Unmasking Your Expression: Expression-Conditioned GAN for Masked Face Inpainting

Sridhar Sola
University of Birmingham
Birmingham, UK
solasridhar@gmail.com

Darshan Gera
Sri Sathya Sai Institute of Higher Learning
Bengaluru, India
darshangera@sssihl.edu.in

Abstract

As face masks continue to be a part of our daily lives, the challenge of reconstructing occluded faces remains relevant. While several approaches have been proposed for removing masks from neutral facial images, few have explored the use of facial expressions as a dominant feature for reconstruction of expressive faces. To address this gap, we propose an expression-conditioned GAN (ECGAN) for reconstructing masked faces with a specified expression. Our approach leverages both the binary segmentation map of the mask and an expression label to generate high-quality images. To train our ECGAN in a supervised manner, we synthesize masked images using the RAFDB dataset to create non-masked-masked pairs of images for training. We evaluate of our approach on the RAFDB test set, demonstrating its effectiveness in generating realistic images that convincingly belong to the given expression class. This is further highlighted by comparing it to a baseline model and a state-of-the-art approach without expression-input. The code is available at https://github.com/SridharSola/ECGAN.

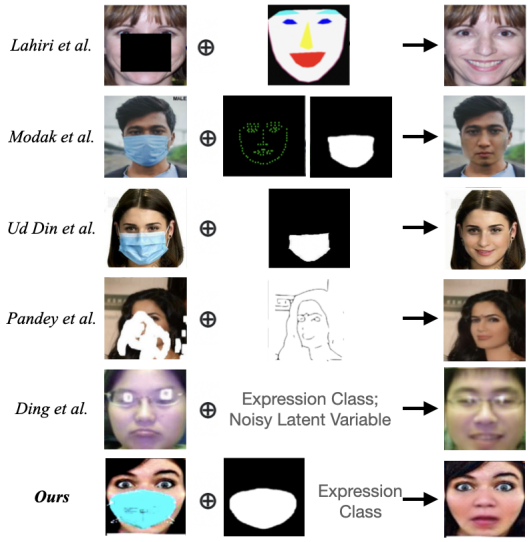

Figure 1. Comparison between the approaches of previous works [5, 6, 17, 21, 23] and ours. To the left of the arrow are features used to generate the output on the right.

1. Introduction

The Covid-19 pandemic has forced people to wear masks, which have remained ubiquitous in society. While masks are necessary for public health, they pose a significant challenge to computer vision systems that rely on facial features, such as face recognition, expression recognition, and gender classification. Since masks cover almost half of the face, important facial features are lost to these systems, making it difficult for them to perform their intended tasks. This issue is likely to persist even after the pandemic recedes, as people may continue to wear masks for personal or cultural reasons.

Various methods have been proposed for reconstructing masked portions of images, with generative adversarial networks (GANs) [8] particularly effective for inpainting holes in images. However, previous works [7, 12, 13, 16] in this area have primarily focused on incorporating contextual information, such as binary segmentation maps and landmark points, to encourage generative models to produce visually realistic content. One major semantic feature that has not been widely used to guide face reconstruction is expression information.

Expression information is essential for conveying emotions and other social signals, making it crucial for tasks such as expression recognition, sentiment analysis, and social robotics. However, facial expressions are notoriously challenging to classify due to the variation in the ways people demonstrate them and the fine face movements involved. While the field of facial expression recognition (FER) has contributed large in-the-wild datasets, such as FERPlus [1], RAFDB [18, 19], AffectNet [22] and recently MSD-E [27],

most works in this area have focused on classification and animation tasks, rather than face reconstruction.

In this paper, we propose a novel expression-conditioned GAN, named ECGAN, that unmasks a masked image, given an expression. The objective of our proposed method is to incorporate expression information into the unmasking process, thereby improving the naturalness and expression fidelity of the generated images. For this work, we use the seven basic expression classes: surprise, fear, disgust, happy, sad, anger, and neutral. Figure 1 shows the different approaches taken toward face inpainting and expression generation. Our approach builds upon previous works, particularly the works of Modak et al. [21] and Ud Din et al. [5], but goes beyond them by explicitly leveraging expression information to guide the reconstruction process.

To achieve this goal, we first generate binary segmentation maps using Mask RCNN [9], and then our generator takes in a masked image, along with the mask segmentation map and an expression class, and inpaints the masked region according to the class. We train our model adversarially with two discriminators, D_{whole} for the whole image and D_{mask} for the masked region. We evaluate the effectiveness of our proposed method on the RAF-DB test set and demonstrate that it generates images with genuine expressiveness and realism.

Overall, our proposed method has the potential to benefit a wide range of applications that rely on facial features, particularly in the context of face recognition, expression recognition, and gender classification. By leveraging large FER datasets and previous works in this area, we demonstrate the effectiveness of our approach and provide evidence that incorporating expression information significantly improves the expressiveness of generated images. Our work highlights the importance of considering expression information in face reconstruction and paves the way for future research in this area.

2. Related Work

Face inpainting and expression editing are important research topics in computer vision, with significant progress made in recent years. Early face inpainting techniques used traditional image processing methods, such as texture synthesis and exemplar-based methods. However, these methods suffered from limitations in terms of generating realistic and natural-looking results.

With the development of deep learning, GAN-based approaches have become popular for face inpainting. One of the earliest works in this direction is that of Pathak et al. [25], who proposed the Context Encoder (CE). Although the CE was not specifically designed for face images, it showed promising results on various datasets and was later extended for face inpainting by several researchers.

Iizuka et al. [15] proposed a globally and locally consistent image completion method that can fill in missing regions in images while preserving the global and local structure of the input. Zheng et al. [30] proposed a pluralistic image completion method that can generate multiple plausible results for a given input image. Wang et al. [29] developed an image inpainting technique via generative multi-column convolutional neural networks, which utilizes the features extracted from multiple convolutional columns to generate high-quality results. Pandey and Savakis [23] proposed an extreme face inpainting technique with sketch-guided conditional GAN, which can generate plausible results by utilizing additional sketch inputs. Hosen et al. [12] proposed a hybrid masked face recognition system through face inpainting, which can recover masked regions and enhance face recognition performance. Ud Din et al. [5] proposed a GAN-based network that detects and removes mask objects in facial images using a two-stage approach, where the first stage produces binary segmentation for the mask region and the second stage synthesizes the affected region with fine details while retaining the global coherency of the face structure. Modak et al. [21] developed a method for extracting and rebuilding the mask region of facial images using landmark detection, Mask R-CNN for mask segmentation.

Expression editing has also been an active area of research, with significant progress made in recent years. Early works on expression editing used traditional methods, such as 3D morphable models and facial landmarks. However, with the development of deep learning, GAN-based approaches have become popular for expression editing.

Siddiqui [26] proposed FEXGAN-META, a facial expression generation method with meta humans. Ding et al. [6] presented ExprGAN, a facial expression editing method with controllable expression intensity. Zhao et al. [3] developed DeepFaceEditing, a deep generation method of human images under arbitrary facial attribute changes. Liu et al. [28] proposed a nonlinear face morphing method that can generate high-fidelity and diverse face images with arbitrary expressions. Park et al. [24] presented SEAN, an image synthesis method with semantic region-adaptive normalization that can generate high-quality images with arbitrary attributes and Choi et al. [4] proposed StarGAN v2, a diverse image synthesis method for multiple domains.

Our proposed method, ECGAN, builds upon these works by considering both the mask position and expression to produce more natural-looking results. Our method is based on a conditional GAN that is trained on a dataset of masked face images with different expressions. The key contribution of our method is the use of expression-guided inpainting to improve the naturalness of the generated images.

Overall, the research on face inpainting and expression editing has seen significant progress in recent years, with various techniques and methods proposed for generating

high-quality and natural-looking results. However, there are still many challenges and limitations in these areas, such as dealing with complex image backgrounds, preserving the identity and personal characteristics of the input images, and ensuring the consistency and correctness of the generated results.

3. Approach

Our approach first generates a binary segmentation mask of the image, which is used, along with the masked image and an expression class, to generate an unmasked image, which is evaluated as real or fake by two discriminators: D_{whole} and D_{mask}. We discuss the architecture for each of these below.

3.1. Mask Segmentation

As in the work of Modak et al. [21], we segment the mask from the image using a Mask RCNN-based architecture. The Mask R-CNN [9] is an extension of Faster R-CNN, which is commonly used to classify and identify various objects in an image. The model has two phases, with object recognition implemented using the Faster R-CNN architecture and semantic segmentation using a Fully Convolutional Network (FCN). With this approach, Mask R-CNN generates a binary object mask, I_{bin}, for each item in the image – in our case, for the mask. This mask is binary; pixels containing the mask are 1 and others are 0.

3.2. Unmasking Generator

The purpose of the editing generator, G_{unmask}, is to fill in the regions left behind after removing the mask in an image, while maintaining structural and appearance consistency with the ground truth image. It has a CNN-based encoder and decoder. The encoder part consists of five convolution blocks, where each block is a convolution layer followed by Leaky relu activation and instance normalization layers. The encoding convolution layers have a kernel size of 3, stride of 1, and no padding. The decoder architecture is the mirror image of the encoder architecture, with deconvolution layers instead of convolution layers, which have a kernel size of 3, stride of 2, and padding of 1.

G_{unmask} also employs shortcut skip connections between the encoder and decoder to combine local information with global information by concatenating the result of deconvolution layers with feature maps from the encoder at the same level, which assists in generating more accurate and consistent results. It also uses a squeeze and excitation block [14] at the output of the first three blocks of the encoder. Additionally, four layers of atrous convolution [2] with rates of 2,4,8,16 are used between the encoder and decoder, which helps capture a large field of view and ensure consistency in the missing part generation with the rest of the face image.

Different from other face inpainting generators, G_{unmask} takes an input image I_{masked}, the binary mask from the mask segmentation generator I_{bin}, and the reshaped embedding of the class label c, to generate an unmasked image, $I_{unmasked}$. Thus,

$$I_{unmasked} = G_{unmask}(I_{masked}, I_{bin}, c) \tag{1}$$

3.3. Discriminators

The two discriminators, D_{whole} and D_{mask}, both have four contracting convolution blocks similar to the encoder part of the generator. As noted in previous works, this is done to encourage the unmasking editor to inpaint more realistic facial features in the masked region, while maintaining an overall coherent reconstructed image. Similar to other works on conditional GANs, our discriminators take in class information given to the generator to further improve the generator's performance on different expression classes. However, D_{mask} does not take in the whole image and takes in the partial unmasked portion of the unmasked image $I_{unmasked}$, called I_{gen}.

3.4. Loss Functions

We train the overall unmasking model by a combination of the following loss functions.

Reconstruction Loss: To penalize the network when it generates images different from the non-masked, we use a reconstruction loss between the generated image $I_{unmasked}$ and the non-masked image pair I_{nm}, by combining L_1 loss and structural similarity loss.

$$L_{rec} = L_1 + L_{ssim} \tag{2}$$

Discriminator Loss: To train the discriminators to correctly identify real and fake inputs we train discriminator losses as below. Suppose M represents masked images and NM represents non-masked images, then we optimize for:

$$\begin{aligned} L_{whole} = &-\mathbb{E}_{I_{nm} \in NM}[\log(D_{whole}(I_{nm}, I_{bin}, c))] \\ &+\mathbb{E}_{I_{masked} \in M}[\log(1 - D_{whole}(G_{unmask} \\ &(I_{masked}, I_{bin}, c), I_{bin}, c))] \end{aligned} \tag{3}$$

The same loss for D_{mask} is used, L_{mask}, except we pass only the masked regions of the non-masked image I_{nm} and the unmasked image $I_{unmasked}$.

Adversarial Loss: We train the generator and the discriminators adversarially using the following losses:

$$\begin{aligned} L_{whole\_adv} = &-\mathbb{E}_{I_{masked} \in M}[\log(D_{whole}(G_{unmask} \\ &(I_{masked}, I_{bin}, c), I_{bin}, c))] \end{aligned} \tag{4}$$

Again, we use a similar loss for D_{mask} called L_{mask_adv}, while passing the required mask area to the discriminator.

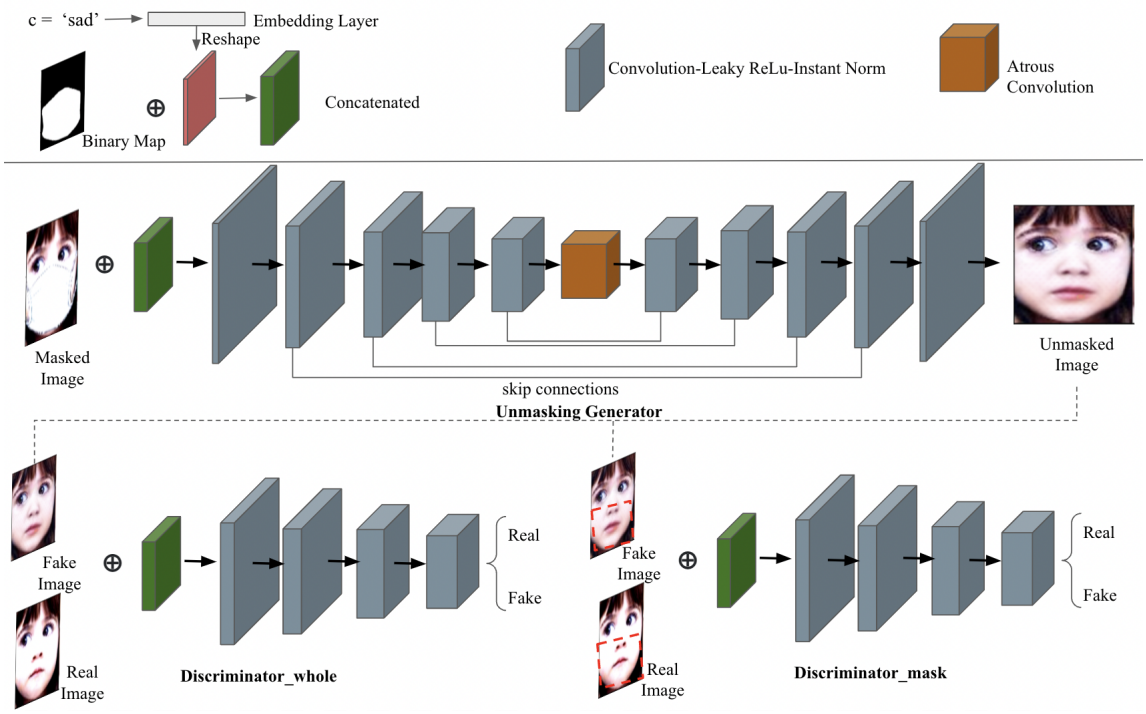

Figure 2. Architecture of ECGAN. The masked image, reshaped class embedding, and binary map are concatenated as input to the generator. The unmasked output is passed to the discriminators along with the original non-masked image.

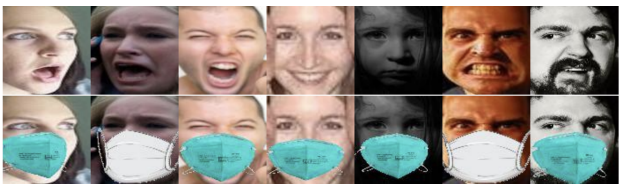

Figure 3. Sample non-masked-masked pairs from RAFDB (left to right: surprise, fear, disgust, happy, sad, anger, and nuetral).

Perceptual Loss: In order to generate natural looking unmasked images, we employ a pretrained VGG-19 network to generate the feature maps of the non-masked image and unmasked image. The perceptual loss computes the distance between the i^{th} layer's feature maps, ϕ_i of I_{nm} and $I_{unmasked}$ as below:

$$L_{per} = \sum_{i=1}^{layers} ||\phi_i(I_{nm}) - \phi_i(I_{unmasked})||^2 \quad (5)$$

Finally, we combine these losses to a total loss we use for training:

$$\begin{aligned} L_{total} = \lambda_{whole} L_{whole} + \lambda_{mask} L_{mask} \\ +\lambda_{whole\_adv} L_{whole\_adv} + \lambda_{mask\_adv} L_{mask\_adv} \\ +\lambda_{rec}(L_{rec} + L_{per}) \end{aligned} \quad (6)$$

This is our final loss conditioned on expression labels.

4. Experiments

Our experiments are designed to test our hypothesis that expression classes provide key information to ECGAN to reconstruct expressive faces. Apart from providing standard metrics for the reconstructed images, we test this using ResNet-18 [10] trained on RAFDB on the unmasked images to determine how distinguishable they are, in terms of expression. These unmasked images are on the RAFDB test set – images not previously seen by the generator during training. We now discuss the implementation details, datasets used for training and testing, and discuss results from our experiments.

4.1. Datasets

RAF-DB: The Real-world Affective Face Database (RAFDB) [18,19] contains 29,762 real-world facial images. RAFDB is annotated using 7 basic emotions: happy, neutral, surprise, sad, angry, disgust, fear. For this work, 12,271 images are used for training and 3,068 images for testing. A face mask was synthetically placed covering the mouth and nose regions for these images using an online tool [1] similar to Covid-19 scenario. Thus, we had 12,271 and 3,068 pairs of non-masked and masked images for training and testing

[1] https://github.com/X-zhangyang/Real-World-Masked-Face-Dataset

respectively. Examples of images from the training set are given in Figure 3.

4.2. Implementation Details

We implement the models in PyTorch. The binary map for the masked images are generated using the mask segmentation network. The unmasking generator and discriminators are loaded with pre-trained weights for face inpainting on CelebA [20]. The pretrained weights are the result of training the architecture without labels – which is same as that of Ud Din et al. [5] – on CelebA for 40,000 iterations. We transform the class label into a tensor the same size as the image using an embedding layer and reshaping it. Then, the masked images, resized to 256×256, are concatenated with the binary mask and embedding tensor as shown in Figure 2. Since we use a batch size of 3, a $3 \times 13 \times 256 \times 256$ tensor is the input to our model. Adam optimizer is used to train all three models – the generator, and both discriminators. The generator, G_{unmask}, and D_{whole} are trained for 20,000 iterations with an initial learning rate of 0.0003. D_{mask} is introduced to the training process after the $20,000^{th}$ iteration, and a total of 45,000 iterations are performed. As for the hyperparameters of L_{total}, we set them according to Ud Din et al. [5] $\lambda_{whole} = 0.3$, $\lambda_{mask} = 0.7$, $\lambda_{whole\_adv} = 0.3$, $\lambda_{mask\_adv} = 0.7$, $\lambda_{rec} = 100$. The training is done on a Tesla P100 GPU. We test ECGAN on the test set of RAFDB – RAFDB_test – and generate Fake_RAFDB_test. To test for expressiveness, we use ResNet-18 pretrained on RAFDB, achieving 85.20% classification accuracy on the unseen RAFDB_test.

4.3. Discussion

The goal of our work is to reconstruct the masked region of the face with an expression. The image generated by ECGAN must accurately reflect the specified expression class, be realistic, and closely resemble the non-masked image pair, in that order of priority. In this section, we discuss the results of our approach with regard to this. Figure 4 shows examples of inpainted images. Clearly, ECGAN can convincingly reconstruct the face, given an expression. We further analyze its performance below.

Expression Fidelity: To determine how well ECGAN generates expressions, we make use of Facial Expression Recognition. The pre-trained ResNet-18 for FER, when tested on Fake_RAFDB_test, achieved a classification accuracy of **78.54%**. Figure 5 shows the resulting confusion matrices. The performance on 'Happy' and 'Neutral' expressions remain largely the same, while there is a dip in the remaining. The worst affected classes are 'Fear' and 'Anger', indicating the difficulty ECGAN had in reconstructing the mouth regions to express these. Apart from being difficult to generate, the number of such images in the training data is also limited. Notably, the performance on the 'Happiness' class **slightly increased**, meaning ECGAN generates recognizably happy faces. This could be because we loaded pretrained weights trained on CelebA which is dominated by neutral and happy images.
To further determine how well ECGAN generates facial features, we visualize the feature map produced by ResNet-18 on t-sne plots, shown in Figure 6. The features of 'Surprise', 'Anger', 'Happy', 'Sad', and 'Neutral' images are similarly distributed for both RAFDB_test and Fake_RAFDB_test. However, the features of "Fear" and "Disgust" are not well separated from the others. Therefore, ECGAN can generate discriminative features in most cases, but generates weak features for 'Fear' and 'Disgust'.

Quantitative Results: While our task of expression-based inpainting is not suitable to be evaluated effectively by structural similarity (SSIM), or Fretchet distance (FID) [11], we provide them, nonetheless. FID is provided using a pre-trained VGG-19 network. The SSIM values provided are (1-ssim), meaning a score closer to 0 is better. In addition, we provide the mean perceptual loss to determine how visually realistic the unmasked images are. The results are provided for each expression in Table 1. While the SSIM scores are low, the FID scores are still high, meaning there is a significant difference in the facial features produced. The perceptual loss score, while not significantly high, is not as low as that obtained in pure face reconstruction works. This is expected as the expression generation task is much harder. Finally, we can correlate the higher scores to the poorer performance of FER in Figure 5, indicating the challenge in generating hard expressions like 'Fear', and 'Anger'.

4.4. Ablation Study

Effect of Label Input
To test how much the class input affects expression reconstruction, we inhibited the label input to ECGAN and generated images on test set of RAFDB set, referred as Fake_RAFDB_test_no_Label. Sample images can be seen in Figure 4(d). Clearly, these images lack the expression component compared to Fake_RAFDB_test Figure 4(c).

As we can see from the t-SNE plot in Figure 6(b-c), the features of Fake_RAFDB_test_no_Label are not as distinct as that of Fake_RAFDB_test. Also, ResNet-18 gets more confused with "Disgust" in this case as seen in Figure 5(c). The poorer performance of FER, in this case, confirms our hypothesis that expression classes significantly aid ECGAN in generating faces with high expression fidelity. The performance metrics are presented in Table 1

Effect of Binary Mask
To determine the impact of the binary segmentation map, we only provide ECGAN with the input image I_{masked} and

Figure 4. (a) Non-masked images from RAFDB_test. (b) Synthetically masked input images. (c) Unmasked image results. (d) Fake_RAFDB_no_Label. (e) Fake_RAFDB_no_Map. (f) Images unmasked by DeepGAN [5]. The class given to the model (when given) is shown on top.

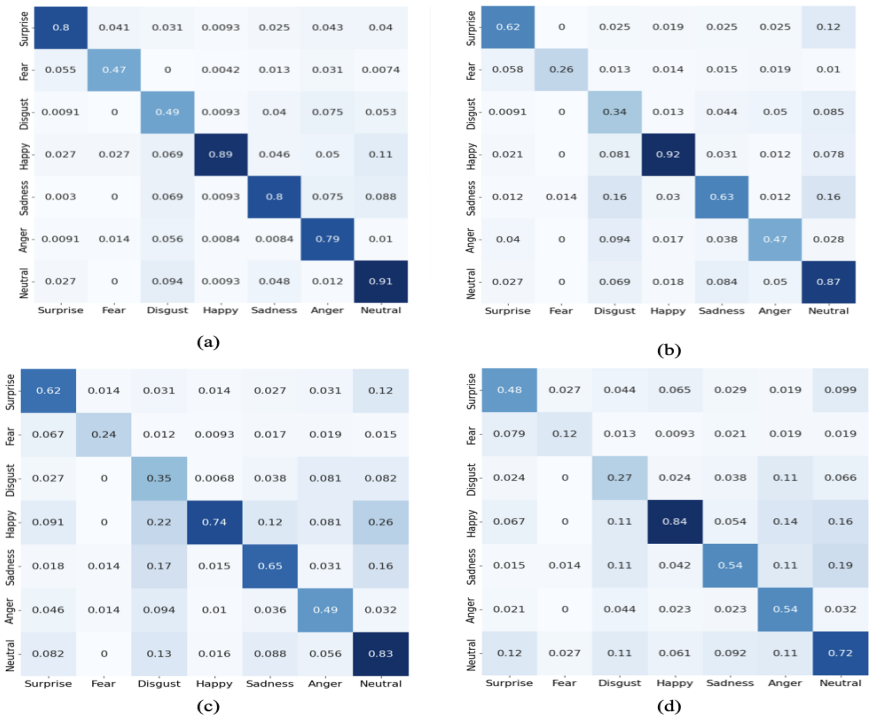

Figure 5. Confusion plots of ResNet-18 on (a) RAFDB_test. (b) Fake_RAFDB_test. (c) Fake_RAFDB_no_Label.(d) Fake_RAFDB_no_Map. Y-axis are true labels and X-axis are predicted ones.

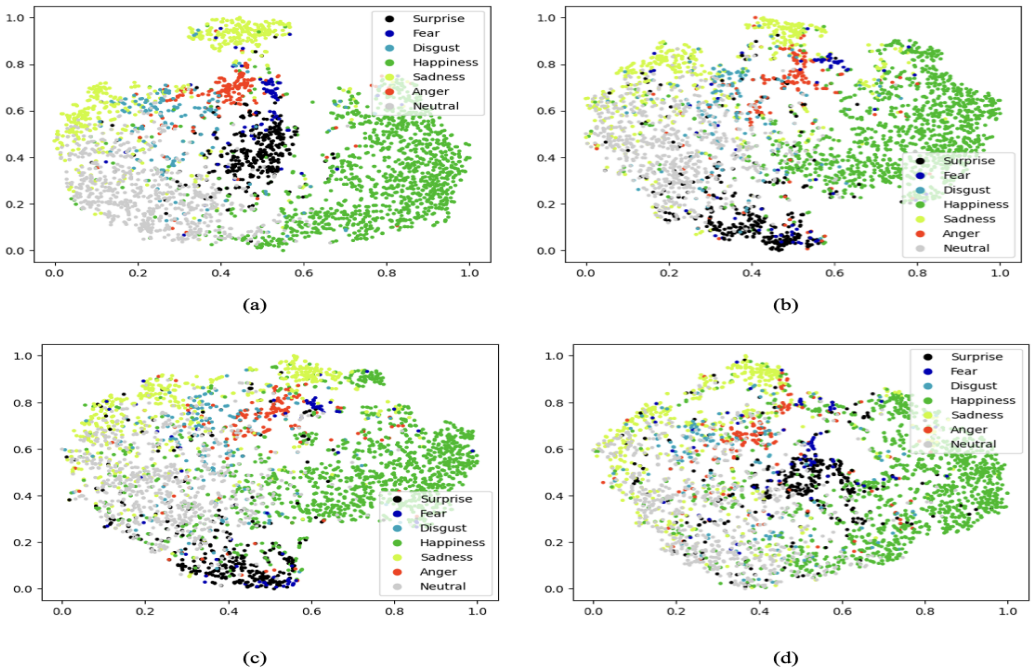

Figure 6. ResNet-18 Features of (a) RAFDB_test, (b) Fake_RAFDB_test, (c) Fake_RAFDB_no_Label, and (d) Fake_RAFDB_no_Map.

	Fake_RAFDB_test			Fake_RAFDB_test_no_Label			Fake_RAFDB_test_no_Map		
Expression	SSIM	FID	Perceptual Loss	SSIM	FID	Perceptual Loss	SSIM	FID	Perceptual Loss
Surprise	0.52	338.92	1.34	0.40	170.76	1.36	0.51	148.61	1.58
Fear	0.64	668.68	1.39	0.44	503.79	1.42	0.50	556.89	1.63
Disgust	0.53	482.43	1.22	0.42	109.93	1.31	0.51	422.26	1.68
Happy	0.53	410.27	1.28	0.42	122.70	1.33	0.52	408.81	1.71
Sad	0.55	429.36	1.32	0.42	179.88	1.34	0.52	370.95	1.71
Anger	0.62	396.95	1.46	0.44	124.41	1.44	0.52	362.88	1.81
Neutral	0.55	244.75	1.21	0.42	139.28	1.27	0.52	328.71	1.72

Table 1. Performance metrics for different generated image sets in terms of Structural Similarity (SSIM), Frechet Inception Distance (FID), and Perceptual Loss.

the class embedding c. The generated images on test set of RAFDB set are referred as Fake_RAFDB_no_Map, shown in Figure 4(e), which we evaluate. The sample images show how artifacts remain in the absence of the binary map which is reflected in the high perceptual loss value in Table 1. The t-SNE plot and resulting confusion matrix can be seen in Figure 6(d) and Figure 5(d) respectively. Therefore, the binary map plays a crucial role in directing the generator to the region that needs to be in-painted.

Expression Recognition Comparison

As the goal of our work is to reconstruct a face with given emotion label, we test whether an emotion classifier is able to detect the emotion correctly on reconstructed face by training ECGAN without conditional label and similarly ECGAN without masked map. This is shown in Figure 7. Clearly, in the absence of mask map and emotion label, there is drop of accuracy of 12.5% -13.5% when tested using images of Fake_RAFDB_test. This demonstrates that both binary map as well as class label play an important role in ECGAN.

4.5. Comparative Study

We perform a comparative study with the architecture of Ud Din et al. [5], which we refer to as DeepGAN. We note that the objectives of their work and ours, though similar, are not the same, and this comparison is done only to give a complete demonstration of how expression labels impact the generation of expressive faces. We train DeepGAN on

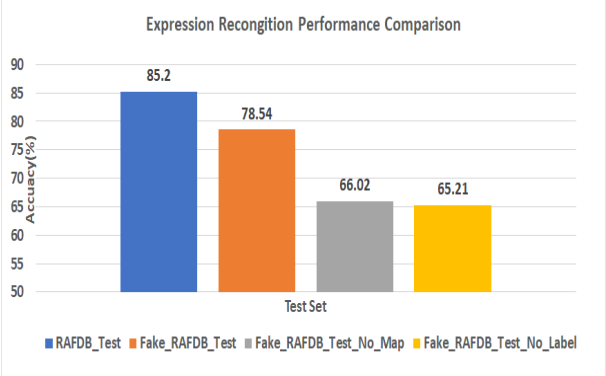

Figure 7. Expression recognition performance comparison on i) RAFDB_test, ii) Fake_RAFDB_test, iii) Fake_RAFDB_no_Label, iv) Fake_RAFDB_no_Map using ResNet-18 trained on RAFDB train set

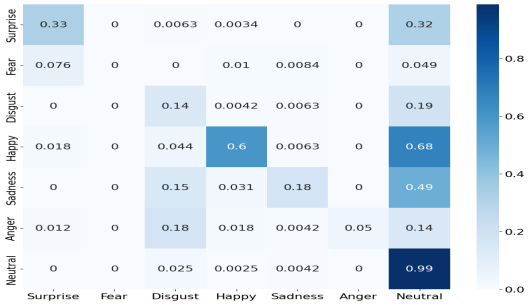

Figure 8. Confusion plot for FER on images generated by DeepGAN. Y-axis are true labels and X-axis are predicted ones.

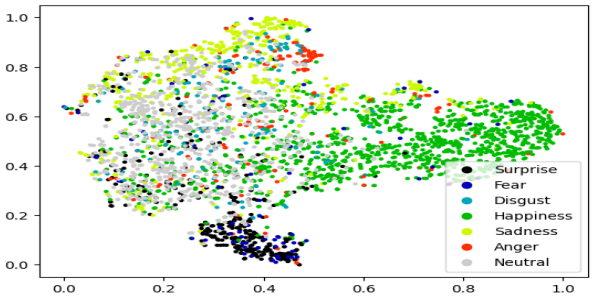

Figure 9. t-SNE plot for ResNet-18 on images generated by DeepGAN.

RAFDB with the same training setup as in Section 4.2. Examples of the unmasked images are shown in Figure 4(f) and the metrics are given in Table 2. From the quantitative results it is evident that the model generates high quality images similar to the ground truth. However, on inspecting the images unmasked by DeepGAN, it becomes clear that the model is unable to retain any of the expression information. In this regard, our model clearly comes out on top. On performing FER comparison, we find that images generated by DeepGAN result in only 55.2% accuracy – **23.3%** lower than the accuracy when tested on our generated images. Further, the confusion matrix and t-SNE plot in Figure 8 and Figure 9 respectively, show that the images generated by DeepGAN are skewed toward the 'neutral' class. These results, along with those in the ablation study in Section 4.4, provide strong support for our hypothesis on the importance of expression labels to generate animated faces.

Expression	SSIM	FID	Perceptual Loss
Surprise	0.08	149.76	0.56
Fear	0.11	411.91	0.63
Disgust	0.08	99.96	0.54
Happy	0.08	118.95	0.55
Sad	0.08	128.49	0.57
Anger	0.11	106.78	0.66
Neutral	0.57	97.98	0.42

Table 2. Metrics for DeepGAN

5. Conclusion

In this paper, we have proposed an expression-conditioned approach to GAN-based face inpainting with an expression. We provided the expression label to the generator and discriminators during training and testing, and demonstrated how this generates not only high quality unmasked faces, but also images that are distinguishable based on expression. The qualitative and quantitative results demonstrate how the expression class significantly guides the generator to successfully reconstruct the masked region with an expression. The unmasking of expression classes with low inter-class variation in the lower face region is superior to classes with high inter-class variation. However, the ablation and comparative study illustrated how the expression labels play a crucial role in generating expressive faces. All in all, this work provides evidence for the efficacy of expression conditioned GANs for further research in expression-aware face inpainting.

References

[1] Emad Barsoum, Cha Zhang, Cristian Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. pages 279–283, 08 2016. 1

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[3] Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. Deepfaceediting: Deep face generation and editing with disentan-

gled geometry and appearance control. *arXiv preprint arXiv:2105.08935*, 2021. 2

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 2

[5] Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi. A novel gan-based network for unmasking of masked face. *IEEE Access*, 8:44276–44287, 2020. 1, 2, 5, 6, 7

[6] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1, 2

[7] Xinyi Gao, Minh Nguyen, and Wei Qi Yan. Face image inpainting based on generative adversarial network. In *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6, 2021. 1

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[12] Md Imran Hosen and Md Baharul Islam. Himfr: A hybrid masked face recognition through face inpainting. *arXiv preprint arXiv:2209.08930*, 2022. 1, 2

[13] Md Imran Hosen and Md Baharul Islam. Masked face inpainting through residual attention unet. In *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5, 2022. 1

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[15] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2

[16] Yefan Jiang, Fan Yang, Zhangxing Bian, Changsheng Lu, and Siyu Xia. Mask removal: Face inpainting via attributes. *Multimedia Tools and Applications*, 81(21):29785–29797, 2022.

[17] Avisek Lahiri, Arnav Jain, Prabir Kumar Biswas, and Pabitra Mitra. Improving consistency and correctness of sequence inpainting using semantically guided generative adversarial network. *arXiv preprint arXiv:1711.06106*, 2017. 1

[18] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019. 1, 4

[19] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017. 1, 4

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[21] Gourango Modak, Shuvra Smaran Das, Md Ajharul Islam Miraj, and Md Kishor Morol. A deep learning framework to reconstruct face under mask. In *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, pages 200–205. IEEE, 2022. 1, 2, 3

[22] Ali Mollahosseini, Behzad Hasani, and Mohammad Mahoor. Affectnet: A new database for facial expression, valence, and arousal computation in the wild. *IEEE Transactions on Affective Computing*, 2017. 1

[23] Nilesh Pandey and Andreas Savakis. Extreme face inpainting with sketch-guided conditional gan. *arXiv preprint arXiv:2105.06033*, 2021. 1, 2

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[25] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting, 2016. 2

[26] J Rafid Siddiqui. Fexgan-meta: Facial expression generation with meta humans. *arXiv preprint arXiv:2203.05975*, 2022. 2

[27] Sridhar Sola and Darshan Gera. Masked student dataset of expressions. *In Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing ICVGIP'22, Gandhinagar, India. ACM, New York, NY, USA*, 2022. 1

[28] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2

[29] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. *Advances in neural information processing systems*, 31, 2018. 2

[30] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 2