International Journal of Innovative Computing, Information and Control
Volume **20**, Number **1**, February **2024**

 ISSN 1349-4198
pp. **283–295**

LOCAL EXPRESSION DIFFUSION FOR FACIAL EXPRESSION SYNTHESIS

Rung-Ching Chen[1], Chayanon Sub-r-pa[1,*], Ming-Zhong Fan[1] and Hui Yu[2]

[1]Department of Information Management
Chaoyang University of Technology
No. 168, Jifeng E. Rd., Wufeng District, Taichung 413310, Taiwan
crching@cyut.edu.tw; s10814150@gm.cyut.edu.tw
*Corresponding author: t5220317@gm.cyut.edu.tw

[2]School of Creative Technologies
University of Portsmouth
Winston Churchill Avenue, Portsmouth, Hampshire, PO1 2UP, United Kingdom
hui.yu@port.ac.uk

Received April 2023; revised September 2023

Abstract. *Facial expression synthesis has gained increasing attention in artificial intelligence applications. Existing methods use the identical facial image as input and generate the whole image for a new facial expression image, which can destroy an important identity/feature from the original image. Psychological research explains that the differences in facial expressions often appear in crucial areas, mainly in the eye and mouth. In this paper, we proposed to generate a new facial expression image from an identical facial image by minimizing the area of generating the image instead of generating the whole image. Our method is based on the Denoising Diffusion Probabilistic Model (DDPM) and text embedding for guiding the generator to produce a new image with design expression. Our method can generate realistic facial expression images while maintaining the identity from the input facial image.*

Keywords: Facial expression synthesis, Image generative model, Denoising diffusion probabilistic model, Text-guided image generator, Text-to-image

1. **Introduction.** Facial information plays a crucial role in numerous artificial intelligence applications and practical purposes, including face recognition [1, 2, 3], face mask detection [4], Facial Expression Recognition (FER) [5, 6, 7, 8, 9], and creating avatar images for virtual and augmented reality [10]. These applications have a significant impact on human society and daily life. Some of these applications rely on a neutral facial expression as input, such as face recognition or search, which can be challenging if the image does not capture a neutral expression. In such cases, facial expression synthesis or manual editing is necessary to modify the presented expression to a neutral one.

Editing facial expressions in photos often requires the expertise of skilled professionals and specialized photo editing software. However, this process can be expedited by utilizing image generation technology. Image-generation software has the capability to create a new facial expression for the entire image, including the background, face, and details of the facial expression. However, this can sometimes result in the removal of certain features or identities from the original image, such as a birthmark, scar, or skin condition.

FER is a complex computer vision task that involves several research fields, including psychology, physiology, and computer science. Recent advances in machine learning technology make it possible to estimate human expression from a full facial image [11, 12].

DOI: 10.24507/ijicic.20.01.283

However, psychological research [13, 14] reveals that the most prominent indicators of human expression are the eyes and mouth. The generator should focus only on these crucial areas to generate the required facial expression from an existing image.

According to psychological research, a practical method for gaining optimal control over the facial expression image generator involves extracting and producing only the essential details from the eyes and mouth. There are different methods to isolate the eye and mouth regions in facial images, including object detection [15, 16, 17], facial landmarks [18], and facial segmentation [19]. The image generative model can be trained with less complexity by utilizing only the extracted regions.

In this paper, we introduce a method for creating synthetic facial expressions while maintaining the identity of the original image. Our approach involves using DDPM [20] as the generator and controlling the expression through text embedding [21]. The outcome is a new facial image with a designed expression that maintains the identity of the original image.

The main contribution of our paper is a new technique for extracting facial expressions, focusing on the eye and mouth regions from original images. And build a model that modifies the extracted area to display a different expression. Overall, our method generates modified images that preserve the person's identity while displaying a different expression, unlike other methods [22, 23, 24] that alter the entire face, which can compromise some of the original identity. Additionally, by focusing on generating small portions of an image, our proposed model becomes smaller and easier to train.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed method and evaluation metrics. Section 4 presents the dataset, implementation, and experiments, then reports and discusses the results. Finally, Section 5 concludes the paper and suggests future work.

2. **Related Work.**

Facial expression synthesis is a generative artificial intelligence capable of generating images in response to prompts or conditions. Most generative models have this basic setup but are different in the details. There are three popular examples of generative model approaches: Generative Adversarial Networks (GANs) [25], Variational Autoencoders (VAEs) [26], and the DDPM [20]. Recent successful text-to-image [27] is an advanced image generator using a DDPM, such as DALL-E, DALL-E2 [28], Imagen [29], and Stable Diffusion [30].

DDPM [20] is a model that can produce high-quality images. The authors demonstrated that a specific parameterization of DDPM is equivalent to denoising score matching across various noise levels during training and annealed Langevin dynamics during sampling, which generates the best quality results. The model learns to generate images by using the denoising approach. DDPM has two steps. 1) The forward process adds random noise to the image over a series of time steps $(t_1, t_2, \ldots, t_n)$, where n is a maximum noise step hyperparameter. 2) The reverse process is to remove the added noise at each time step and denoise the samples in the backward direction $(t_n, t_{n-1}, \ldots, t_1)$.

DDPM utilizes a Convolution Neural Network (CNN) in the U-Net [31] architecture to manage the created image by learning and predicting the noise added to the input images. U-Net processes from a given noise image with known timestep (t) by progressively lowering the feature map resolution and increasing the resolution. To control the generator to design images with DDPM, conditional U-Net [32] and text embedding [21] approaches can be applied.

Contrastive Language-Image Pre-training (CLIP) [21] introduces text embedding to the image generator. CLIP is a neural network trained on various (image, text) pairs. It can be instructed in natural language to predict the most relevant text snippet, given an

image, without directly optimizing for the task, like the zero-shot capabilities of GPT-2 [33] and GPT-3 [34]. Text embedding is included in U-Net after down-sampling and up-sampling blocks with a self-attention mechanism [35].

Transformer-based models rely heavily on the self-attention mechanism [35]. To enhance the quality of images, self-attention was introduced to GANs by [36]. This enables the modeling of attention-driven, long-range dependencies for generating images. Traditional GANs create high-resolution details based on spatially local points in lower-resolution feature maps. In diffusion-based models, self-attention is incorporated into the architecture of the model. The DDPM's U-Net incorporates a self-attention layer in every downsampling and upsampling block.

DDPM has limitations regarding high-resolution images as it requires more computational power for training and generating images. However, it can still be applied as a pre-processing and post-processing technique to improve image resolution. For post-process, Imagen [29] used Super-Resolution GAN (SRGAN) [37] to upscale the output images from DDPM. Stable Diffusion utilizes an auto-encoder model [27] for pre-processing and post-processing. The stable Diffusion model compresses the input image to latent space, which DDPM then uses to learn how to denoise the space rather than the input image or random noise. The resulting latent space can be decoded to create high-resolution images.

The state-of-the-art model uses text guidance [21] to control the output images. It is designed to create a new image based on the provided prompt. However, to create new facial expression images based on an existing image, using text guidance may cause the loss of crucial features or alter the original identity of the face. Local and Global Perception Generative Adversarial Network (LGP-GAN) [38] is an approach focusing on facial expression synthesis to alter the expression that is present in the input image. They proposed to crop the essential areas, such as the eyes and mouth, and employ GANs to produce the necessary expression image. The generated image would then be merged with the original image using GANs. LGP-GAN uses three distinct GANs to create the new facial expression image.

This paper proposes a new method for generating required facial expressions inspired by LGP-GAN [38]. Our proposed generative model is based on DDPM, whereas LGP-GAN uses GANs. We extracted the eye and mouth portions of the facial image and used them as input to produce a new image. Since the cropped and generated areas were small, our method did not necessitate an extra step to enhance the resolution of the generated image.

3. **Local Expression Diffusion.**

In this section, we describe our proposed local expression diffusion, which includes data preprocessing, text-to-image for local facial expression, and evaluation metrics.

3.1. **Data preprocessing.**

Generating new facial expression images while keeping the identity of the original image is challenging. To achieve this, our method follows the guidelines from psychology research [13, 14] that facial expressions often vary in the eye and mouth areas. Therefore, we propose minimizing facial generation in these regions.

To isolate a facial image's eye and mouth regions, we rely on facial landmark data to construct a bounding box around them. Specifically, we determine the location of the left and right eyes for the eye bounding box and then calculate the top left coordinates, width, and height of the bounding box using Equations (1)-(4), where a_1 is the location of the right eye, a_2 is the location of the left eye, m_x and m_y is the fixed margin, x is an x-axis of the top left of boundary box, y is a y-axis of the top left of boundary box, w is the width of boundary box, and h is the height of boundary box.

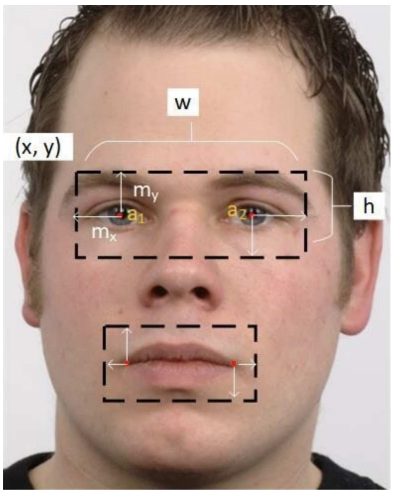

FIGURE 1. Boundary box for the eye and mouth

Figure 1 shows the visualization of the boundary box obtained from Equations (1)-(4). For this experiment, the values for m_x and m_y are set to 50 for the eye, while the mouth has $m_x = 10$ and $m_y = 40$.

$$x = a_1.x - m_x \tag{1}$$

$$y = a_1.y - m_y \tag{2}$$

$$w = (a_2.x + m_x) - (a_1.x - m_x) \tag{3}$$

$$h = (a_1.y + m_y) - (a_1.y - m_y) \tag{4}$$

3.2. **Text-to-image with DDPM.** Our method utilizes text-to-image technology to create a specific design expression for the eye and mouth. We input the extracted eye and mouth from the original image and the expression as a text caption into our model. In Figure 2, we present the steps of our method. Initially, we extract the eye and mouth regions from the original image. Then, we introduce Gaussian noise to this region and input it into the noise prediction model and the target expression in text-embedding format. The noise prediction model is trained separately to predict the added noise in the image. We continue the denoising process until the added noise is removed from the input image, resulting in a new image that reflects the target expression. We merge the

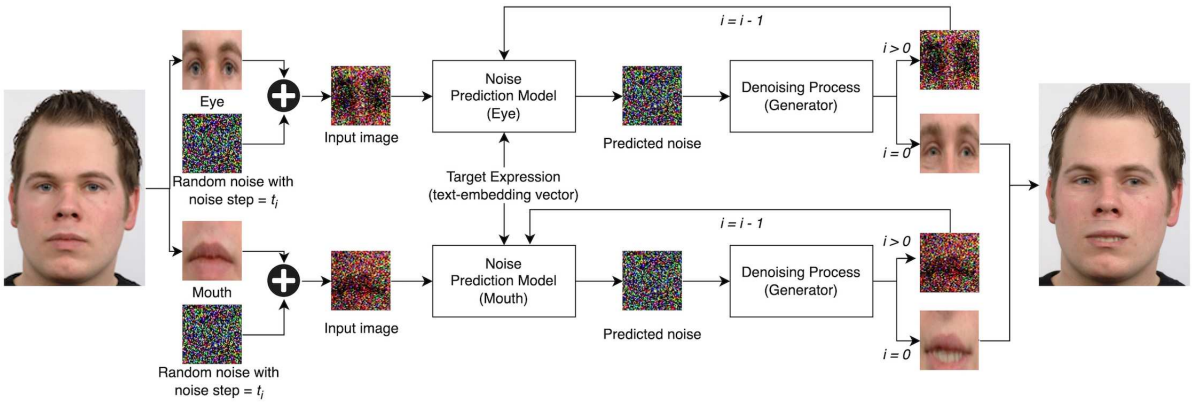

FIGURE 2. Local expression diffusion framework

generated images with the original image to get the new facial expression image precisely at the extracted location.

$$l_n = \begin{cases} \dfrac{0.5(x_n - y_n)^2}{\beta}, & \text{if } |x_n - y_n| < \beta \\ |x_n - y_n| - 0.5 \times \beta, & \text{otherwise} \end{cases} \tag{5}$$

To train the model, we add generated noise to the original image. The resulting image and an expression in text format are then used as input for the model. The target expression is then encoded to text-embedding vector and incorporated into the model using a transformer attention block. This block includes a self-attention mechanism that helps the model identify the relationship between the important regions in the image. The predicted noise from the model is compared to generated noise for loss function with smooth L1 loss (Equation (5)) where x_n is generated noise, y_n is predicted noise, and β is a fixed value of 1.0, then we optimize the model with back-propagation. The trained model's predicted noise can be used to denoise the noise image to the original.

Figure 3 shows the forward diffusion and generative denoising process, image after adding noise in different *noise_step*. To create a new facial expression image using an existing image, we introduce noise with the appropriate noise step of t_i. The maximum noise step is t_n, and n equals 500. A low t_i is utilized to preserve the feature of the original image. In the denoising process, we send the target expression to the model instead of the real expression.

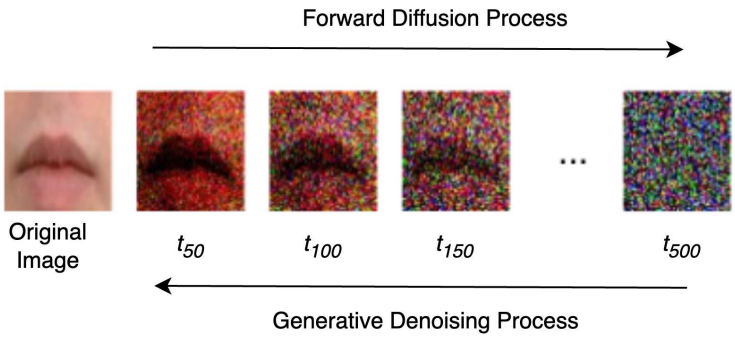

FIGURE 3. An example of generating new images through the denoising process using DDPM

To preserve the structure of an image, noise from the lower noise step (such as t_{100} from Figure 3) can be incorporated. When denoising, the model utilizes the target expression as a reference to modify the image. The noise step is a hyperparameter that allows for adjustment of the ratio to maintain the original image structure while producing the new expression image.

3.3. **Evaluation metrics.** Our experiment used two distinct metrics to quantitatively analyze our findings. These metrics are known as Inception Score (IS) [39] and Fréchet Inception Distance (FID) [40]. They are commonly used to evaluate the quality of images that have been synthesized.

IS is a mathematical algorithm to gauge the quality of AI-generated images. The score produced by the IS algorithm can range from zero (indicating the poorest quality) to infinity (indicating the best quality). The Inception Score algorithm considers two factors: Quality and Diversity. The pre-trained Inception network generates a probability distribution for the image to calculate an IS score.

To calculate the IS score for samples x_i, first construct an empirical marginal class distribution (6) where $p(y|x)$ is the conditional probability of the image being the given object x and y, and N is the number of sample images taken from the model. Then an IS score can be computed by (7), where $D_{KL}(p||q)$ is KL-divergence between the distributions p and q.

$$\hat{p}(y) = \frac{1}{N}\sum_{i=1}^{N} p(y|x_i) \tag{6}$$

$$IS = \exp\left(\frac{1}{N}\sum_{i=1}^{N} D_{KL}(p(y|x_i)||\hat{p}(y))\right) \tag{7}$$

FID metric measures the distance between feature vectors of real and generated images. It also uses a pre-trained Inception network to create a probability distribution for the image. FID calculates the mean and covariance between the feature vectors of real and generated images, FID evaluates the quality of a generated image from a model. Higher scores on FID indicate better performance.

$$FID = ||\mu_1 - \mu_2|| + Tr\left(\sigma_1 + \sigma_2 - 2\sqrt{\sigma_1 \times \sigma_2}\right) \tag{8}$$

FID is calculated using (8), where μ_1 and σ_1 represent the mean and covariance of the training data, while μ_2 and σ_2 represent the mean and covariance of the test data. Additionally, the trace linear algebra operation is denoted by Tr.

4. **Experiment.**

4.1. **Dataset.**

Our experiment uses the Radboud Faces Dataset (RaFD) [41] dataset. The RaFD consists of 4,824 images with a size of 681 × 1024 pixels collected from 67 participants in laboratory settings. Each subject has the image from 5 angles, 3 eye directions, and 8 expressions. To simplify the experiment, we use only images from the front face. We train our facial generative model with 65 subjects (1,339 images). We exclude two subjects as test sets. The facial images in this database are labeled by eight facial expressions, including angry, contemptuous, disgusted, fearful, happy, neutral, sad, and surprised.

4.2. **Implementation.**

In preprocessing, all the input facial images were first aligned and cropped to the size of 320×320 according to the facial landmarks detected by YuNet [15] Ultra-High-Performance Face Detection in OpenCV, which is an open-source library for detecting the location of a face and five facial landmarks. After we obtained the eye and mouth location, we calculated and fined the boundary box for each area in the original image, then cropped and resized the eye and mouth images from the original size image into 64×64 pixels.

To encode expression caption to text-embedding, we pre-calculate the embedding for eight expression labels using a pre-trained sentence-transformers model [42, 43] to map sentences and paragraphs to a 768-dimensional dense vector space.

We have set the maximum noise step for local expression diffusion to t_{500}. The model has been trained and optimized using Adam Optimizer [44] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The training process starts with a learning rate of 0.001, and we have implemented the CosineAnnealingLR [45] scheduler to reduce learning. The model has been trained for 500 epochs.

The trained model's weights are saved into two different versions, the original weight from training as denoted as the UNet-based model, and the updated original weight with Exponential Moving Average (EMA) [46] denoted as EMA-based model. EMA is

frequently used in image-generative models to enhance performance by updating the weight from the trained model.

4.3. **Qualitative evaluations.** To conduct qualitative evaluations, we divided the results into three sections. First, we analyzed the output images by varying the noise step (t_i) parameter, which determines the amount of noise added to the input image. This helped us to establish a suitable value for the parameter. Second, we compared the visual representations of images derived from different sources and target expressions to determine the success and failure of altering expressions from output images. Finally, we compared the images produced by our method to the state-of-the-art models.

4.3.1. *Images generated from input images with varying of t_i.* The results of our image generation method are shown in Figure 4. The original image displays an angry expression, while the target expression is happy. Creating an input with a low noise step (e.g., t_{100} and t_{200}) aims to send the structure of the original image to the model, and then the denoising process can manipulate the output image with that structure. The result using high noise step (e.g., t_{300} and t_{400}) clearly shows the target expression but they form the border.

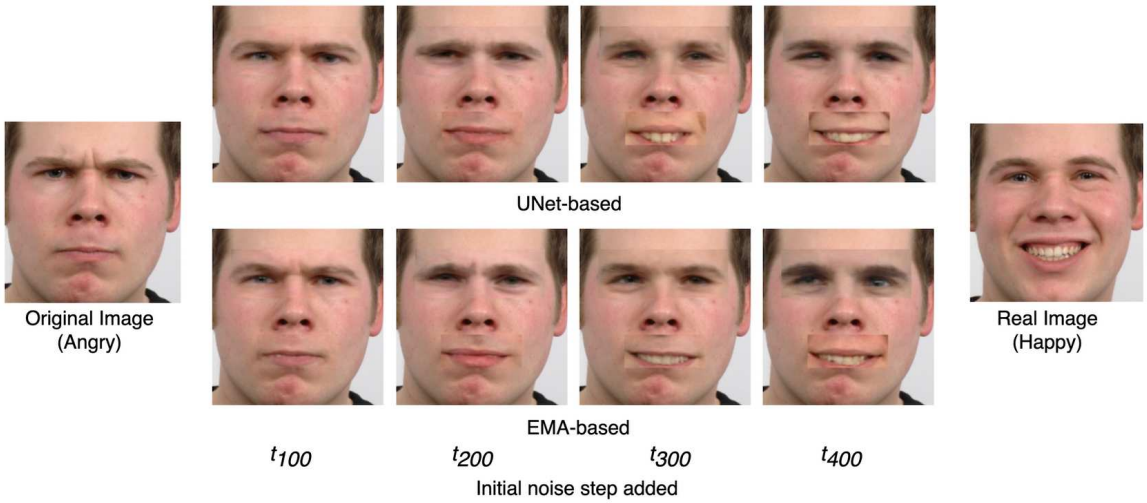

FIGURE 4. Generated facial expression image where original expression is angry and target expression is happy (upper: UNet-based, lower: EMA-based)

The output images can be affected by various noise step parameters. A low noise step generates images similar to the original with a slight change in the target expression. On the other hand, a high noise step can show the target expression but may lack information from the original image, such as skin color. This can result in a visible border when the generated image is merged with the original image.

The generated images from the UNet-based and EMA-based models are quite similar. However, the EMA-based model produces better results when merging the generated image with the original to form a complete facial image. This is because the borders are more defined and less affected by noise than in the UNet-based model.

4.3.2. *Comparison of the visual representations of images derived from different sources and target expressions.* The images in Figure 5 display different facial expressions generated from various sources and target expressions. The eight expressions included are angry, contemptuous, disgusted, fearful, happy, neutral, sad, and surprised.

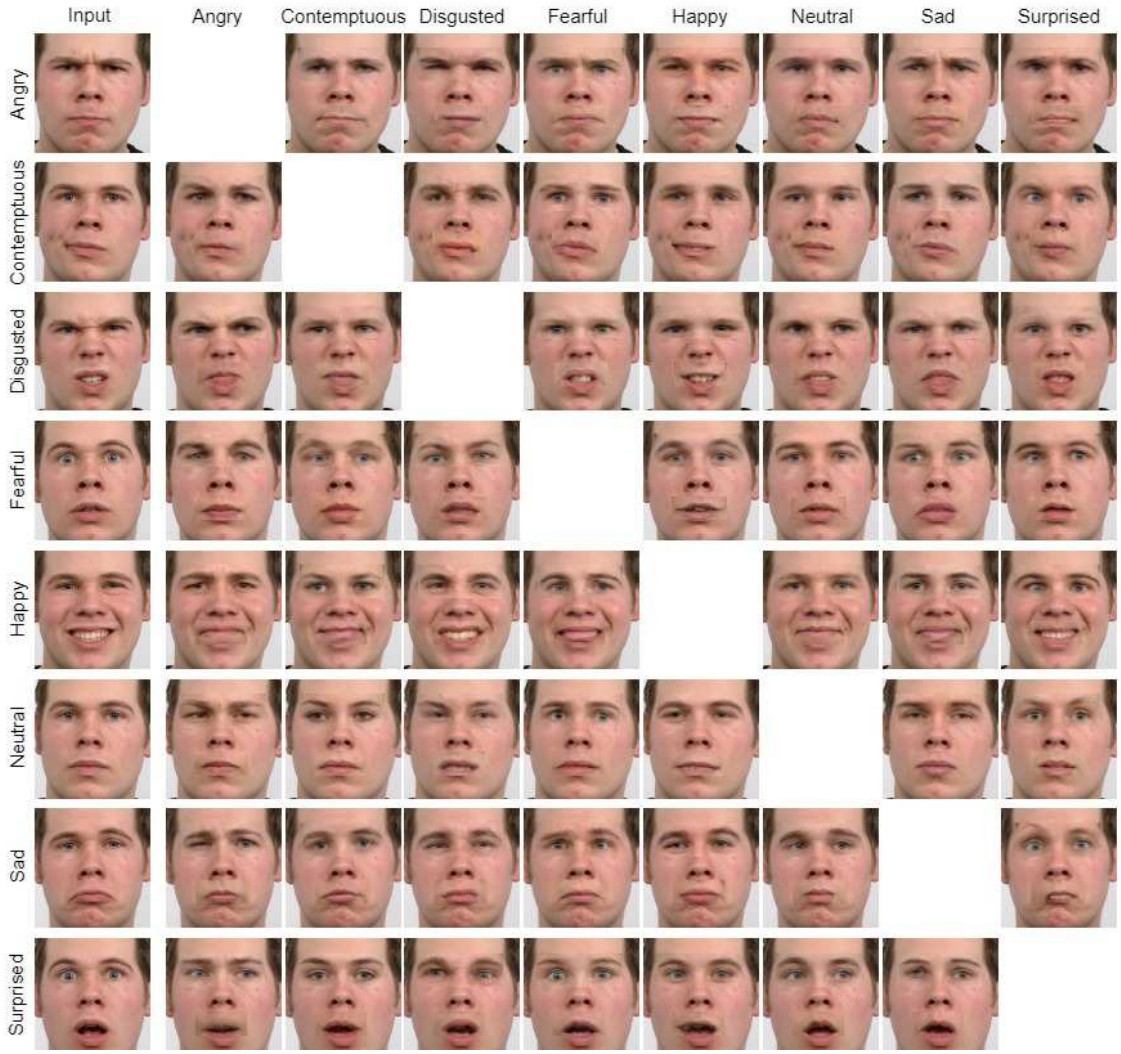

FIGURE 5. Local expression diffusion outputs using noise step t_{250}. Input expressions (from top to bottom): Angry, contemptuous, disgusted, fearful, happy, neutral, sad, and surprised.

Our method has shown promising results in creating new facial expression images. Moreover, we found that the source expression is essential in transforming the original image into the desired target expression. The output image confirms the targeted expressions when using a neutral expression as the input. However, certain expressions like happiness that involve the cheek area (which our approach does not modify) may not be accurately portrayed. Thus, the final image may still appear to have a happy expression in the cheek area.

4.3.3. *Comparison of the images produced by our method to the state-of-the-art models.* In this experiment, we compare the images generated by our method and the state-of-the-art method for facial expression synthesis from happy expression to contempt. The state-of-the-art models we used include StarGAN [22], AttGAN [23], GANimation [24], and LGP-GAN [38]. StarGAN and AttGAN are methods that focus on editing facial attributes, while GANimation is a method that uses a specific GAN network architecture to achieve impressive results in facial expression synthesis. LGP-GAN, on the other hand, generates the eye and mouth separately before generating the full facial image with another GAN-based model.

Figure 6 shows the comparison of output from state-of-the-art and our model. Ours produces more realistic facial images than StarGAN, AttGAN, and GANimation based on our results. From the results, LGP-GAN generates a better visualization of contempt expression. This is because our method does not modify the cheek area from the original image. However, other models modify the full facial image, which can remove the small detail that is the identity of the facial image from the original image.

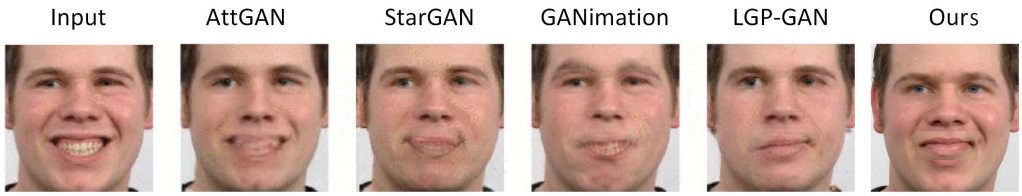

FIGURE 6. Comparison to state-of-the-art models (input expression: happy, target expression: contempt)

4.4. **Quantitative evaluation.**

4.4.1. *Inception score.* The results of the experiment's IS comparison between state-of-the-art and our method are shown in Table 1. A higher IS score indicates better image quality. Both LGP-GAN and our method received the highest scores, while the other methods scored lower. These results demonstrate that our method can generate realistic facial expressions with intricate details. However, our method using an EMA-based model did not perform well.

TABLE 1. IS of our method and other methods

Method	IS
StarGAN [22]	1.29
AttGAN [23]	1.26
GANimation [24]	1.30
LGP-GAN [38]	1.31
Our method with UNet-based model	1.31
Our method with EMA-based model	1.27

4.4.2. *Fréchet Inception Distance.* FID is a metric that evaluates the similarity between real and generated images in a given dataset. However, when using our method, which only generates the eye and mouth regions, calculating FID based on full facial images can lead to biased results. This is because the real-image content in full facial images is over 50%. To avoid this bias, we only report FID values generated by our method for different image portions.

Table 2 presents the FID scores. A low score indicates the generated images are realistic and comparable to real samples, without noise, blur, or unrealistic elements. Our FID results indicate that both the UNet-based and EMA-based models demonstrate our generator's ability to produce realistic images.

TABLE 2. FID from our method in different generated areas

Area	UNet-based model	EMA-based model
Eye	0.4922	0.4843
Mouth	0.3823	0.3872

5. **Conclusion.** Our research introduces a novel technique called "local expression diffusion" to alter facial images to design expressions. Our proposed system extracts the critical features of the eyes and mouth from the source image, generates a new image with the desired expression, and then integrates it back into the original image. Through our experiments, we have demonstrated that this method can generate new facial expressions while retaining the identity of the original image. However, we have observed that some combinations of source and target expressions may result in unrealistic facial images. Moreover, our method is currently limited to datasets in controlled environments. Therefore, we will conduct further experimentation and analysis to expand our method's applicability to facial expressions in the wild datasets.

In future work, we aim to determine the most appropriate shape for modifying the nose and cheek area in certain facial expressions of the original image. Our experiment has shown that the cropping area should not be restricted to a rectangular shape but should be expanded. Additionally, we intend to improve the quality of the resulting image by incorporating a new post-processing technique into our model.

Acknowledgment. This paper is supported by the Ministry of Science and Technology, Taiwan. The Nos are NCST-111-2622-E-324-002- and NCST-112-2221-E-324-011-MY2, Taiwan.

REFERENCES

[1] W.-C. Cheng, H.-C. Hsiao and D.-W. Lee, Face recognition system with feature normalization, *International Journal of Applied Science and Engineering*, vol.18, pp.1-9, 2021.

[2] O. M. Parkhi, A. Vedaldi and A. Zisserman, Deep face recognition, *British Machine Vision Conference*, 2015.

[3] S. I. Serengil and A. Ozpinar, LightFace: A hybrid deep face recognition framework, *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp.1-5, 2020.

[4] W.-C. Cheng, H.-C. Hsiao, Y.-Q. Hong and D.-Y. Wang, Masked face recognition based on FaceNet and genetic algorithm, *International Journal of Applied Science and Engineering*, vol.20, pp.1-10, 2023.

[5] Y. Guo, Y. Xia, J. Wang, H. Yu and R.-C. Chen, Real-time facial affective computing on mobile devices, *Sensors*, vol.20, no.3, 2020.

[6] Z. Lian, Y. Li, J.-H. Tao, J. Huang and M.-Y. Niu, Expression analysis based on face regions in real-world conditions, *International Journal of Automation and Computing*, vol.17, no.1, pp.96-107, 2019.

[7] X. Sun and M. Lv, Facial expression recognition based on a hybrid model combining deep and shallow features, *Cognitive Computation*, vol.11, no.4, pp.587-597, 2019.

[8] S. Zhang, H. Yu, T. Wang, J. Dong and T. D. Pham, Linearly augmented real-time 4D expressional face capture, *Information Sciences*, vol.545, pp.331-343, 2021.

[9] Y. Wang, X. Dong, G. Li, J. Dong and H. Yu, Cascade regression-based face frontalization for dynamic facial expression analysis, *Cognitive Computation*, vol.14, no.5, pp.1571-1584, 2021.

[10] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang and H. Yu, Realistic facial expression reconstruction for VR HMD users, *IEEE Transactions on Multimedia*, vol.22, no.3, pp.730-743, 2020.

[11] X. Wang, X. Hao and K. Wang, Facial expression recognition based on multi-branch adaptive squeeze and excitation residual network, *International Journal of Innovative Computing, Information and Control*, vol.17, no.3, pp.735-751, 2021.

[12] G. R. Naufal, R. Kumala, R. Martin, I. T. A. Amani and W. Budiharto, Deep learning-based face recognition system for attendance system, *ICIC Express Letters, Part B: Applications*, vol.12, no.2, pp.193-199, 2021.

[13] M. Allen, *The SAGE Encyclopedia of Communication Research Methods*, SAGE Publications, Inc., 2017.

[14] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[15] Y. Feng, S. Yu, H. Peng, Y.-R. Li and J. Zhang, Detect faces efficiently: A survey and evaluations, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol.4, no.1, pp.1-18, 2022.

[16] H. Peng and S. Yu, A systematic IoU-related method: Beyond simplified regression for better localization, *IEEE Transactions on Image Processing*, vol.30, pp.5032-5044, 2021.

[17] D. Qi, W. Tan, Q. Yao and J. Liu, YOLO5Face: Why reinventing a face detector, in *Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science*, L. Karlinsky, T. Michaeli and K. Nishino (eds.), Cham, Springer Nature Switzerland, 2023.

[18] X. Dong, Y. Yang, S.-E. Wei, X. Weng, Y. Sheikh and S.-I. Yu, Supervision by registration and triangulation for landmark detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.43, no.10, pp.3681-3694, 2021.

[19] K. Khan, R. Khan, K. Ahmad, F. Ali and K. Kwak, Face segmentation: A journey from classical to deep learning paradigm, approaches, trends, and directions, *IEEE Access*, vol.8, pp.58683-58699, 2020.

[20] J. Ho, A. Jain and P. Abbeel, Denoising diffusion probabilistic models, *Advances in Neural Information Processing Systems*, vol.33, pp.6840-6851, 2020.

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, Learning transferable visual models from natural language supervision, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol.139, pp.8748-8763, 2021.

[22] 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, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8789-8797, 2018.

[23] Z. He, W. Zuo, M. Kan, S. Shan and X. Chen, AttGAN: Facial attribute editing by only changing what you want, *IEEE Transactions on Image Processing*, vol.28, no.11, pp.5464-5478, 2019.

[24] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu and F. Moreno-Noguer, GANimation: Anatomically-aware facial animation from a single image, in *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (eds.), Cham, Springer International Publishing, 2018.

[25] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, Analyzing and improving the image quality of StyleGAN, *Proc. of CVPR*, 2020.

[26] L. Pinheiro Cinelli, M. Araújo Marins, E. A. Barros da Silva and S. Lima Netto, *Variational Autoencoder*, Springer International Publishing, Cham, pp.111-149, 2021.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, High-resolution image synthesis with latent diffusion models, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.10674-10685, 2022.

[28] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever, Zero-shot text-to-image generation, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol.139, pp.8821-8831, 2021.

[29] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet and M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding, *arXiv Preprint*, arXiv: 2205.11487, 2022.

[30] A. Blattmann, R. Rombach, K. Oktay, J. Müller and B. Ommer, Retrieval-augmented diffusion models, *Advances in Neural Information Processing Systems*, vol.35, pp.15309-15324, 2022.

[31] O. Ronneberger, P. Fischer and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*, N. Navab, J. Hornegger, W. Wells and A. Frangi (eds), Cham, Springer International Publishing, 2015.

[32] G. M. Brocal and G. Peeters, Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations, *arXiv Preprint*, arXiv: 1907.01277, 2019.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., Language models are unsupervised multitask learners, *OpenAI Blog*, vol.1, no.8, 2019.

[34] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language models are few-shot learners, *Advances in Neural Information Processing Systems*, vol.33, pp.1877-1901, 2020.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, vol.30, 2017.

[36] H. Zhang, I. Goodfellow, D. Metaxas and A. Odena, Self-attention generative adversarial networks, *International Conference on Machine Learning*, pp.7354-7363, 2019.
[37] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang and W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.105-114, 2017.
[38] Y. Xia, W. Zheng, Y. Wang, H. Yu, J. Dong and F.-Y. Wang, Local and global perception generative adversarial network for facial expression synthesis, *IEEE Transactions on Circuits and Systems for Video Technology*, vol.32, no.3, pp.1443-1452, 2022.
[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen and X. Chen, Improved techniques for training GANs, *Advances in Neural Information Processing Systems*, vol.29, 2016.
[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, pp.6629-6640, 2017.
[41] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk and A. van Knippenberg, Presentation and validation of the radboud faces database, *Cognition & Emotion*, vol.24, no.8, pp.1377-1388, 2010.
[42] N. Reimers and I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, *Conference on Empirical Methods in Natural Language Processing*, 2019.
[43] K. Song, X. Tan, T. Qin, J. Lu and T.-Y. Liu, MPNet: Masked and permuted pre-training for language understanding, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020.
[44] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *The 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.
[45] I. Loshchilov and F. Hutter, SGDR: Stochastic gradient descent with warm restarts, *International Conference on Learning Representations*, 2017.
[46] Z. Cai, A. Ravichandran, S. Maji, C. Fowlkes, Z. Tu and S. Soatto, Exponential moving average normalization for self-supervised and semi-supervised learning, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.194-203, 2021.

Author Biography

Rung-Ching Chen (Member, IEEE) received the B.S. degree from the Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, in 1987, the M.S. degree from the Institute of Computer Engineering, National Taiwan University of Science and Technology, in 1990, and the Ph.D. degree in Computer Science from the Department of Applied Mathematics, National Chung Hsing University, in 1998.

Prof. Chen is currently a Distinguished Professor with the Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan. He is listed in the Top 2% Scientists Worldwide in A.I. by Stanford University. His research concerns include network technology, pattern recognition, knowledge engineering, the Internet of Things, data analysis, and artificial intelligence.

Chayanon Sub-r-pa received the B.S. degree from the Department of Computer Science, Kasetsart University, Thailand, in 2005, the M.S. degree from the Department of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Thailand, in 2010, and the Ph.D. degree from the Department of Software and Information Science, Iwate Prefectural University, Japan, in 2017.

Dr. Chayanon is currently a Postdoctoral Researcher at the Department of Information Management, Chaoyang University of Technology, Taiwan. Before joining Chaoyang University of Technology, he worked as a Lecturer with the King Mongkut's Institute of Technology Ladkrabang. His research interests are computer networks, road networks, intelligent transportation systems, computer vision, and generative AI.

Ming-Zhong Fan received the B.S. degree from the Department of Information Management, Chaoyang University of Technology, Taiwan, in 2023. He is currently a Master's student at the Department of Information Management, Chaoyang University of Technology, Taiwan. His research focuses on deep learning, convolutional neural networks, and generative AI.

Hui Yu (Senior Member, IEEE) received the Ph.D. degree from the Brunel University London, Uxbridge, U.K., in 2009. He is a Professor of visual computing with the University of Portsmouth, Portsmouth, U.K. He worked with the University of Glasgow, Glasgow, U.K., and Queens University Belfast, Belfast, Northern Ireland, before joining the University of Portsmouth in 2012. His research interests include vision, creative computing and AI with the applications in 4-D facial expression of emotion, human-machine interaction, VR/AR, and video analysis. He serves as an Associate Editor for *IEEE Transactions on Human-Machine Systems and Neurocomputing* Journal.