EXPRGAN AND THE PROPOSED MODEL

-	Metric	ExprGAN	Proposed without unsharp masking	Proposed
CNN Expression Classifier	Accuracy	91%	94%	97%
PCA Based reconstruction	Accuracy	Recognition accuracy of 96% with 10 eigen faces and above	Recognition accuracy of 100% with 14 eigen faces	Recognition accuracy of 100% with 10 eigen faces
FID	FID Score	91.32	84.73	79.10

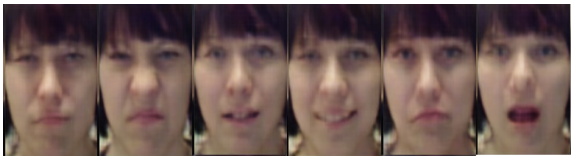


(a) ExprGAN (b) Proposed

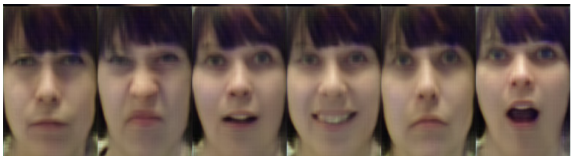
Fig. 5. Generated images after training 200 epochs



(a) Ground Truth

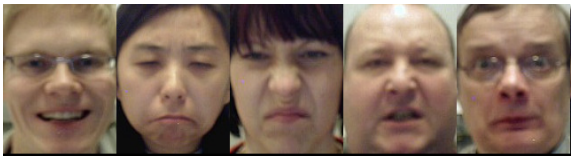


(b) ExprGAN

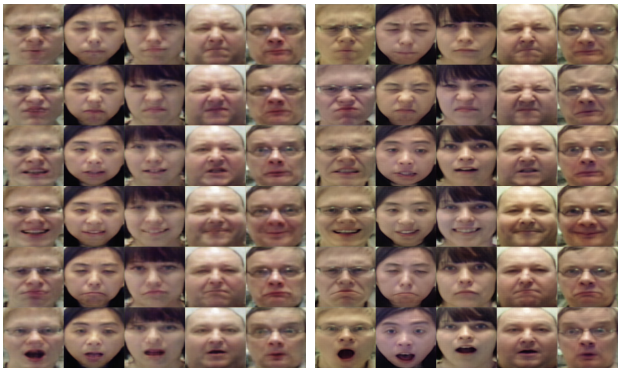


(c) Proposed

Fig. 6. Comparison of outputs generated by ExprGAN and the proposed model for a single test image



(a) Ground Truth



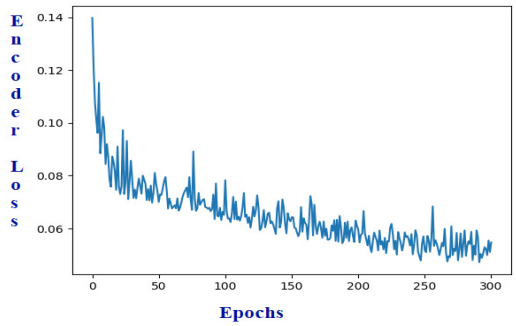
(b) ExprGAN (c) Proposed

Fig. 7. Comparison of outputs generated by ExprGAN and the proposed model

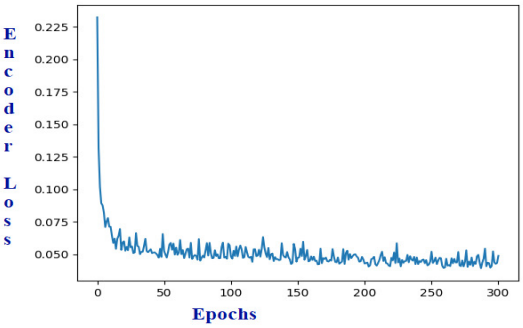
under the FIST scheme No. $SR/FST/ET-I/2017/68$.

REFERENCES

[1] Albert Mehrabian. Communication without words. In *Communication theory*, pages 193–200. Routledge, 2017.
[2] Rafael Bidarra, Robert Schaap, and Kim Goossens. Growing on the inside: Soulful characters for video games. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*, pages 337–344, 2010.
[3] Zhu Xinyue, Liu Yifan, Li Jiahong, Wan Tao, and Qin Zengchang. Emotion classification with data augmentation using generative adversarial networks. In *Advances in Knowledge Discovery and Data Mining*, pages 349–360, Cham, 2018. Springer International Publishing.
[4] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[5] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *AAAI*, pages 6781–6788. AAAI Press, 2018.
[6] Daw-Tung Lin and Han Huang. Facial expression morphing and animation with local warping methods. In *Proceedings 10th International Conference on Image Analysis and Processing*, pages 594–599, 1999.



(a) ExprGAN



(b) Proposed

Fig. 8. Loss Curve Comparison

[7] Qingshan Zhang, Z. Liu, Gaining Quo, D. Terzopoulos, and Heung-Yeung Shum. Geometry-driven photorealistic facial expression synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 12(1):48–60, 2006.

[8] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. Generating facial expressions with deep belief nets. *Affective Computing, Emotion Modelling, Synthesis and Recognition*, pages 421–440, 2008.

[9] Behnaz Nojavanasghari, Yuchi Huang, and Saad Khan. Interactive generative adversarial networks for facial expression generation in dyadic interactions. *arXiv preprint arXiv:1801.09092*, 2018.

[10] J Rafid Siddiqui. Fexgan-meta: Facial expression generation with meta humans. *arXiv preprint arXiv:2203.05975*, 2022.

[11] Zhifeng Wang, Zhenghui Wang, Chunyan Zeng, Yan Yu, and Xiangkui Wan. High-quality image compressed sensing and reconstruction with multi-scale dilated convolutional neural network. *Circuits Syst. Signal Process.*, 42(3):1593–1616, sep 2022.

[12] Ramakrishna Prabhu, Xiaojing Yu, Zhangyang Wang, Ding Liu, and Anxiao Jiang. U-finger: Multi-scale dilated convolutional network for fingerprint image denoising and inpainting. In *Inpainting and Denoising Challenges*, pages 45–50. Springer, 2019.

[13] Xiaoyu Dong, Longguang Wang, Xu Sun, Xiuping Jia, Lianru Gao, and Bing Zhang. Remote sensing image super-resolution using second-order multi-scale networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3473–3485, 2021.

[14] Yuxiang Dai and Peixian Zhuang. Compressed sensing mri via a multi-scale dilated residual convolution network. *Magnetic resonance imaging*, 63:93–104, 2019.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[16] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, abs/1411.1784, 2014.

[17] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, editors, *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 37–49. PMLR, 02 Jul 2012.

[18] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *ICLR Workshop*, 2015.

[19] Thang Vu, Tung Luu, and Chang Yoo. In *Perception-Enhanced Image Super-Resolution via Relativistic Generative Adversarial Networks*, pages 98–113, 01 2019.

[20] Andrea Vedaldi Omkar M. Parkhi and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[21] Charvi Jain, Kshitij Sawant, Mohammed Rehman, and Rajesh Kumar. Emotion detection and characterization using facial features. In *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–6, 2018.

[22] Xie Lanchi, Guo Jingjing, and Li Zhihui. A novel unsharp mask sharpening method in preprocessing for face recognition. In *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC)*, pages 378–381, 2015.

[23] Oulu-casia. `https://www.v7labs.com/open-datasets/oulu-casia`. Accessed on: 2022-01-10.

[24] Matthew A. Turk and Alex Pentland. Face recognition using eigenfaces. *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.