Generation of High-Resolution Facial Expression Images Using a Super-Resolution Technique and Self-Supervised Guidance

Tatsuya Hanano [1], Masataka Seo [2], and Yen-Wei Chen [1,*]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan
[2] Osaka Institute of Technology University, Osaka, Japan; Email: masataka.seo@oit.ac.jp (M.S.)
*Correspondence: chen@is.ritsumei.ac.jp (Y.W.C.)

***Abstract*—The recent spread of smartphones and social networking services has increased the means of seeing images of human faces. Particularly, in the face image field, the generation of face images using facial expression transformation has already been realized using deep learning–based approaches. However, in the existing deep learning–based models, only low-resolution images can be generated due to limited computational resources. Consequently, the generated images are blurry or aliasing. To address this problem, we proposed a two-step method to enhance the resolution of the generated facial images by combining a super-resolution network following the generative model, which can be considered a serial model, in our previous work. We further proposed a parallel model that trains a generative adversarial network and a super-resolution network through multitask learning. In this paper, we propose a new model that integrates self-supervised guidance encoders into the parallel model to further improve the accuracy of the generated results. Using the peak signal-to-noise ratio as an evaluation index, image quality was improved by 0.25 dB for the male test data and 0.28 dB for the female test data compared with our previous multitask-based parallel model.**

***Keywords*—image processing, deep learning, facial expression transformation, generative adversarial networks, super resolution**

I. Introduction

Human face images are now widely used in various fields, including advertisements for sales promotions in the industrial world and diagnostic reference materials in the medical industry. Consequently, technologies for generating fictitious human face images by changing a person's gender, age group, and facial expressions have been attracting attention, and related face image conversion technologies have been actively developed. Particularly, facial expression generation from a single facial image is widely applied in entertainment and social communication. Achieving facial expression changes is possible using Pix2Pix [1, 2] or other generative adversarial networks (GANs) [3, 4]. However, deep learning or GAN-based methods cannot generate high-resolution images due to limited computational resources. To solve this problem, we proposed a two-step method [5], which comprises Pix2Pix and a super-resolution (SR) model, as well as its improved end-to-end model [6], in which a facial image is generated using Pix2Pix and a high-resolution image is generated via an SR network. The weakness of these methods was their complexity because two models (generation and SR models) were used in the training and testing phases. Therefore, we proposed a parallel model with multitask learning [7]. Unlike previous methods [5, 6], this method uses both models only in the training phase. In the test phase, only the generation model of the main task is used to generate high-resolution facial expression images from low-resolution expressionless facial images. Compared with conventional methods, this method improved accuracy while reducing computational costs. Inspired by Wang *et al.* [8], we proposed another approach, which can be considered as an improved cGAN with two encoders [9]. An additional self-supervised guidance encoder (SGE) is integrated into the cGAN to extract high-frequency features by cropping patches from the original high-resolution images.

In this paper, we incorporate the SGE into the multitask-based parallel model [7] to further improve facial image generation accuracy. In the proposed method, two SGEs are adopted for the facial expression generation task and the super-resolution task, respectively. In each task, the features obtained from the SGE are input to each layer of the corresponding decoder stepwise, which is expected to improve the generation accuracy and the super-resolution accuracy.

II. Related Work

A. Pix2Pix

To date, Pix2Pix [2] has been used for facial expression transformation, which is based on the conditional adversarial network [10], a type of GAN [11]. GAN comprises two neural networks: the generator and the discriminator. These two networks are trained on each other to output an image similar to the training image. The framework of the generator is shown in Fig. 1 and

Manuscript received December 4, 2022; revised April 22, 2023; accepted May 10, 2023.

comprises eight encoder layers and eight decoder layers. The upper number indicates the number of output channels in each layer. The generator uses a network called "U-net" [12] to avoid information bottlenecks. U-net is a network that adds a "skip connection" to the encoder–decoder network between each layer (Fig. 1). The framework of the discriminator is shown in Fig. 2; the discriminator is constructed similarly to the encoder layer of the generator.

Pix2Pix uses "conditional image/image pairs of expressionless and expression images" as training data and learns the correspondence between them. By learning the relationship between images in this way, generating one image from another is possible. In this study, we used this network as a baseline to generate images of facial expression changes.

The objective function of Pix2Pix is shown in Eq. (1), which is the objective function of cGAN (Eq. (2)) plus an L1 loss (Eq. (3)). Eq. (1) represents a minimax game where the generator tries minimizing the objective function, and the discriminator tries maximizing it. G represents the generator, and D represents the discriminator; ϕ_G and ϕ_D represent the parameters of G and D, respectively, where x is the training image, y is the real image, $D(\phi_D; \mathbf{x}, \mathbf{y})$ is the output of the discriminator with the training and real images as input, and $D\left(\phi_D; \mathbf{x}, G(\phi_G; \mathbf{x})\right)$ is the output of the discriminator with the training and generated images as input. $G(\phi_G; \mathbf{x})$ is the image generated by the generator. The limitation of the Pix2Pix (cGAN)-based method is that it can only input low-resolution images due to limited computational resources and cannot generate high-resolution images.

$$\min_G \max_D L_{P2P}(D, G) = L_{cGAN}(D,\ G) + \lambda L_{L1}(G) \quad (1)$$

$$L_{cGAN}(D, G) = E_{\mathbf{x},\mathbf{y}}[\ln D(\phi_D; \mathbf{x}, \mathbf{y})] + E_{\mathbf{x}}\left[\ln\left(1 - D\left(\phi_D; \mathbf{x}, G(\phi_G; \mathbf{x})\right)\right)\right] \quad (2)$$

$$L_{L1}(G) = E_{\mathbf{x},\mathbf{y}}\left[\|\mathbf{y} - G(\phi_G; \mathbf{x})\|_1\right] \quad (3)$$

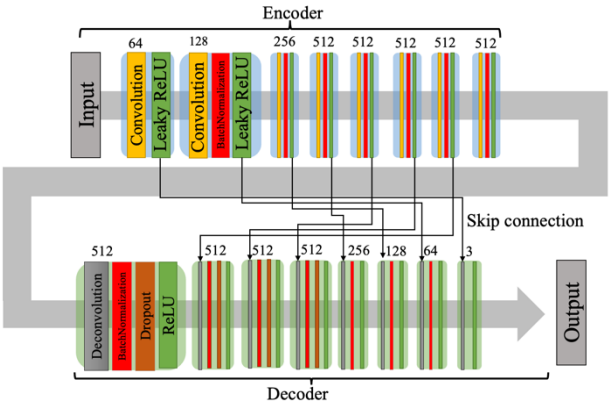

Figure 1. The network structure of the generator.

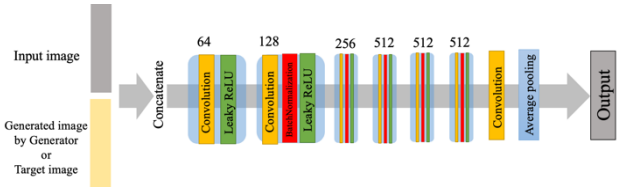

Figure 2. The network structure of the discriminator.

B. *Serial Model (Pix2Pix + SR)*

To solve the low-resolution problem, we proposed a two-step approach [5] and its end-to-end two-step model [6] in our previous work (Figs. 3 and 4). The two-step approach comprises Pix2Pix [2] and a SR model (SRCNN) [13], as well as its improved end-to-end model [6].

In the first step, the Pix2Pix generates a low-resolution expression image from an expressionless image. Subsequently, the generated low-resolution expression image is interpolated and enlarged using Bicubic. In the second step, the generated low-resolution expression image is enhanced to a high-resolution expression image using the SR model (SRCNN). In the two-step approach [5], Pix2Pix and SRCNN are trained separately. In the end-to-end two-step model [6], Pix2Pix and SRCNN are trained in an end-to-end fusion using Eq. (4) and (5) as objective functions. In Eq. (4), G, D, and S represent the generator, discriminator, and SR, respectively. The first term of Eq. (4) is the objective function for Pix2Pix (Eq. (1)), and the second term is for SR (Eq. (5)). λ is a hyperparameter. In Eq. (5), S represents the SR model, and ϕ_s and ϕ_G represent parameters of the SRCNN and generator, respectively. Eq. (5) represents the mean squared error of the reconstructed image $S(\phi_s; G(\phi_G; \mathbf{x}))$ and the corresponding high-resolution expression image $\mathbf{y}$. $\mathbf{x}$ is the low-resolution expressionless image.

$$\min_{G,S} \max_D L_{total}(D, G, S) = L_{P2P}(D,\ G) + \lambda L_{SR}(G, S) \quad (4)$$

$$\min_S L_{SR}(S) = E_{\mathbf{x},\mathbf{y}}\left[\|\mathbf{y} - S(\phi_S; \mathbf{x})\|_2\right] \quad (5)$$

The limitation of this approach is that it is a two-step method, and two CNNs are used to generate high-resolution expression images.

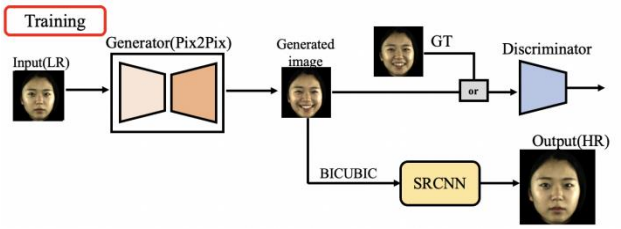

Figure 3. Overview of the two-step approach (Pix2Pix + SR) during training.

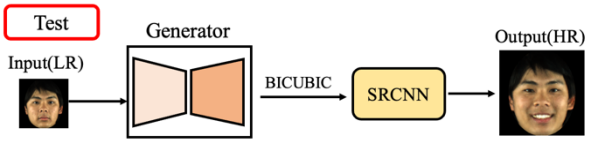

Figure 4. Overview of the two-step approach (Pix2Pix + SR) during inference.

C. Multitask-Based Parallel Model (Pix2Pix + SR)

We further proposed an end-to-end parallel model that integrates the SR network into the existing Pix2Pix pipeline [7]. Specifically, the model is trained as parallel multitask learning, with Pix2Pix as the main task and SR as an auxiliary task. Fig. 5 shows the network during training, and Fig. 6 shows the network during inference. Notably, although this method uses two models (generation and SR models) in the training phase (Fig. 5), it only uses the generation model of the main task to generate the high-resolution facial expression image in the inference phase (Fig. 6). We used a model in which the encoders of the two streams are common and share weights. This allows the SR network to be removed during inference, which is expected to improve the resolution of the generated images without incurring extra computational costs. The weakness of this method is that it produces an image with less blurring overall than the serial model; however, it does not adequately generate small-edge information.

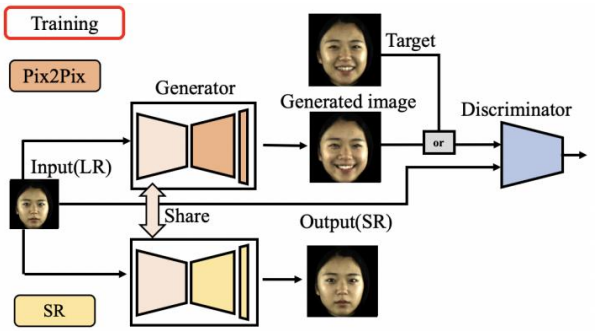

Figure 5. Overview of the multitask-based parallel model during training.

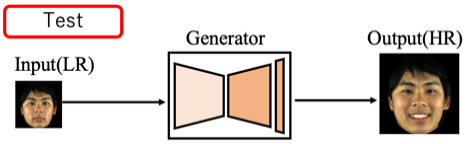

Figure 6. Overview of the multitask-based parallel model during inference.

III. The Proposed Method (Multitask Model with SGEs)

A. Overview

In this paper, we propose a multitask-based model with additional SGEs [8, 9], which extracts high-frequency features by cropping patches from the original high-resolution image. In the proposed method, two SGEs are introduced on both sides of the generation and SR models, respectively. The SGE is introduced to ensure that each model (generation and SR models) can generate highly accurate images. Fig. 7 shows the network for training, and Fig. 8 shows the network for inference. Two encoders are applied in this study. The *Encoder* is used for image generation, which inputs a low-resolution expressionless image. The *SGEs* on the Pix2Pix and SR sides are used to guide high-frequency information, whose inputs are patch images of the high-resolution expressionless image of the same size as the low-resolution image. The outputs (features) of two encoders are combined and fed into the decoder to generate a high-resolution expression image. We used a model in which the encoders of the two streams (generation and SR models) are common and share weights during training. Since the SGE can extract features of the original high-resolution image and input it as guidance, the proposed method can generate high-resolution facial expression images from low-resolution expressionless facial images during inference with a single-generation model without using SR networks. Compared with the conventional Pix2Pix and multitask-based parallel approaches, the proposed method can significantly capture high-resolution information by introducing additional SGE. Compared with the existing two-step serial approaches (Pix2Pix + SR), the proposed method is a single-step method like Pix2Pix during inference.

The generator parts of the generation and SR models use the U-net structure, which consists of an encoder, decoder, and skip connection. Additionally, the SGE uses the U-net structure with a shared decoder, where features obtained from the SGE are input to each layer of the corresponding decoder in stages. This allows information from deeper and shallow layers to be added, which is expected to improve the image generation accuracy and super-resolution accuracy.

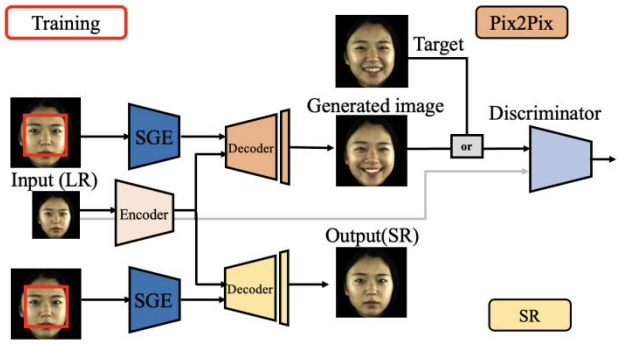

Figure 7. Overview of the multitask-based model with the SGE model for training.

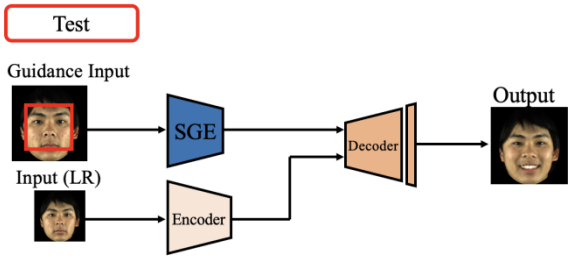

Figure 8. Overview of the multitask-based model with the SGM model for inference.

B. Self-Supervised Guidance Encoder

The SGE generates highly accurate images by extracting high-frequency features from the original high-resolution image and inputting them into the decoder [8, 9]. Due to limited computational resources, the encoder's

input is a downsampled image of the original high-resolution image. Therefore, the advantage is that the original high-resolution image can be used. For the patch portion, the center portion (1/4 of the original image) is cropped instead of randomly. In this paper, the guidance image (256×256×3) is cropped from the high-resolution image (512×512×3). The SGE consists of multiple residual blocks and MaxPool. The output size of the intermediate layer is adjusted so that the obtained features can be input to each layer of the decoder. The residual block is comprised of 3×3 and 1×1 convolutional layers. The network structure of the SGE and the residual block is shown in Fig. 9.

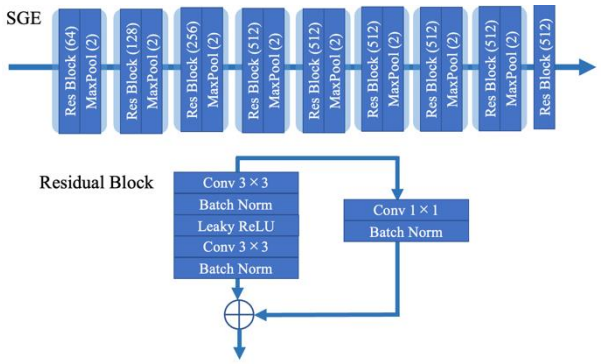

Figure 9. The network structure of the SGE and residual block.

C. Contributions of the Proposed Method

The main contributions are summarized as follows:

(1) The proposed method uses two encoders (the generation model encoder and the SGE on Pix2Pix), which is expected to improve the accuracy of the generation results without using the SR network during inference. The generation model encoder uses a low-resolution image that is downsampled from a high-resolution image due to limited computational resources, whereas the SGE has the advantage of using the original high-resolution image. Additionally, although this paper introduces SGE to the baseline Pix2Pix (cGAN), the proposed framework is generic and can be integrated into any GAN-based model.

(2) Adopting the U-net structure for SGE makes improving the overall accuracy and the accuracy of detailed edge information possible. Simply combining features obtained from the generation model encoder and SGE and inputting them into the decoder is insufficient in generating highly accurate images. This could be because only the deep layer features obtained using SGE are used, which can be expected to improve the accuracy of small-edge information, such as around the eyes and nose, but does not add shallow layer information, such as structural information. Therefore, to improve overall blur and small-edge information, the proposed method adopts a U-net structure in which features obtained from the SGE are directly inputted to a decoder and combined stepwise in each layer of the corresponding decoder. This allows for the addition of information from the deeper (detailed information, such as eyes and teeth) and shallow layers (large structural information, such as facial contours), thereby improving the overall accuracy and the accuracy of detailed parts.

D. The Objective Function in the Proposed Method

In this study, we added the objective function of SR (Eq. (4)) to that of Pix2Pix (Eq. (1)). Consequently, the objective function to be optimized in the proposed method is shown in Eq. (6). G, D, and S represent the generator, discriminator, and SR, respectively. Notably, λ is a hyperparameter. In this research, λ is selected as 100 based on experiments.

$$\min_{G,S} \max_{D} L_{total}(D,G,S) = L_{P2P}(D,\ G) + \lambda L_{SR}(G,S) \quad (6)$$

IV. Experiment and Comparison

A. Preprocessing for Database Creation and Data Augmentation

This experiment required an image dataset of facial expression changes. The subjects were asked to change their facial expressions from straight faces to smiling faces in the 74 frames. For all 74 frames, the face images were extracted, except during the expression change. Frames 1 to 10 were straight faces before the change of expression, and Frames 65 to 74 were smiling faces after the change of expression was complete (Fig. 10) The 1st and 65th frames, the 2nd and 66th frames, and the 10th and 74th frames are images of straight faces and their corresponding smiling faces, respectively. Although there were minute changes in the images, the degree of facial expression is the same and can be regarded as a type of data augmentation.

Additionally, data augmentation was performed on each data. The original image obtained by shooting was 1200×1600 pixels, and the central part was cropped for experiments because the outer parts are unnecessary. The cropping area is shown in Fig. 11. Suppose the distance between the eyes was d, the cropping area is defined as 3.2 d from the top of each eye to the edge of the image and 6.4 d in both length and width (Fig. 11). Subsequently, the image was resized to ensure that 6.4 d was 512 pixels in height and width. A normal random number with mean 0 and standard deviation 1 was generated for each coordinate and size to be cropped, and data augmentation was performed by shifting the cropping position. Moreover, we added a horizontally flipped image with a probability of 1/2 and a scaled image to increase the training data by 20 times.

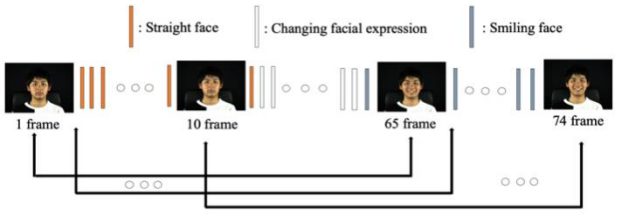

Figure 10. Overall flow of dataset creation.

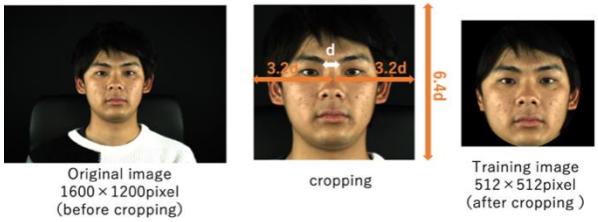

Figure 11. Cropping process of dataset.

B. Experimental Details

Pairs of low-resolution expressionless images (input image) and their high-resolution expression images (teacher image) were used for training. Pairs of 41 people (16 males and 25 females) were used as training data. We performed data augmentation to increase the training data by 20 times. Therefore, we used 8,200 pairs of input and teacher images (41 people × 10 frames × 20 times data augmentation). Only low-resolution expressionless images (one male and one female image), which are not included in the training dataset, were used as the test dataset. Experimental conditions are shown in Table I.

All the models are optimized by Adam [14], with the initial learning rate set to 2e–4.

TABLE I. EXPERIMENTAL CONDITIONS

Epoch	30
Batch size	20
Input size	256 × 256 pixels
Output size	512 × 512 pixels

C. Results

The experimental results, representing images of the input, output from the test image, and ground truth, are shown in Fig. 12. The image size in (a) is 256×256 pixels, and that in (b) and (c) is 512×512 pixels. Although some parts of the image, such as the boundary lines of the teeth, are a little less clear than the ground truth, we generated a high-resolution facial expression image from a low-resolution expressionless facial image.

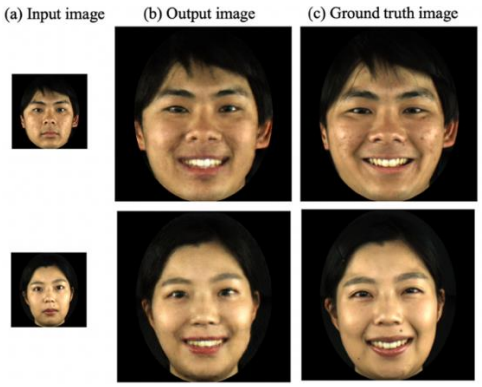

Figure 12. Generated high-resolution facial expression images. Top: the male sample; bottom: the female sample. Left: input image; middle: output image; right: ground truth image.

D. Comparison and Evaluation

We compared the proposed method with existing methods in Fig. 13. Fig. 13(a)–(e) are the output results of (a) the low-resolution images generated by Pix2Pix and interpolated and enlarged by bilinear [2], (b) the two-step serial model [5], (c) the end-to-end serial model [6], (d) the multitask learning–based parallel model [7], and (e) the proposed method. The existing serial model is confirmed to not only fail to generate small-edge information, such as around the nose and teeth, but also result in an overall blurry image. Additionally, the existing parallel model produces images with reduced blurring overall but does not produce fine details well. Conversely, we confirmed that the proposed method adequately generates the small-edge areas around the nose, especially the teeth, and produces an image with limited blurring overall.

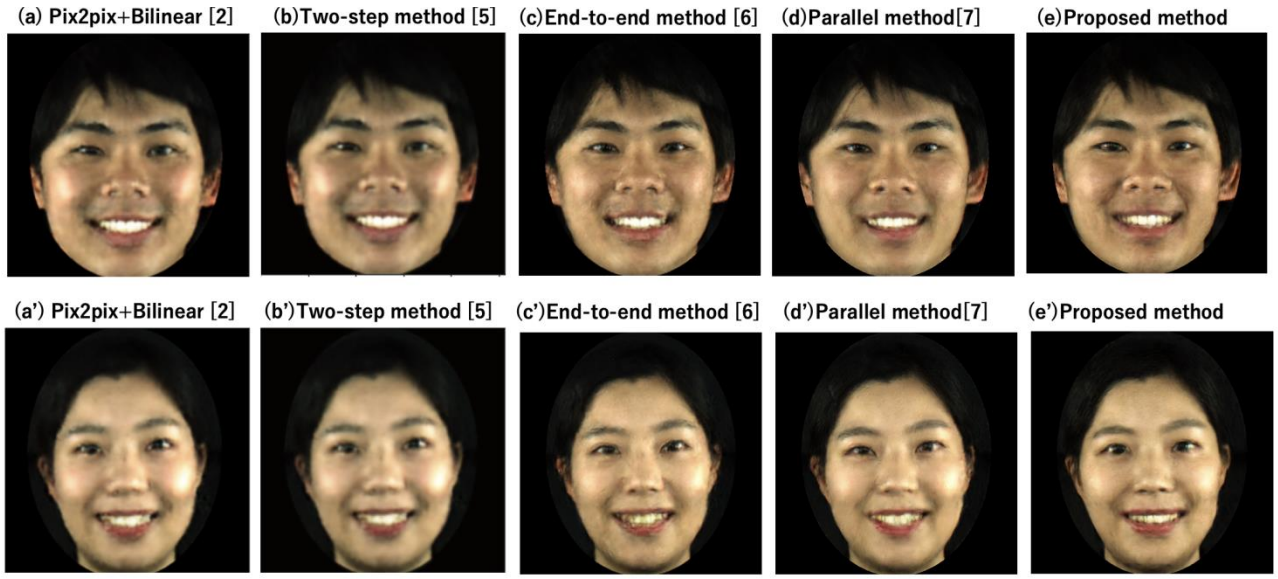

Figure 13. Comparison of the proposed method with existing methods for the male ((a)–(e)) and female samples ((a')–(e')).

We evaluated the image quality using the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). When used as an indicator of image quality evaluation, the PSNRs of the original and generated images are calculated. The formula for PSNR is shown in Eq. (7), where MSE is the mean squared error of the differences between the pixels of the reconstructed image and the ground truth image. MAX is the maximum possible pixel value in the image. The formula for SSIM is shown in Eq. (8), where the reconstructed image is represented by *x* and the ground truth image is represented by *y*. By measuring the mean (μ_x, μ_y), variance (σ_x^2, σ_y^2), and covariance (σ_{xy}) of neighboring pixels based on brightness, contrast, and structure, the SSIM index contains correlation with both individual pixels and pixels nearby. The PSNR index is very sensitive to shifting pixel positions. The PSNR score reduces dramatically even for a single pixel shift. Whereas the SSIM index takes into account the surrounding pixels for calculating the score, it is invariant to pixel shift.

Quantitative comparison using PSNR and SSIM indexes are shown in Table II, and the proposed method is shown to have achieved the best result. The PSNR was improved by 0.25 dB for the male test data and 0.28 dB for the female test data compared with the end-to-end parallel method [7], which is the second-best result. In addition, the SSIM was improved by 0.01 for the male test data and 0.007 for the female test data. We demonstrated the effectiveness of SGE from both subjective and quantitative assessments.

$$PSNR = 10 \log_{10} \frac{MAX^2}{MSE} \tag{7}$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8}$$

TABLE II. QUANTITATIVE COMPARISON OF ACCURACY (PSNR AND SSIM) AND NUMBER OF CNNS (DURING INFERENCE)

Method	Number of CNNs	PSNR (Male / Female)	SSIM (Male / Female)
Pix2Pix + Bilinear [2]	1	23.79 dB / 25.81 dB	0.809 / 0.851
Two-step method (Pix2Pix + SR) [5]	2	23.93 dB / 25.84 dB	0.814 / 0.853
End-to-end serial method (Pix2Pix + SR) [6]	2	24.39 dB / 26.17 dB	0.820 / 0.858
End-to-end parallel method (Pix2Pix + SR) [7]	1	24.68 dB / 26.41 dB	0.822 / 0.862
Proposed method	1	24.93 dB / 26.69 dB	0.832 / 0.869

E. Comparison of Computational Costs

We calculate the computational cost during the inference to observe the model complexity. We report the number of model parameters and compared it with other methods. Table III presents the results. The conventional Serial method [6] uses SRCNN as the SR network. The SR network used in the proposed method is an Encoder and Decoder structure. The number of parameters is compared with methods which have the same SR network structure. The proposed method has 44% more parameters during inference than [2] and [7], since the designed SGE Encoder module is incorporated. However, our proposed method with SGE modules has 27% fewer parameters than the Serial method.

TABLE III. COMPUTATIONAL COST ANALYSIS

Method	Parameters (During inference)
Pix2Pix + Bilinear [2]	54,414,979
Serial method (Pix2Pix + SR) [6]	108,890,918
Parallel method (Pix2Pix + SR) [7]	54,475,939
Proposed method	78,743,102

F. Ablation Study

We perform ablation studies to examine the efficiency of the designed modules employing different components. We incorporate the designed SGE module with the Pix2Pix side (Model-2) and SR side (Model-3). We also report the results of experiments without (Model-4) and with the skip (residual) connection (Model-5). The PSNR and SSIM scores for these various component configurations are shown in Table IV. In comparison to the baseline method, the accuracy of Models 2 and 3 has improved. This illustrates the importance of the designed Self-Supervised Guidance Encoder (SGE), which automatically focuses on the crucial feature information. When compared to the baseline, the performance is poor when there are no skip (residual) connections, proving that simply passing the feature maps from the SGE Encoder to the Decoder is insufficient to emphasize the important information. The finest performance demonstrates the necessity of skip connections between the SGE Encoder and Decoder (Model-5). The skip connections enable efficient flow of both shallow layer feature information (finer attributes) and deep layer feature information (coarse attributes), resulting in improved accuracy.

TABLE IV. ABLATION EXPERIMENTS TO DEMONSTRATE THE EFFICIENCY OF VARIOUS COMPONENTS USING PSNR AND SSIM INDEX

	Parallel method [7] (baseline)	SGE (Pix2Pix side)	SGE (SR side)	skip connection	PSNR (Male / Female)	SSIM (Male / Female)
Model- 1	✓				24.68dB / 26.41dB	0.822 / 0.862
Model-2	✓	✓		✓	24.80dB / 26.51dB	0.824 / 0.865
Model-3	✓		✓	✓	24.83dB / 26.55dB	0.828 / 0.865
Model-4	✓	✓	✓		24.49dB / 26.28dB	0.825 / 0.862
Model-5 (Proposed method)	✓	✓	✓	✓	24.93dB / 26.69dB	0.832 / 0.869

V. Conclusion

In this paper, we proposed an accurate and efficient multitask learning—based method using an SGE to generate high-resolution facial expression images. The SGE was effective in improving the accuracy of the generated results and achieved higher accuracy than the existing method. We experimented with Pix2Pix as our baseline generation model to validate the effectiveness of the SGE in this paper, but the SGE can also be implemented in any GAN-based generation model.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Tatsuya Hanano, Yen-Wei Chen and Masataka Seo conducted the research; Tatsuya Hanano wrote the paper; Yen-Wei Chen and Masataka Seo revised the paper. All authors had approved the final version.

Funding

This work was supported in part by the Grant in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant Nos. 21H04903, 21K11936 and 20K11867.

Acknowledgment

The authors wish to thank Mr. Hongyi WANG of Zhejiang University, China for his advice on self-supervised guidance. The authors also wish to thank Dr. Rahul Jain for his kind English proof.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efos "Image-to-image translation with conditional adversarial networks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5967–5976.

[2] Y. Kawai, M. Seo, and Y.-W. Chen, "Automatic generation of facial expression using generative adversarial nets," in *Proc. 8th Global Conference on Consumer Electronics*, 2018, pp. 278–280.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[4] Y. Choi *et al.*, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8789–8797.

[5] Y. Kawai, M. Seo, and Y.-W. Chen, "A static2dynamic GAN model for generation of dynamic facial expression images," in *Proc. IEEE 38th International Conference on Consumer Electronics*, USA, 2020.

[6] T. Hanano, M. Seo, and Y.-W. Chen, "Automatic generation of high-resolution facial expression images with end-to-end models using Pix2Pix and super-resolution convolutional neural network," in *Proc. Global Conference on Consumer Electronics*, 2021, pp. 798–801.

[7] T. Hanano, M. Seo, and Y.-W. Chen, "Accurate and efficient generation of high-resolution facial expression images by multi-task learning using generative adversarial networks," in *Proc. Global Conference on Consumer Electronics*, 2022.

[8] H. Wang *et al.*, "Patch-free 3D medical image segmentation driven by super-resolution technique and self-supervised guidance," in *Proc. of International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)*, LNCS 12901, 2021, pp. 131–141.

[9] T. Hanano, M. Seo, and Y.-W. Chen, "An improved cGAN with self-supervised guidance encoder for generation of high-resolution facial expression images," in *Proc. of IEEE 41st International Conference on Consumer Electronics (ICCE)*, 2023.

[10] M. Mirza and S. Osindero. "Conditional generative adversarial nets," arXiv preprint arXiv: 1411. 1784, 2014

[11] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. International Conference on Neural Information Processing System*, 2014, pp. 2672–2680.

[12] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)*, Springer, Cham, 2015, pp. 234–241.

[13] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," arXiv:1609.05158v2, Sep. 2016.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations*, 2015.