



US-GAN: on the importance of ultimate skip connection for facial expression synthesis

Arbish Akram¹ · Nazar Khan¹

Received: 14 April 2022 / Revised: 1 September 2022 / Accepted: 6 April 2023 /
Published online: 6 June 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

We demonstrate the benefit of using an ultimate skip (US) connection for facial expression synthesis using generative adversarial networks (GAN). A direct connection transfers identity, facial, and color details from input to output while suppressing artifacts. The intermediate layers can therefore focus on expression generation only. This leads to a lightweight US-GAN model comprised of encoding layers, a single residual block, decoding layers, and an ultimate skip connection from input to output. US-GAN has $3\times$ fewer parameters than state-of-the-art models and is trained on 2 orders of magnitude smaller dataset. It yields 7% increase in face verification score (FVS) and 27% decrease in average content distance (ACD). Based on a randomized user-study, US-GAN outperforms the state of the art by 25% in face realism, 43% in expression quality, and 58% in identity preservation.

Keywords Facial expression synthesis · Generative adversarial network · Skip connection, Image-to-image translation, Residual block

1 Introduction

Facial expression synthesis is an image-to-image translation task, which aims to change the expression of a given image to a desired one. Photorealistic facial expression synthesis can be useful for data augmentation to train face and expression recognition models. It is also useful for producing animations for human-computer interaction and entertainment. Image synthesis has received considerable attention with the arrival of generative adversarial networks (GANs) [13] and its conditional variant [24]. A GAN consists of two networks, a generator and a discriminator. It employs an adversarial learning scheme to train both networks. The generator tries to fool the discriminator by producing realistic-looking fake images while the discriminator learns to distinguish real images from fake ones. GANs have

✉ Arbish Akram
arbishakram@pucit.edu.pk

Nazar Khan
nazarkhan@pucit.edu.pk

¹ Department of Computer Science, University of the Punjab, Lahore, Pakistan

been utilized to solve many image related problems including image synthesis [26], image-to-image translation [16, 36], image super-resolution [3, 5], and facial attribute editing [12, 33, 35]. Recent GAN-based methods have shown impressive results on facial expression synthesis tasks [6, 27, 31] as well. Despite producing photorealistic and plausible results, these models have two limitations. First, they require large datasets to synthesize photorealistic expressions. When trained on smaller datasets, [1, 6, 18], and [4] have shown that existing models introduce color degradation and noticeable artifacts in the synthesized expressions. Second, the majority of existing methods [6, 7, 27, 33] share a common architecture [17] in their generator which incurs a significantly large computational cost and prohibits deployment on resource-constrained devices.

In this paper, we aim to build a simpler, smaller, and effective facial expression synthesis model called Ultimate Skip Generative Adversarial Network (US-GAN). The generator of the proposed US-GAN consists of four modules: encoding layers, single residual block, decoding layers and an ultimate skip connection as illustrated in Fig. 1. The encoding layers enable the network to encode key facial details. The residual block helps to refine these details. The decoding layers then decode high-level facial details from this latent encoding. Finally, we propose to directly link the input image to an output image with a skip connection, called the ultimate skip connection. We hypothesize that the inclusion of an ultimate skip connection in the generator of a GAN will lead to improved transfer of identity, facial details, and color details from input to output. Since the generator will be relieved from the task of producing such details, it will utilize its parameters to learn expression mappings only and not suffer from color degradation and block artifacts.

The contributions of this work can be summarized as follows:

1. Incorporation of an ultimate skip connection improves preservation of identity, facial details, and color details while inducing convincing expressions.
2. The ultimate skip connection leads to a model with three times fewer parameters that can be trained using two orders of magnitude smaller dataset than the state-of-the-art GANimation model.
3. US-GAN qualitatively outperforms the state of the art in realism, mapped expression and identity preservation by 25%, 43%, and 58% respectively.
4. Quantitatively, US-GAN improves face verification score (FVS) by 7% and reduces average content distance (ACD) by 27% compared to the state of the art.
5. US-GAN generalizes well on out-of-dataset facial images.

The rest of the paper is structured as follows. Section 2 reviews state-of-the-art image-to-image translation and facial expression synthesis models. The proposed US-GAN

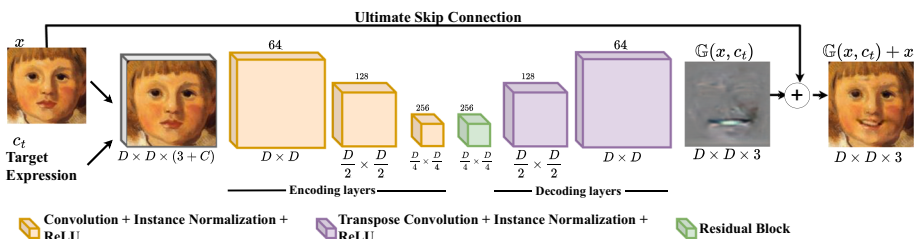


Fig. 1 The generator of the proposed framework. US-GAN takes a face image with any expression and the target expression label as input and produces the output image with the target expression. It consists of four modules: i) encoding layers, ii) single residual block, iii) decoding layers, and iv) an ultimate skip connection

architecture is explained in Section 3 along with loss functions, datasets, and evaluation metrics. Qualitative as well as quantitative experimental results are presented in Section 4. Section 5 discusses the effectiveness of our proposed method and we conclude in Section 6.

2 Related work

2.1 Image-to-image translation

Generative Adversarial Network (GAN) [13] and its conditional variant (cGAN) [24] have been successfully utilized for image synthesis and image-to-image translation problems. The Pix2pix model [16] employed a conditional GAN with ℓ_1 image reconstruction loss to enforce generated samples to be close to target images. In order to learn the mapping between two domains using paired datasets, they utilized U-Net [29] and PatchGAN [16] in the generator and discriminator, respectively. Zhu et al. [36] introduced a cycle consistency loss to perform cross-domain mapping using an unpaired dataset. Their network, named CycleGAN, contains two generators and two discriminators to learn the cyclical, cross-domain mappings. They adopted Johnson et al.'s architecture [17] in their generator. While Pix2pix and CycleGAN can be used to learn the mappings between facial expressions, these networks fail to produce realistic expressions as demonstrated in [18] where the smaller size of facial expression synthesis datasets is suggested as a possible reason for their failure.

2.2 Facial expression synthesis

GANs have been widely used for facial expression synthesis. Ding et al. [8] proposed ExprGAN to synthesize facial expressions with controllable intensity. However, it fails to preserve the identity details of the input image. The GC-GAN model [28] induces a desired facial expression on an input image. To learn mappings among multiple expressions, StarGAN [6] learns mappings among multiple expressions using a single, shared generator. When trained on a small dataset, the authors reported color degradation artifacts. The GANimation model [27] trained on the large EmotioNet [10] dataset extracts action units from a target face and transfers them to an input face. The quality of expressions is highly dependent on the extracted action units. Liu et al [21] proposed an encoder-decoder architecture, called STGAN, with symmetric skip connections and selective transfer units for facial attribute manipulation. The Cascade-EF GAN [34] generates sharper and realistic images by employing local and global attention. It focuses in a progressive manner. For out-of-dataset images, it was fine-tuned on the large AffectNet [25] dataset. In summary, current state-of-the-art, GAN-based, facial expression synthesis models require larger datasets for inducing satisfactory expressions on in- and out-of-dataset images. Quality significantly degrades when trained on smaller datasets. In contrast, our proposed US-GAN method produces realistic expressions on both in- and out-of-dataset images by employing only hundreds of images for training.

3 Materials and methods

In this work, we aim to learn a mapping to transform an input image $x \in \mathcal{R}^{D \times D \times 3}$ with known original expression c_o , to an output image $y \in \mathcal{R}^{D \times D \times 3}$ with a target expression

c_t . Both expression vectors c_o and c_t are 1-hot encodings from C expression classes. The proposed US-GAN consists of two modules: a generator and a discriminator.

3.1 Generator

Figure 2 presents an overview of the complete US-GAN pipeline consisting of the following four modules.

3.1.1 Encoding layers

There are three encoding layers in our proposed network. The first layer takes an input volume x of size $D \times D \times (3 + C)$ and generates a feature volume of size $D \times D \times 64$. The second encoding layer takes this feature volume, performs strided convolution for downsampling and produces an output volume of size $\frac{D}{2} \times \frac{D}{2} \times 128$. The third encoding layer takes this downsampled feature volume, performs strided convolution again and provides an encoded volume of size $\frac{D}{4} \times \frac{D}{4} \times 256$.

3.1.2 Residual block

Residual connections [14] assist deep networks in learning better representations by directly transferring a layer's input to a later layer's output so that only the residual transformation needs to be modelled. This makes the learning task easier. For expression synthesis, residual connections help to preserve facial details. We employ only one residual block in the body of our network. The residual block consists of two 3×3 convolutional layers of 256 channels, two instance normalization layers [32] and a ReLU layer [11]. Each convolutional layer is followed by an instance normalization layer. ReLU activation function is applied to the output of the first instance normalization layer. The input of a block is directly added after the second instance normalization layer.

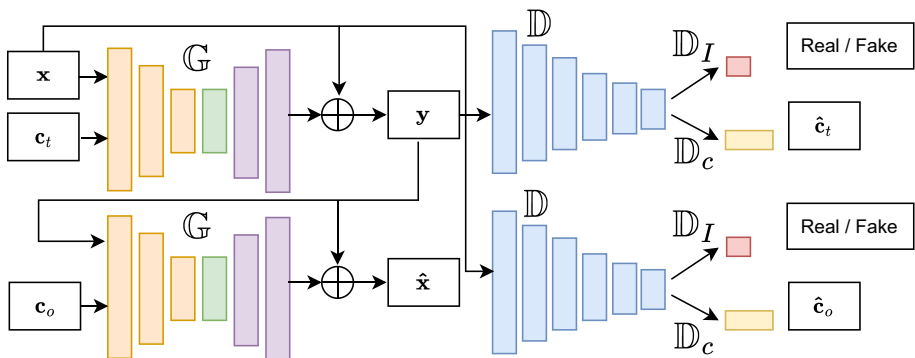


Fig. 2 US-GAN consists of two modules: a generator (G) and a discriminator (D). G takes an input image (x) and a target expression vector (c_t) and generates an output image (y) with the target expression. The same network G is then used to reconstruct the input image \hat{x} from y and original expression vector (c_o). Discriminator head D_I learns to distinguish real images x from fake images y while head D_c classifies discriminator input into an expression class

3.1.3 Decoding layers

The first decoding layer takes the $\frac{D}{4} \times \frac{D}{4} \times 256$ volume produced by the residual block and performs fractionally-strided convolutions to produce an upsampled volume of size $\frac{D}{2} \times \frac{D}{2} \times 128$. The second decoding layer takes these upsampled features and applies fractionally-strided convolution again to produce an upsampled volume of size $D \times D \times 64$. The last layer transforms these feature maps into an output image $\mathbb{G}(x, c_t)$ of size $D \times D \times 3$.

3.1.4 Ultimate skip connection

For large training sets, it is possible to produce rich enough encodings and powerful enough decoders [6]. However, for smaller training sets, if the input is transferred directly to the output via an ultimate skip connection, then the encoding and decoding tasks become easier. The output of the generator is obtained by adding the input to the output of the last decoding layer as

$$y = \mathbb{G}(x, c_t) + x. \quad (1)$$

This allows the encoding and decoding layers to focus purely on the expressions as can be seen in Fig. 3. Since input details relating to identity, facial features, and overall color are already transferred via the ultimate skip connection, the parameters of the generator only learn to produce the residual expression $\mathbb{G}(x, c_t)$. Similar ideas have been explored for image restoration [23] and facial attribute editing [30].

3.2 Discriminator

The discriminator \mathbb{D} transforms its input image into a feature volume of size $\frac{D}{26} \times \frac{D}{26} \times 2048$ through a sequence of six layers of 4×4 strided convolution filters. Starting from 64 channels in the first layer, each convolution layer doubles the the number of channels. Each convolution is followed by LeakyReLU activation. The feature volume after the sixth layer is converted into a score \mathbb{D}_I that is interpreted as the chance of the discriminator's

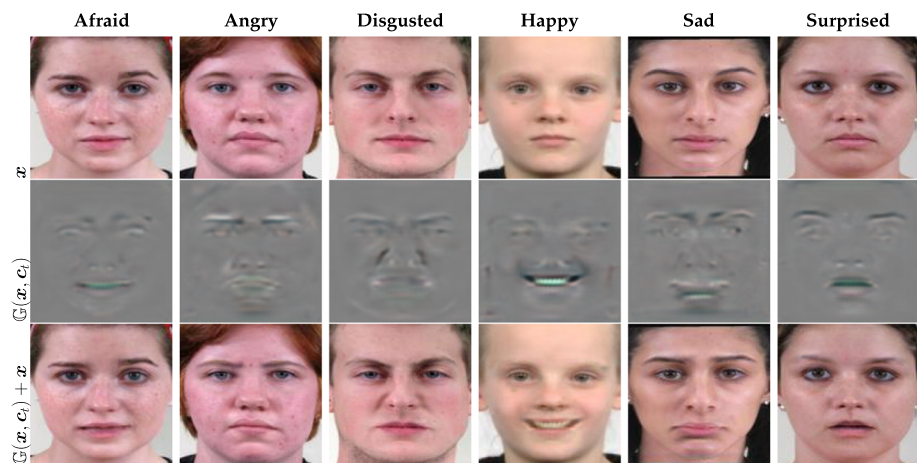


Fig. 3 Illustration of residual and final output image synthesized by our proposed US-GAN. The network is encouraged to generate only expression-related details since the ultimate skip connection carries over identity, facial, and color details directly from the input image

input being a real image. In parallel, the volume is also converted into a $C \times 1$ vector of probabilities representing expression of the discriminator's input image.

3.3 Loss formulations

The proposed model is trained by minimizing a combination of three loss functions.

3.3.1 Adversarial loss

In order to avoid training instability and generate higher quality images, we used Wasserstein adversarial loss [2], defined as

$$\mathcal{L}_A = E [\mathbb{D}_I(\mathbf{x})] - E [\mathbb{D}_I(\mathbf{y})] - \lambda_{gp} E [(\|\nabla_{\bar{\mathbf{x}}}\mathbb{D}_I(\bar{\mathbf{x}})\|_2 - 1)^2], \quad (2)$$

where $\mathbb{D}_I(\cdot)$ is proportional to the the probability of it's input image being real, λ_{gp} is the gradient penalty coefficient, and $\bar{\mathbf{x}}$ is a uniform random linear combination of \mathbf{x} and \mathbf{y} .

3.3.2 Image reconstruction loss

Let $\hat{\mathbf{x}}$ be the reconstruction of the original image \mathbf{x} generated from the fake expression \mathbf{y} as

$$\hat{\mathbf{x}} = \mathbb{G}(\mathbf{y}, \mathbf{c}_o) + \mathbf{y} \quad (3)$$

In order to softly enforce that the face in the input \mathbf{x} and the generated image \mathbf{y} correspond to the same person, we utilize the cycle reconstruction loss [36] defined as

$$\mathcal{L}_R = E [\|\mathbf{x} - \hat{\mathbf{x}}\|_1] \quad (4)$$

3.3.3 Expression classification loss

We define a multiclass cross-entropy loss between target expression \mathbf{c}_t and classified expression $\hat{\mathbf{c}}_t$ of the generated image \mathbf{y} as

$$\mathcal{L}_C^F = E [-\log \mathbb{D}_c(\mathbf{c}_t | \mathbf{y}, \hat{\mathbf{c}}_t)]. \quad (5)$$

and between original expression \mathbf{c}_o and classified expression $\hat{\mathbf{c}}_o$ of the input image \mathbf{x} as

$$\mathcal{L}_C^R = E [-\log \mathbb{D}_c(\mathbf{c}_o | \mathbf{x}, \hat{\mathbf{c}}_o)] \quad (6)$$

The idea is to penalize any deviation between target and classified expression of \mathbf{y} and between original and classified expression of \mathbf{x} .

3.3.4 Overall GAN objectives

The overall objective functions for discriminator \mathbb{D} and generator \mathbb{G} can be written as

$$\mathcal{L}_{\mathbb{D}} = -\mathcal{L}_A + \lambda_C \mathcal{L}_C^R \quad (7)$$

$$\mathcal{L}_{\mathbb{G}} = \mathcal{L}_A + \lambda_C \mathcal{L}_C^F + \lambda_R \mathcal{L}_R \quad (8)$$

where λ_C and λ_R denote weights of classification and reconstruction losses, respectively. Minimizing $\mathcal{L}_{\mathbb{D}}$ encourages the discriminator to improve it's ability to i) differentiate between real and fake images, and ii) classify the expression of it's input image. Minimizing $\mathcal{L}_{\mathbb{G}}$ encourages the generator to produce fake images that i) are hard to distinguish from real images, ii) have the desired expression, and iii) preserve the identity of the input image.

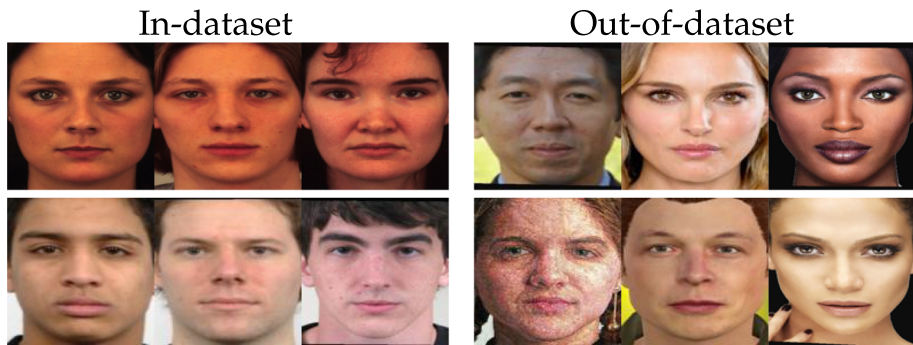


Fig. 4 Example images from in- and out-of-dataset

3.4 Implementation details

We train the proposed US-GAN model from scratch for 350 epochs using the Adam optimizer [19] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, learning rate 0.0001 and batch size of 8. Following [6], we set $\lambda_C = 1$, $\lambda_R = 10$ and $\lambda_{gp} = 10$ for all experiments.

3.5 Dataset

We trained our model using three publically available datasets KDEF [22], RaFD [20] and CFEE [9]. The KDEF dataset consists of 490 images of seven universal expressions collected from 35 male and 35 female participants. RaFD contains 8,040 facial expression images collected from 67 participants from five different angles. We used only 469 frontal images in our experiments. The CFEE dataset contains 5,060 compound facial expressions images of 230 participants. We used 1,610 images from this dataset. In total, we use 2,569 images from the three datasets for seven facial expressions. We used 90% images for training and the rest for testing. All facial images are center-cropped and resized to 128×128 . To evaluate the effectiveness of our proposed model on out-of-dataset images, celebrities, paintings and avatar images are downloaded from the Internet. These images are significantly different from the distribution of the training datasets. Some example images from in- and out-of-dataset are shown in Fig. 4.

3.6 Evaluation metrics

We compare models in terms of i) number of learnable parameters, and ii) size of training sets. We compare the quantitative performance of models in terms of the following two metrics.

1. Average Content Distance (ACD) is the squared Euclidean distance between the features $\phi(x)$ of the input image and features $\phi(y)$ of the generated image. The features $\phi(\cdot)$ are extracted using a face classifier¹.

$$ACD(x, y) = \|\phi(x) - \phi(y)\|_2^2 \quad (9)$$

¹https://github.com/ageitgey/face_recognition

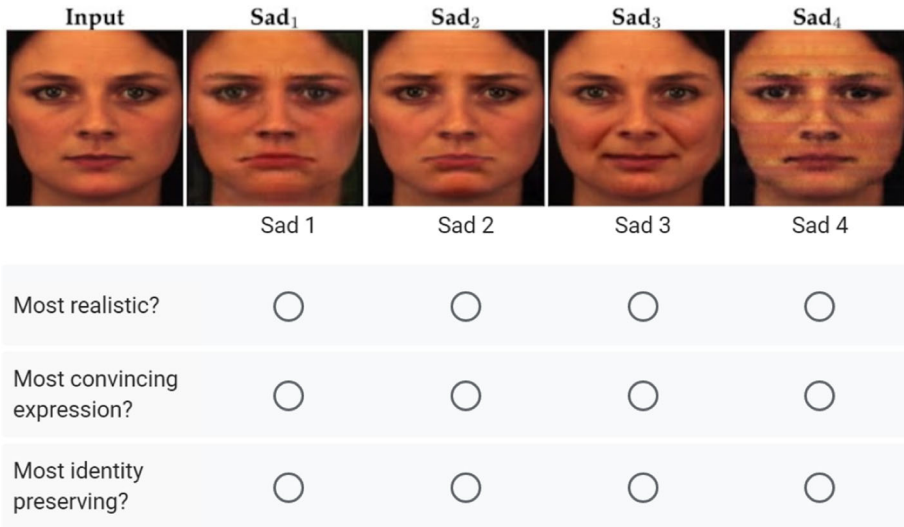


Fig. 5 A sample from the user study. One input image in neutral expression along with its four manipulated versions were placed in random order in one row and human evaluators were asked to vote for the best-synthesized image in terms of i) realism, ii) mapped expression, and iii) identity preservation

2. Face Verification Score (FVS) computes the similarity between input and synthesized images using Face++² and returns a value between 0 and 100 to indicate the likeness between two faces.

For qualitative comparison, we performed a user study to compute user preference percentages for StarGAN, STGAN, GANimation as well as our proposed US-GAN. Eighty users participated in this user study. Eighteen input face images from in- and out-of-dataset were randomly selected for evaluation. Synthesized expressions for these images were generated using StarGAN, STGAN, GANimation and the proposed US-GAN. One input image in neutral expression was placed on the left. Its four manipulated versions were placed next to it in random order (as shown in Fig. 5). The human evaluators were not made aware of the source algorithm for any manipulated image and were asked to select one image for each of the following three questions.

1. Which image looks most realistic (without considering the expression)?
2. Which image has the most convincing expression?
3. Which image best preserves the identity of the input face?

3.7 Ablation study

We conducted ablation studies to investigate the effectiveness of the *ultimate skip connection* and the *number of residual blocks in the generator* of the proposed US-GAN.

²<https://www.faceplusplus.com/>

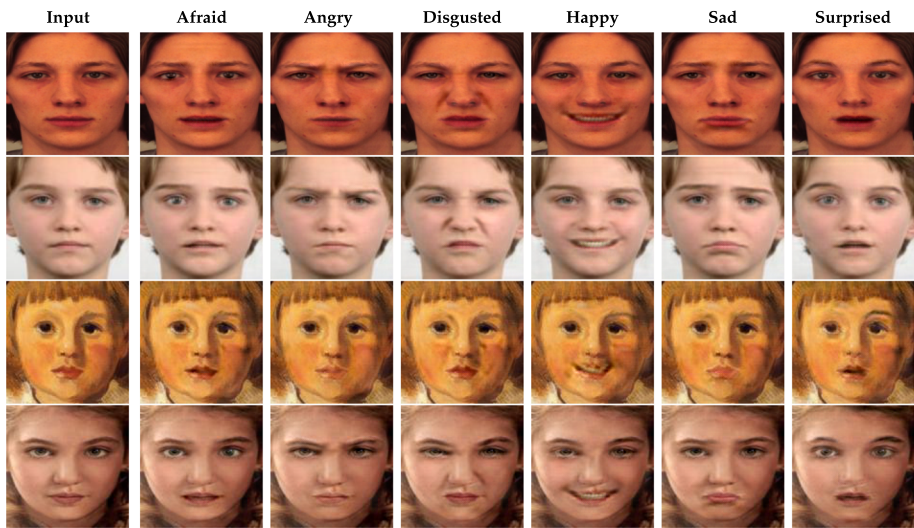


Fig. 6 Facial expression synthesis results on in-dataset (top two rows) and out-of-dataset (last two rows) testing images. The proposed method trained on RaFD, KDEF and CFEE datasets generates plausible expressions while preserving identity and retaining facial details. Despite being trained on a smaller dataset, these results demonstrate both in-dataset and out-of-dataset generalization strength of the proposed method

3.7.1 Ultimate skip connection

To qualitatively demonstrate the usefulness of the ultimate skip connection, we designed the following two experiments:

1. Train US-GAN with and without³ ultimate skip connection to observe the difference.
2. Train STGAN [21] with and without ultimate skip connection to observe the difference.

3.7.2 Number of residual blocks

In order to validate the effectiveness of residual blocks in the bottleneck of the US-GAN generator, US-GAN is trained with one and six residual blocks, respectively. We denote US-GAN model with R residual blocks as US-GAN- R , where $R \in \{1, 6\}$. The comparison with six residual blocks is motivated by the state of the art such as StarGAN [6] and GANimation [27] that uses six residual blocks.

4 Results

In this section, we conduct extensive experiments to evaluate the performance of our proposed method. We first discuss baseline details in Section 4.1. Qualitative as well as quantitative evaluation on these results are presented in Sections 4.2 and 4.3.

³Technically, the model should no longer be called US-GAN in this case.

4.1 Baselines

We compare our proposed US-GAN with three state-of-the-art, multi-domain, facial expression synthesis models, StarGAN [6], STGAN [21] and GANimation [27]. For StarGAN and STGAN, we use the code and hyperparameter settings provided by the authors and trained on the same combined dataset (KDEF, RaFD and CFEE) used to train US-GAN. For GANimation, we used a model pre-trained on the large EmotioNet dataset [10] for 30 epochs.

4.2 Qualitative evaluation

Generalization of US-GAN on out-of-dataset imagery can be observed from Figs. 6 and 7. The ultimate skip connection helps to transfer input image details so that the network

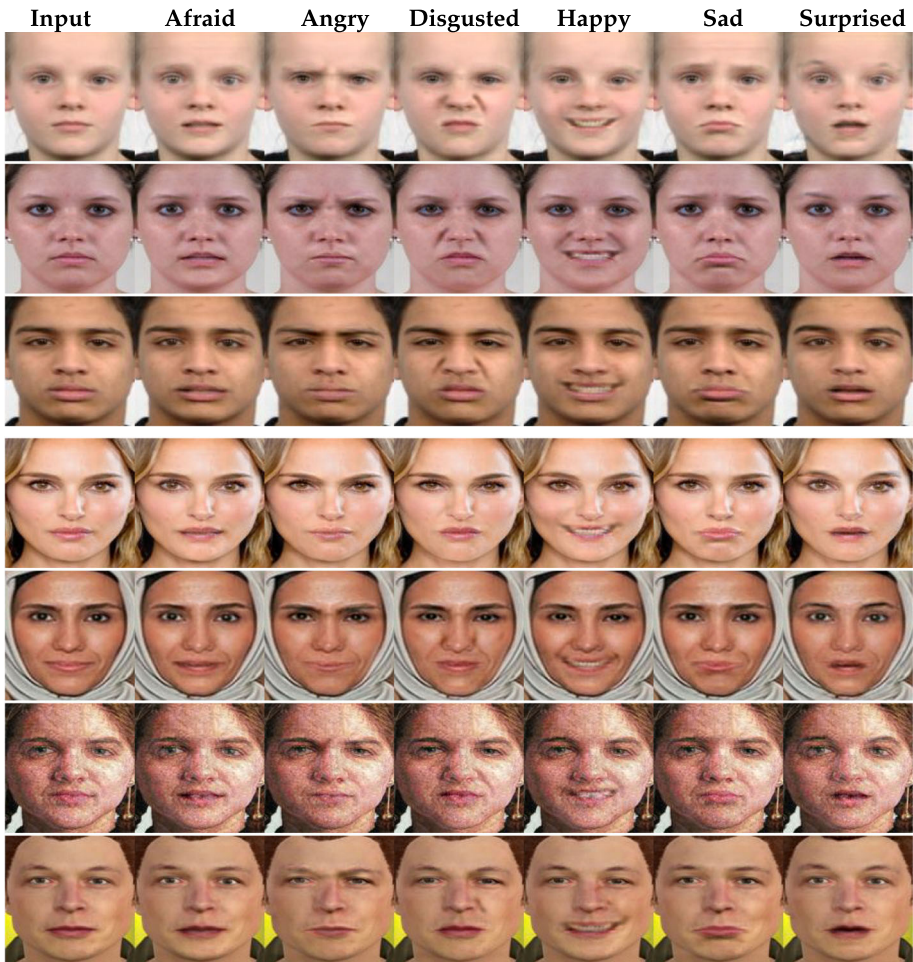


Fig. 7 Expressions synthesised by US-GAN on **Rows 1 to 3**: in-dataset, and **Rows 4 to 7**: out-of-dataset images. Due to ultimate skip connection, the proposed method preserves input image details and colors and induces convincing expressions

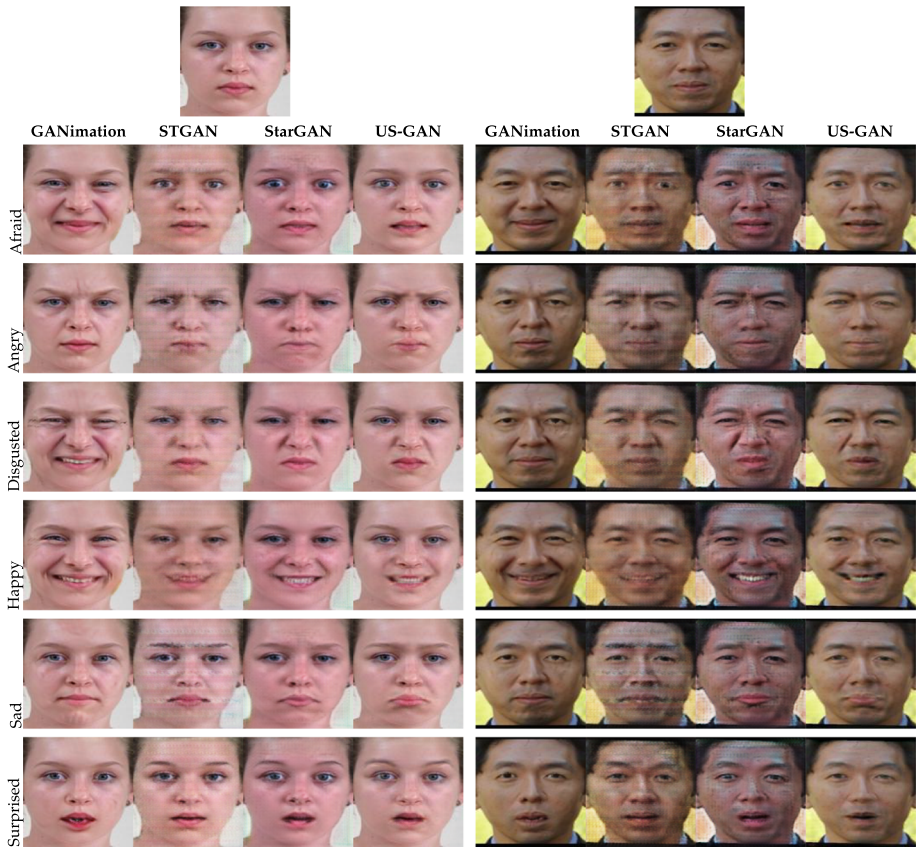


Fig. 8 Comparison of facial expression synthesis results obtained by proposed US-GAN and other state-of-the-art models. **Left:** An in-dataset image. **Right:** An out-of-dataset image. GANimation [27] introduces aging and other noticeable artifacts on all in- and out-of-dataset images. STGAN [21] introduces pseudo-periodic artifacts. StarGAN [6] introduces a pinkish bias and for out-of-dataset images, generates artifacts. The proposed US-GAN successfully synthesizes expressions while preserving identity, facial details, and color details

parameters only learn to focus on generating expressions. This leads to realistic expressions and preserved identities, facial details, and color details. Comparison with three state-of-the-art facial manipulation models including StarGAN [6], STGAN [21] and GANimation [27] is presented in Fig. 8. While existing models may induce realistic expressions on both in- and out-of-dataset images, GANimation introduces strong artifacts around the eyes, nose and mouth, StarGAN fails to recover the true input image colors, and STGAN introduces pseudo-periodic artifacts on the synthesized images. In comparison, the proposed US-GAN successfully introduces the desired expression without adding irrelevant changes.

4.3 Quantitative evaluation

Table 1 shows that our proposed method has three times fewer parameters and is trained on two orders of magnitude smaller dataset than StarGAN and GANimation. Compared to STGAN, the proposed method has an order of magnitude fewer parameters. The values of

Table 1 Compared to the state of the art, the proposed US-GAN has more than three times fewer parameters and is trained on two orders of magnitude smaller dataset. It yields the best identity preservation (lowest ACD and highest FVS) between inputs and outputs

Method	# of parameters ↓	# of images ↓	ACD ↓	FVS ↑
US-GAN	2.5×10^6	2.5×10^3	0.2832	94.19 ± 1.11
StarGAN [6]	8.5×10^6	2.1×10^5	0.5660	91.71 ± 1.95
STGAN [21]	7.4×10^7	2.0×10^5	0.3725	92.05 ± 2.81
GANimation [27]	8.5×10^6	4.1×10^5	0.3899	87.98 ± 8.67

ACD and FVS indicate that US-GAN is most effective at preservation of identity and other features of the input.

The summarized results of the user study are provided in Fig. 9. Compared to the second best performing model, US-GAN yielded 25% improvement in realism, 43% improvement in plausibility of mapped expressions, and 58% improvement in identity preservation.

4.4 Ablation study

4.4.1 Ultimate skip connection

We demonstrate in Fig. 10 that the ultimate skip connection directly leads to preservation of input image details. This is true for the proposed US-GAN method as well as for STGAN [21]. The second row contains US-GAN results. When the ultimate skip connection is removed, the third row shows that the corresponding model fails to preserve input image colors and introduces some artifacts. For STGAN as well, the fourth row shows that

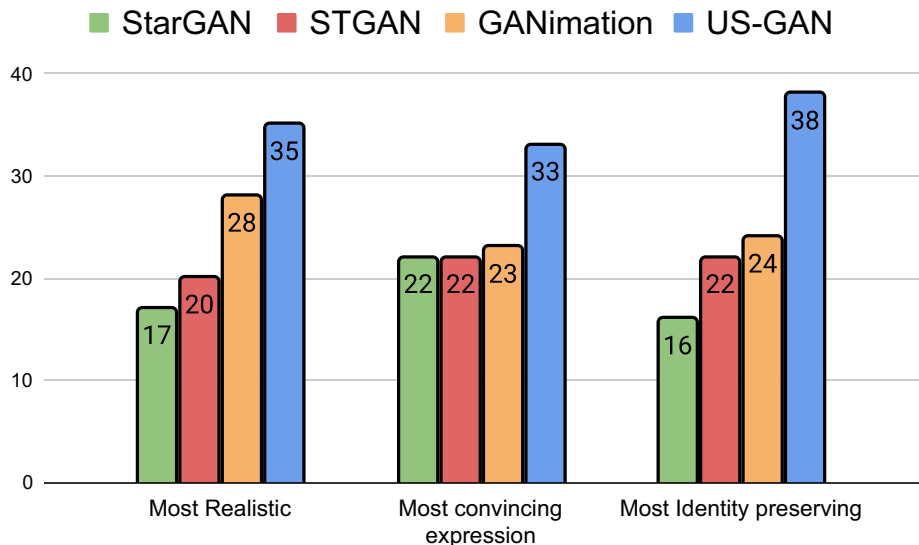


Fig. 9 User study results for facial expression synthesis. The proposed US-GAN outperforms StarGAN [6], STGAN [21] and GANimation [27] based on user preferences

the addition of an ultimate skip connection leads to clear reduction in artifacts and improves the transfer of facial details from input to output.

4.4.2 Number of residual blocks

Existing models such as StarGAN [6] and GANimation [27] use six residual blocks in the generator. Figure 11 demonstrates that, powered by the ultimate skip connection, even one residual block can produce plausible, identity- and color-preserving transformations. In other words, the ultimate skip connection reduces the need for many parameters, which can help improve generalization.

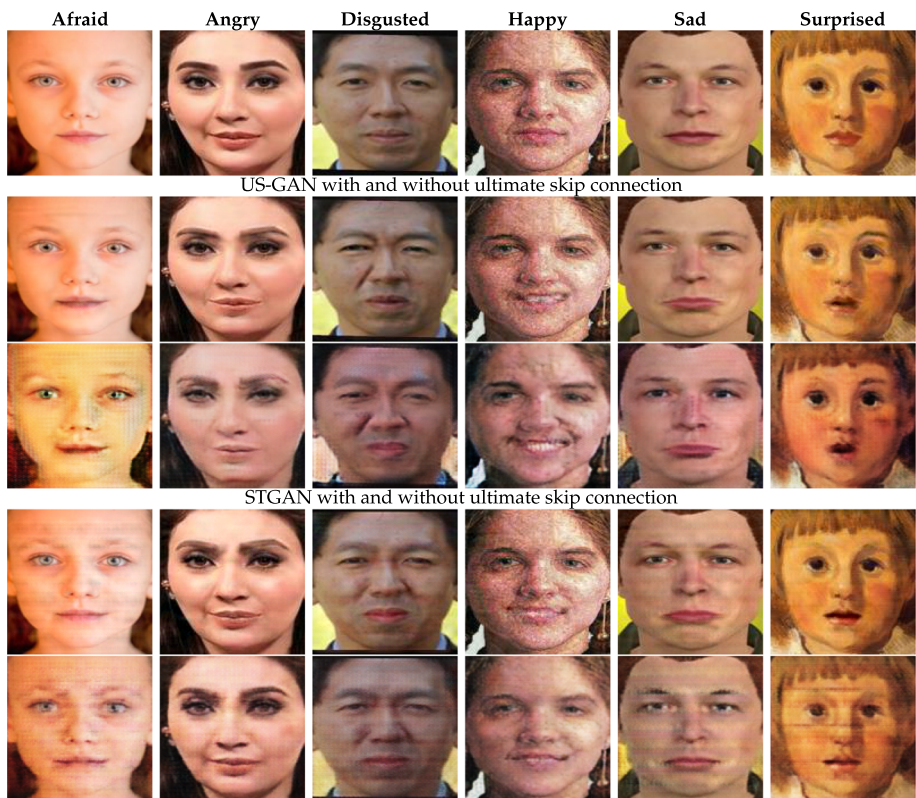


Fig. 10 Impact of ultimate skip connection on generated images. **First row:** Input images. **Row 2:** US-GAN results. **Row 3:** US-GAN trained without ultimate skip connection. **Row 4:** STGAN [21] trained after adding ultimate skip connection. **Row 5:** STGAN results. The introduction of the ultimate skip connection helps to preserve the input image as well as overall color details while introducing convincing expressions in the proposed US-GAN. In the last row, we can see that STGAN fails to preserve the color details and introduces noise-like artifacts in the synthesized images without ultimate skip connection

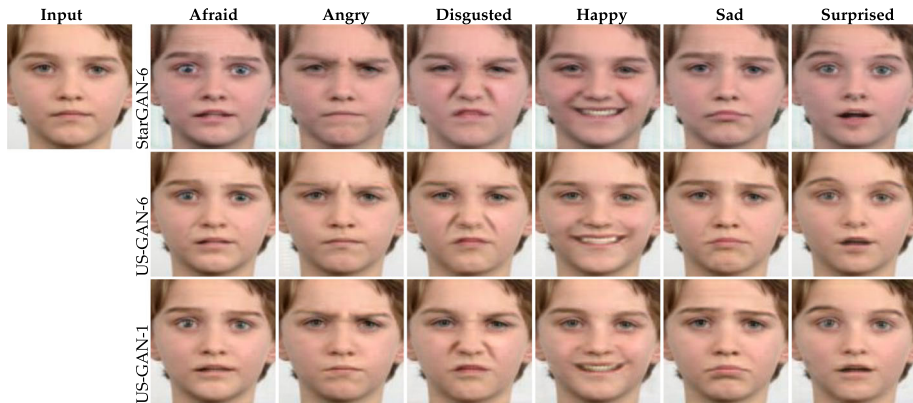


Fig. 11 **First row:** StarGAN with six residual blocks failed to preserve the true colors of an input image. **Second row:** US-GAN with six residual blocks produced sharper expressions with better preserved facial and color details. **Third row:** US-GAN with only one residual block did not suffer from significant drop in quality of results since most of the heavy lifting related to facial and color details is already carried out by the ultimate skip connection

5 Discussion and future directions

We now address a few questions raised by our results and place our results in context of existing work.

Why an ultimate skip connection? Skip connections have already been used between encoding and decoding layers [21] since they allow learning of easier residuals of intermediate tasks within deep networks. In hindsight, it seems natural to apply residual learning on the original end-to-end expression synthesis task.

The ultimate skip connection has been shown to be fundamentally important for expression synthesis since it performs the heavy lifting of transferring non-expression related details from the input to the output at no cost. This leaves the learnable parameters of the generator to focus on pure expression synthesis only.

This can be viewed from another perspective of residual learning as well. Residual learning such as the popular ResNet model [14] works by learning easier, residual sub-problems within deep network layers instead of complete transformations. By incorporating a direct ultimate skip connection from input to output, we have applied residual learning to the original, end-to-end expression synthesis problem and made it easier to solve.

Since the problem has been made easier, we have solved it using fewer parameters (only one residual block in US-GAN instead of six in competing models). As a consequence, US-GAN has been shown to have better generalization on out-of-dataset imagery.

Are residual blocks even necessary? Another question raised, but not answered, by our current work is whether residual blocks are necessary. This is because, even one residual block powered by an ultimate skip connection produced results better than StarGAN and GANimation models that both contain six residual blocks. Perhaps increasing direct skip

connections between encoding/decoding layers in the manner of UNet [29] or DenseNet [15] can be more effective than residual blocks.

What is the limitation of the proposed ultimate skip connection? Although the ultimate skip connection helps to recover the input image details, it slightly decreases the expressiveness of expressions as can be observed from Fig. 10. The ultimate skip connection helps to synthesize realistic expressions at the cost of weakened expression manipulation ability. This can perhaps be alleviated by gating the skip connection through attention mechanisms [3].

6 Conclusion

We have proposed US-GAN, a smaller and more effective model for facial expression synthesis. Our primary contribution is demonstrating the benefit of an ultimate skip connection which transfers identity, facial, and color details directly from input to output. This eases the task of the generator that can then focus on inducing expressions only. It also helps to reduce the number of learnable parameters, such as multiple residual blocks, which improves generalization. Compared to state-of-the-art models, US-GAN has more than three times fewer parameters, is trained on two times smaller dataset. Based on ACD and FVS metrics, US-GAN generates realistic expressions while best preserving identity and details of the input face. US-GAN also outperforms the state of the art in terms of image and expression realism and identity preservation based on responses from human evaluations. Our results indicate that the ultimate skip connection is fundamentally important for the facial expression synthesis task. The proposed method can potentially be extended by exploring the use of intermediate skip connections as an alternative to residual blocks, and incorporating spatial attention to improve expressiveness of results.

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Akram A, Khan N (2021) Pixel-based facial expression synthesis. In: International conference on pattern recognition. IEEE, pp 9733–9739
2. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: International conference on machine learning, pp 214–223
3. Chen C, Gong D, Wang H, Li Z, Wong K-YK (2020) Learning spatial attention for face super-resolution. *IEEE Trans Image Process* 30:1219–1231
4. Chen Y-C, Xu X, Jia J (2020) Domain adaptive image-to-image translation. In: IEEE Conference on computer vision and pattern recognition, pp 5274–5283
5. Chen Y, Tai Y, Liu X, Shen C, Yang J (2018) FSRNet: end-to-end learning face super-resolution with facial priors. In: IEEE Conference on computer vision and pattern recognition, pp 2492–2501

6. Choi Y, Choi M, Kim M, Ha J-W, Kim S, Choo J (2018) StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE Conference on computer vision and pattern recognition, pp 8789–8797
7. d'Apolito S, Paudel DP, Huang Z, Romero A, Van Gool L (2021) GANmut: learning interpretable conditional space for gamut of emotions. In: IEEE Conference on computer vision and pattern recognition, pp 568–577
8. Ding H, Sricharan K, Chellappa R (2018) ExprGAN: facial expression editing with controllable expression intensity. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
9. Du S, Tao Y, Martinez AM (2014) Compound facial expressions of emotion. *Proc Natl Acad Sci* 111(15):E1454–E1462
10. Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM (2016) EmotioNet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: IEEE Conference on computer vision and pattern recognition, pp 5562–5570
11. Fukushima K (1975) Cognitron: a self-organizing multilayered neural network. *Biol Cybern* 20(3):121–136
12. Gao Y, Wei F, Bao J, Gu S, Chen D, Wen F, Lian Z (2021) High-fidelity and arbitrary face editing. In: IEEE Conference on computer vision and pattern recognition, pp 16115–16124
13. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on computer vision and pattern recognition, pp 770–778
15. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition
16. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: IEEE Conference on computer vision and pattern recognition, pp 1125–1134
17. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European Conference on computer vision. Springer, pp 694–711
18. Khan N, Akram A, Mahmood A, Ashraf S, Murtaza K (2020) Masked linear regression for learning local receptive fields for facial expression synthesis. *Int J Comput Vis* 128(5):1433–1454
19. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)
20. Langner O, Dotsch R, Bijlstra G, Wigboldus DanielHJ, Hawk ST, Van Knippenberg AD (2010) Presentation and validation of the Radboud Faces Database. *Cogn Emotion* 24(8):1377–1388
21. Liu M, Ding Y, Xia M, Liu X, Ding E, Zuo W, Wen S (2019) Sgan: a unified selective transfer network for arbitrary image attribute editing. In: IEEE Conference on computer vision and pattern recognition, pp 3673–3682
22. Lundqvist D, Flykt A, Öhman A (1998) The karolinska directed emotional faces - KDEF, CD ROM. Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, Stockholm, Sweden
23. Mao X, Shen C, Yang Y-B (2016) Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Adv Neur Inform Process Syst* 29:2802–2810
24. Mirza M, Osindero S (2014) Conditional generative adversarial nets. [ArXiv:1411.1784](https://arxiv.org/abs/1411.1784)
25. Mollahosseini A, Hasani B, Mahoor MH (2017) AffectNet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans Affect Comput* 10(1):18–31
26. Perarnau G, van de Weijer J, Raducanu B, Álvarez JM (2016) Invertible conditional GANs for image editing. [ArXiv:1611.06355](https://arxiv.org/abs/1611.06355)
27. Pumarola A, Agudo A, Martinez AM, Sanfeliu A, Moreno-Noguer F (2020) GANimation: one-shot anatomically consistent facial animation. *Int J Comput Vis* 128(3):698–713
28. Qiao F, Yao N, Jiao Z, Li Z, Chen H, Wang H (2018) Geometry-contrastive generative adversarial network for facial expression synthesis. [ArXiv:1802.01822](https://arxiv.org/abs/1802.01822)
29. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241
30. Shen W, Liu R (2017) Learning residual images for face attribute manipulation. In: IEEE Conference on computer vision and pattern recognition, pp 4030–4038
31. Tang J, Shao Z, Ma L (2021) EGGAN: learning latent space for fine-grained expression manipulation. *IEEE MultiMedia*
32. Ulyanov D, Vedaldi A, Lempitsky V (2016) Instance normalization: the missing ingredient for fast stylization. [ArXiv:1607.08022](https://arxiv.org/abs/1607.08022)

33. Wu P-W, Lin Y-J, Chang C-H, Chang EY, Liao S-W (2019) RelGAN: multi-domain image-to-image translation via relative attributes. In: IEEE International conference on computer vision, pp 5914–5922
34. Wu R, Zhang G, Lu S, Chen T (2020) Cascade EF-GAN: progressive facial expression editing with local focuses. In: IEEE Conference on computer vision and pattern recognition, pp 5021–5030
35. Yi Z, Zhang H, Tan P, Gong M (2017) DualGAN: unsupervised dual learning for image-to-image translation. In: IEEE International conference on computer vision, pp 2849–2857
36. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International conference on computer vision, pp 2223–2232

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.