Face image-sketch synthesis via generative adversarial fusion

Jianyuan Sun [a,b], Hongchuan Yu [c,*], Jian J. Zhang [c], Junyu Dong [d], Hui Yu [e], Guoqiang Zhong [d,*]

[a] Department of Computer Science and Technology, Qingdao University, Qingdao 266071, China
[b] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK
[c] National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, UK
[d] Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
[e] School of Creative Technologies, University of Portsmouth, Portsmouth PO1 2DJ, UK

Face image-sketch synthesis is widely applied in law enforcement and digital entertainment fields. Despite the extensive progression in face image-sketch synthesis, there are few methods focusing on generating a color face image from a sketch. The existing methods pay less attention to learning the illumination or highlight distribution on the face region. However, the illumination is the key factor that makes the generated color face image looks vivid and realistic. Moreover, existing methods tend to employ some image preprocessing technologies and facial region patching approaches to generate high-quality face images, which results in the high complexity and memory consumption in practice. In this paper, we propose a novel end-to-end generative adversarial fusion model, called GAF, which fuses two U-Net generators and a discriminator by jointly learning the content and adversarial loss functions. In particular, we propose a parametric tanh activation function to learn and control illumination highlight distribution over faces, which is integrated between the two U-Net generators by an illumination distribution layer. Additionally, we fuse the attention mechanism into the second U-Net generator of GAF to keep the identity consistency and refine the generated facial details. The qualitative and quantitative experiments on the public benchmark datasets show that the proposed GAF has better performance than existing image-sketch synthesis methods in synthesized face image quality (FSIM) and face recognition accuracy (NLDA). Meanwhile, the good generalization ability of GAF has also been verified. To further demonstrate the reliability and authenticity of face images generated using GAF, we use the generated face image to attack the well-known face recognition system. The result shows that the face images generated by GAF can maintain identity consistency and well maintain everyone's unique facial characteristics, which can be further used in the benchmark of facial spoofing. Moreover, the experiments are implemented to verify the effectiveness and rationality of the proposed parametric tanh activation function and attention mechanism in GAF.

1. Introduction

Face image-sketch synthesis aims to generate a face sketch from an input face image, or vice versa, which are widely applied in many areas such as digital entertainment and law enforcement. Specifically, we are concerned about the sub-problem of generating face image from a sketch, since it is one of the important technologies of legal safety supervision. For example, due to the unavailability of suspect face images, face sketches are drawn manually using professional software in terms of the witness description, and are further used for police hunting (Klare, Li, & Jain, 2011; Wang & Tang, 2009; Wang, Tao, Gao, Li, & Li, 2014). Another increasing public concern is the face spoofing attacks to face recognition systems (Peng, Wang, Li, & Gao, 2021; Saez Trigueros, Meng, & Hartnett, 2021). When verifying the security of a face recognition system, it tends to produce inaccurate verification results using the sketches to attack face recognition system. Because it is difficult to match two different image modalities, i.e. face images and sketches (Galea & Farrugia, 2018; Peng et al., 2021). One feasible solution is to generate a face image from the query sketch and then attack the face recognition systems. Challenging issues are raising, i.e. (1) if and what extent face features can be regenerated from sketches; (2) what extent face features extracted from face sketches can be inverted to obtain the real face image.

In recent years, many researchers have devoted efforts to study the face image-sketch synthesis. Although the extensive progressives in face image-sketch synthesis (Peng, Wang, Li, & Gao, 2020) have been achieved, there are few methods focusing on generating a face image from a sketch. The classic method is to divide a face into a set of overlapped patches with different scales, and train a machine learning model to generate the face image from a query sketch (Wang & Tang, 2009; Xiao, Gao, Tao, & Li, 2009). Xiao et al. (2009) use the embedded hidden Markov model (EHMM) (Nefian & Hayes, 1999) to learn the nonlinearity of image-sketch pair with less training samples. Recently, generative adversarial networks (GANs) and their variants have been employed in learning the nonlinear relationship of image-sketch pairs, and yielded promising results (Peng et al., 2016; Sangkloy, Lu, Fang, Yu, & Hays, 2017; Sannidhan, Prabhu, Robbins, & Shasky, 2019; Wang, Sindagi, & Patel, 2018; Yi, Liu, Lai, & Rosin, 2019; Zhu, Li, Wang, & Gao, 2019). These methods usually regard the task of face image-sketch synthesis as an image-to-image translation problem. The raising issue is that the generated face images cannot reach high resolution using the GANs model alone (Sangkloy et al., 2017). Because of the effectiveness of the StackGANs in improving the resolution of the generated images (Zhang, Xu, & Li, 2017), it is becoming more and more popular in face image-sketch synthesis task (Sannidhan et al., 2019; Wang et al., 2018; Zhu et al., 2019). In particular, Zhu et al. (2019) propose a deep collaborative framework with two opposite networks to synthesis face images from sketches, whose structure is similar to the existing CycleGAN (Zhu, Park, Isola, & Efros, 2017). Furthermore, Duan, Chen, Wu, Cai, and Lu (2021) propose a multi-scale gradient self-attention residual learning framework, which strengthens the constraint of facial features using the self-attention mechanism. However, these methods usually have a high complexity due to a large number of parameters involved in deep networks. For example, the methods (Peng et al., 2016; Yi et al., 2019) are used to learn the features of individual facial regions (e.g. eyebrows, eyes, nose and mouth) by a set of sub-networks. The method (Yu et al., 2020) tends to use the image preprocessing method to obtain the different facial pixel-wise labeling masks as the input of the neural network for learning. This process can hardly be end-to-end due to the traditional image preprocessing methods that are difficult to combine with neural networks for end-to-end learning. The other methods are used to increase the depth of networks, i.e. at least 60 convolutional layers in a network model (Zhu et al., 2019). Moreover, all of existing methods ignore learning and controlling the illumination distribution over the faces. It often happens that the face area of the generated face images has no illumination highlight (Duan et al., 2021; Wang et al., 2018; Xiao et al., 2009) or has stronger illumination highlight (Nefian & Hayes, 1999; Sannidhan et al., 2019; Wang & Tang, 2009; Zhu et al., 2019) compared with the ground truth face image. It is not difficult to find that a generated face image looks like a fake image or even a color sketch due to the lack of learning method to control the illumination distribution over the face in the existing methods of generating the color facial image from an input sketch. In fact, the illumination highlight is the key factor that makes the generated color face image looks vivid and realistic.

To tackle the high complexity and illumination highlight issues, we propose a novel end-to-end robust adversarial fusion network that fuses two U-Net generators and a discriminator by jointly learning the content and adversarial loss functions. Specifically, to learn and control illumination highlight distribution for the generated face image, we proposed a parametric tanh activate function, which is integrated between the two U-Net generators by designing an illumination distribution layer. Moreover, we fuse the attention mechanism into the second U-Net generator to refine the features of the face regions instead of the existing methods (Peng et al., 2016; Yi et al., 2019; Yu et al., 2020) that employ a set of sub-network to learn the features of individual facial regions (e.g. eyebrows, eyes, nose and mouth). This effectively reduces the computational complexity and is easier to be applied.

The main contributions of this paper include:

- We propose a novel end-to-end adversarial fusion network model, called GAF, that fuses two U-Net generators and a discriminator by jointly learning the content and adversarial loss functions for the task of generating the color facial image from an input facial sketch. In particular, the attention mechanism is fused into the second U-Net generator. Unlike the traditional methods employing a set of sub-network to learn the features of individual facial regions (e.g. eyebrows, eyes, nose and mouth), we fuse the attention mechanism into the second U-Net generator to refine the features of the face region details and keep the identity consistency, which can effectively reduce the computational complexity and is easier to be applied.
- To learn and control illumination highlight distribution over the faces, we propose an illumination activation function in the field of generating the facial image from the facial sketch, i.e., a new extension of tanh termed parametric tanh (ptanh). Unlike the slope of the standard tanh activation function is not adjustable, ptanh can be trained using back-propagation with other layers jointly. It makes that learning and controlling the illumination highlight distribution is an ease problem in the convolution network. In GAF, the ptanh is integrated between the two generators by an illumination distribution layer. In the illumination distribution layer, the ptanh is used to learn the highlight for the face region of the generated face image.
- The qualitative and quantitative experiments on the public benchmark datasets prove that the proposed GAF method is superior to the existing image-sketch synthesis deep learning methods in synthesis face image quality (FSIM) and face recognition accuracy (NLDA). Meanwhile, the good generalization ability of GAF has also been verified on the public sketch-photo synthesis datasets. In particular, to verify the reliability and authenticity of face images generated using GAF, we use the generated face image to attack the well-known face recognition system, FaceNet (Schroff, Kalenichenko, & Philbin, 2015), to quantitatively evaluate whether the generated face image and the ground truth face image are the same people. The result shows that the face images generated by GAF can keep the identity consistent and well maintain everyone's unique facial characteristics. Moreover, the experiments are implemented to verify the effectiveness and rationality of the proposed parametric tanh activation function and attention mechanism in GAF.

2. Related work

The task of face image-sketch synthesis has been receiving more and more attention. However, it still remains challenging due to great geometrical deformations and large texture difference between face images and sketches. In this section, specifically, previous researches on the task of generating face images from sketch inputs are reviewed.

To tackle the task of generating a color face image from a sketch, some methods based on the traditional machine learning algorithms are proposed (Wang & Tang, 2009; Xiao et al., 2009). For example, Xiao et al. (2009) propose a synthesis face image algorithm that embedded the hidden Markov model (EHMM) to learn the complex nonlinear relationship between photos and sketches. In recent years, deep learning models have been widely used in many fields and achieved state-of-the-art results, specially the generative adversarial networks (GANs) model. GANs showed a significant success in the task of the image-to-image style translation (Isola, Zhu, Zhou, & Efros, 2017; Zhu et al., 2017). Therefore, researchers try to use GANs and their variants to solve the face image-sketch synthesis task. For example, Sangkloy et al. first use an adversarial deep architecture network to generate realistic images from the sketch inputs, where the sketches are with sparse color scribbles (Sangkloy et al., 2017). Moreover, Di et al. combine the variational auto-encoder and conditional GANs (cGANs) to preserve the attributes of the generated face image and improve the quality of the overall image (Di & Patel, 2018). However, there was also some deformations in the facial parts, such as the serious deformations and aliasing defects over the mouth and hair regions. The reason for deformation is that the training process of the original GANs and cGANs model is unstable. To improve the resolution of the generated image, the StackGANs model was proposed (Zhang et al., 2017). Furthermore, Zhu et al. (2017) propose the CycleGAN by adding an inverse mapping and a cycle consistency loss function to tackle unpaired images. Subsequently, the StackGANs and CycleGAN are also widely applied in the face photo-sketch synthesis task. For example, Wang et al. (2018) propose a novel synthesis framework with the Multi-Adversarial Networks (PS2-MAN) based on CycleGAN, which iteratively generates low resolution to high resolution images in an adversarial way. Sannidhan et al. (2019) combine the trained Convolution Neural Network and a conditional Generative Adversarial Network (cGANs) to generate the synthesis sketch and photo image at the same time. In particular, Yi et al. (2019) propose a hierarchical GANs model that the generator including the global network and six local networks for the whole images and the individual facial regions, respectively. Since face photos can be described using features from different face regions, Peng et al. (2016) present a novel multiple representations-based method that combines multiple representations to represent an image patch. In particular, the multiple filters were used to the multiple features of face images, and the Markov networks were used to exploit the interacting relationships between the sketch and face image. However, the generated images of the work (Peng et al., 2016) often have serious illusions. Moreover, the face patch methods usually need some extra calculation process due to each face patch needs to be processed separately. To reduce the extra calculation process of the patch methods, Zhu et al. (2019) propose a deep collaborative framework with two opposite networks to synthesis face image-sketch and designed a collaborative loss for the two opposite mappings. This deep collaborative framework is similar to the existing CycleGAN (Zhu et al., 2017). In addition to the limited generalization ability, these face image-sketch methods usually have a higher complexity due to a large number of parameters involved in a complex or deep network architecture. Moreover, these existing methods ignore learning and controlling the illumination distribution over the face. It is easy to observe that the face area of generated face images has no illumination highlight (Duan et al., 2021; Wang et al., 2018; Xiao et al., 2009) or has stronger illumination highlight (Sannidhan et al., 2019; Wang & Tang, 2009; Zhu et al., 2019) compared with the ground truth image. The illumination highlight is the key factor that directly affects the visual quality of the generated face image.

In this paper, to reduce the complexity and generate high-quality face images from the sketch inputs, we propose a novel end-to-end robust generative adversarial fusion network called GAF based on the cGANs (Isola et al., 2017), which fuses two U-Net generators and a discriminator by jointly learning the content and adversarial loss functions where the U-Net generators fuse a parametric tanh activation function and the attention mechanism. Here, we propose a parametric activation function to learn and control illumination highlight distribution over faces, which is integrated between the two U-Net generators by an illumination distribution layer. Moreover, we fuse the attention mechanism into the second U-Net generator of GAF to keep the identity consistent and refine the generated facial details. The attention mechanism is very popular at present (Oktay et al., 2018), which is widely used in the fields of classification (Chen & Shi, 2021; Wang et al., 2017; Zhu, Li, Yang, & Ye, 2020), machine translation (Vaswani et al., 2017), image captioning (Anderson et al., 2018) etc. Duan et al. (2021) prove the effectiveness of the self-attention mechanism on strengthening the constraint of facial features and more robust to the interference of background and other factors. Existing research has confirmed that the attention mechanism fusion network model can help the network model suppress the features unrelated to learning tasks and enhance the features related to learning tasks at the same time.

3. Proposed method

In this section, the proposed GAF model is described in detail. Moreover, the illumination distribution layer that contains the parametric tanh activation function (ptanh) is introduced. The principle of ptanh to learn and control illumination highlight distribution over the faces is explained. Finally, the effectiveness of the attention mechanism fuses into the second U-Net generator of GAF is presented.

3.1. Overview of the proposed GAF

Fig. 1 shows the pipeline of the proposed end-to-end GAF model. The GAF fuses of two U-Net generators U and A, and a convolutional discriminator D by joining learning the content loss and adversarial loss function. In particular, we design an illumination distribution layer between generator U and generator A to learn and control illumination highlight distribution over the face. In the illumination distribution layer P, a novel activate function is used, i.e., the parametric tanh activate function (ptanh function), to learn and control illumination. We have discovered that the illumination distribution layer can do a better job when it works on image intensities instead of image features. So we force the first U-Net generator U to generate an initial face image Y_U instead of latent feature maps, which is then used as the input of the illumination distribution layer P. Moreover, we fuse the attention mechanism into the second U-Net generator A to keep the identity consistent and refine the generated facial details.

To generate face images from sketches, the training face sketch-image pairs are equipped at the GAF input sketch X and the ground truth photo Y respectively. The U-Net generator U extracts the global features of the input sketch X to generate the initial face image Y_U. Then, the face Y_U goes into the illumination distribution layer (ptanh function) to output the face image Y'_U with the highlight over the face and good contrast. After that, the Attention U-Net further refine the generated face image Y'_U. To this end, GAF loss function contains two terms to generate the initial colored face image Y_U and the final colored face image Y_A. Correspondingly, the adversarial losses L_{adv} for Y_U and Y_A are defined as follows:

$$\begin{aligned} L_{adv}&(U, D) \\ &= \mathbb{E}_{x\sim p_{data}(x)}[\log D(Y) + \log(1 - D(U(x)))] \\ &= \mathbb{E}_{x\sim p_{data}(x)}[\log D(Y) + \log(1 - D(Y_U))], \end{aligned} \tag{1}$$

$$\begin{aligned} L_{adv}&(U, P, A, D) \\ &=\mathbb{E}_{x\sim p_{data}(x)}[\log D(Y) + \log(1 - D(UPA(x)))] \\ &=\mathbb{E}_{x\sim p_{data}(x)}[\log D(Y) + \log(1 - D(Y_A))], \end{aligned} \tag{2}$$

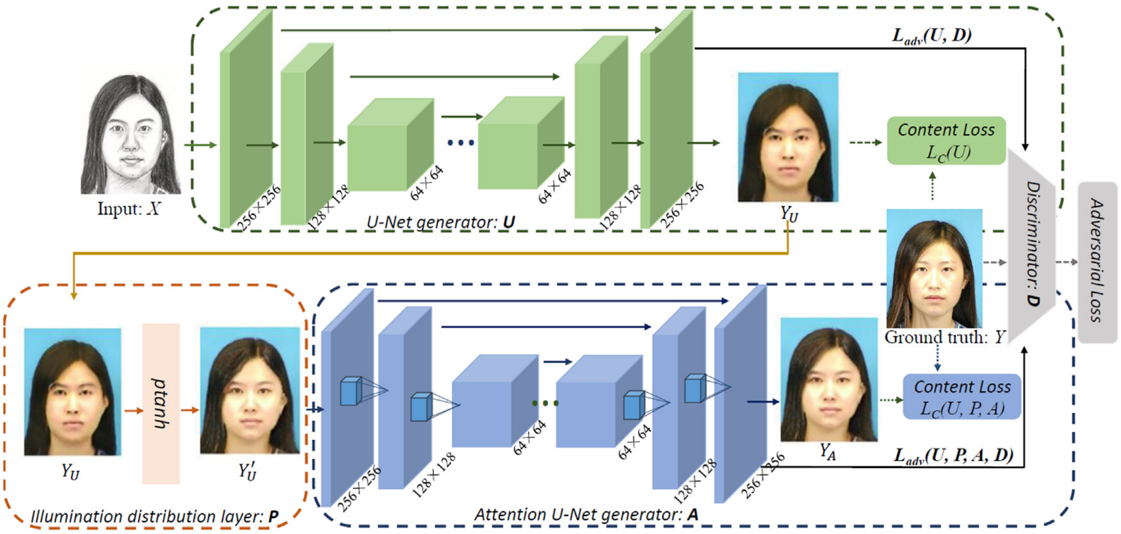

Fig. 1. Framework of GAF. We input the sketch X to generate the initial face image Y_U using the U-Net generator. Then, the initial face image Y_U as the input of the illumination distribution layer to obtain the face image Y_U' with the highlight in the face area and a good contrast. Finally, the Attention U-Net is used to refine the face image Y_U' and generated the final face image Y_A. The overall object function contains the content loss and adversarial loss between the initial face image Y_U, the final generated image Y_A after refine and the ground truth face image Y.

where Y is the target face image, U, P, A and D represent the U-Net generator, the illumination distribution layer, the attention mechanism based generator and the discriminator, respectively.

To make the generated face images Y_U and Y_A close to the ground truth image Y in content features. The content losses of the generated face image L_c for Y_U and Y_A are defined as follows:

$$L_c(U) = \|U(X) - Y\| = \|Y_U - Y\|, \tag{3}$$

$$L_c(U, P, A) = \|UPA(X) - Y\| = \|Y_A - Y\|. \tag{4}$$

Then the overall objective function of GAF can be formulated as follows:

$$\begin{aligned} U^*, P^*, A^* = \arg\min_{U,P,A}\max_{D}(L_{adv}(U, D) + \lambda L_c(U) \\ + L_{adv}(U, P, A, D) + \lambda L_c(U, P, A)), \end{aligned} \tag{5}$$

where λ is a parameter to balance the adversarial loss and the content loss. To optimal the overall objective of the proposed GAF, we separate the overall objective function to optimize the discriminator, the illumination distribution layer and the generators:

$$D^* = \arg\max_{D}(\log D(Y) + \log(1 - D(Y_U))), \tag{6}$$

$$U^* = \arg\min_{U}(\log(1 - D(Y_U)) + \lambda\|Y_U - Y\|), \tag{7}$$

$$D^* = \arg\max_{D}(\log D(Y) + \log(1 - D(Y_A))), \tag{8}$$

$$U^*, P^*, A^* = \arg\min_{U,P,A}(\log(1 - D(Y_A)) + \lambda\|Y_A - Y\|). \tag{9}$$

The optimization process of GAF is shown in Algorithm 1.

Algorithm 1: Optimization process of GAF

```
    **Input**: Initialized the generators $U$, $A$, the
        illumination distribution layer $P$, the
        discriminator $D$, nEpoches $= k \in R$.
    **Output**: Optimized $U$, $A$, $P$, $D$.
    **for** $i = 1$ to $k$ do
        Generate the initial face image $Y_U$ by $U$.
        Optimize discriminator $D$ based on Eq. (6).
        Fix discriminator $D$ and optimal U-Net generator
        $U$ by solving Eq. (7).
        Refine the initial face image $Y_A$ by $U$, $P$, $A$.
        Optimize discriminator $D$ based on Eq. (8).
        Fix discriminator $D$ and optimize $U$, $P$, $A$ based on
        Eq. (9).
    **end for**
```

3.2. Illumination distribution layer

Illumination highlight is the key factor that makes the generated face image look realistic and close to the real face image. Inspired by the work (Zhang, Ji, Hu, Gao, & Lin, 2018), we define a slope of the function mapping input pixels to output pixels. However, the existing of the slope of the standard activation functions like tanh, sigmoid or ReLU, which cannot be adjusted for repeated training and learning in convolutional networks. To design an adjustable re-mapping function of the illumination highlight distribution, we propose a new parametric tanh (ptanh) function. The ptanh function is defined as:

$$f(x) = tanh(mx) = \frac{e^{mx} - e^{-mx}}{e^{mx} + e^{-mx}}, \tag{10}$$

where x represents the input, m is a learnable parameter controlling the slope of the function to learn the illumination highlight distribution over the face. The proposed ptanh can be trained by using back propagation with other convolution layers jointly. The updating of parameter m is derived from the chain rule as follows:

$$\frac{\partial L}{\partial m} = \frac{\partial L}{\partial f(x)}\frac{\partial f(x)}{\partial m}, \tag{11}$$

where L is the objective function, $\frac{\partial L}{\partial f(x)}$ is the gradient propagated from the deeper convolution layers. By derivation, we get the gradient of m as follows:

$$\frac{\partial f(x)}{\partial x} = m[1 - \frac{(e^{mx} - e^{-mx})^2}{(e^{mx} + e^{-mx})^2}], \tag{12}$$

$$\frac{\partial f(x)}{\partial m} = x[1 - \frac{(e^{mx} - e^{-mx})^2}{(e^{mx} + e^{-mx})^2}]. \tag{13}$$

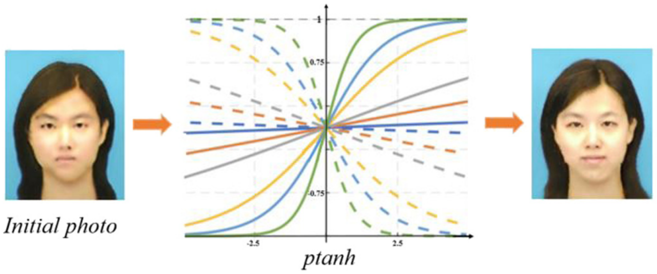

Fig. 2. Illustration of the illumination distribution layer.

The ptanh learns and controls illumination highlight for the initial generated face image, as shown in Fig. 2. From Fig. 2, it can be seen that the illumination distribution layer uses the ptanh activate function to adjust the contrast of the initial face image.

3.3. Attention U-Net in GAF

The effectiveness of the attention mechanism on strengthening the constraint of facial features and more robust to the interference of background and other factors has been proved (Duan et al., 2021). To refine the details of the generator output Y_U, inspired by the existing Attention U-Net (Oktay et al., 2018), we fuse the attention mechanism into the second U-Net generator of GAF. Current face detail refinement methods are used to involve additional sub-networks (or sub-methods) to learn the outline and details of different face regions such as eyebrows, eyes, nose and mouth, separately (Peng et al., 2016; Yi et al., 2019; Yu et al., 2020). Herein, we show that the same goal can be achieved by integrating attention gates (AGs) into the U-Net model. It does not require the training of additional sub-networks. Thus, it does not result in additional computational burden. Moreover, AGs can suppress feature responses in irrelevant background regions (Oktay et al., 2018).

The attention gate model gives different weights to features by the attention coefficients according to specific learning tasks. Let x^l be the activation map of a specified convolutional layer $l \in \{1, \ldots, L\}$. For each layer, AG computes attention coefficients $\alpha^l \in [0, 1]$ to identify salient image regions and prune feature responses to preserve the activations only relevant to the specific task. Attention gating is to scale the input features x^l with the α^l through element-wise multiplication. Moreover, there is a global gating signal g which provides information to AGs to disambiguate task-irrelevant feature content in the x^l. In U-Net architecture, g is collected from the skip connection at the $l-1$ layer. The output of AG $\hat{x}^l$ is concatenated with the x^l. It is observed that applying the feature information extracted from coarse-scale layer to attention gating in skip connections can progressively suppress feature responses outside the face area.

To demonstrate the effect of the AGs in GAF model, we show the response region of attention mechanism to face image features in our network. This is indeed to purposely further enhance the features that the AGs focus on so that the highlighted features are prominent in the output Y_A. For comparison, we also add a test of replacing the Attention U-Net with the usual U-Net in GAF model and show the results in Fig. 3. Fig. 3(c) is the visualization of the attention mechanism response to image features in GAF. The response and suppressing areas of the AGs can be clearly noted in Fig. 3(c). That is, the AGs diminish low-level background blue color features, which highlight the features of the face region and hair. Fig. 3(a) is generated by GAF replacing the Attention U-Net with the usual U-Net. Compared to Fig. 3(b) generated by the GAF model, it can be noted that the facial regions are very blurred. Fig. 3(b) is clearer and closer to the ground truth Fig. 3(d) than Fig. 3(a), which justifies that the AGs enhance the features of facial and hair regions and optimize the generated face images.

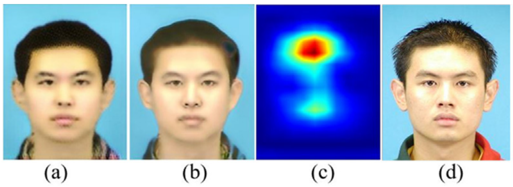

Fig. 3. The effectiveness of Attention U-Net for generating face image. (a) is generated by GAF model replacing the Attention U-Net with U-Net; (b) is obtained by using GAF model; (c) is the visualization of the attention mechanism response to image features in GAF. (d) is the ground truth.

4. Experiments and results

In this section, we carry out a variety of performance tests and generalization experiments on the public face sketch-photo synthesis datasets to demonstrate the outstanding performance of the proposed GAF model. In particular, a face verification evaluation scheme is used to evaluate the quality of the generated face images in Section 4.3. Moreover, the factors that affect the performance of the proposed GAF are analyzed by the experiments.

4.1. Experimental datasets

To verify the effectiveness and generalization of the proposed GAF. Experiments are conducted on the public face image-sketch synthesis datasets[1]: the CUFS dataset (Wang & Tang, 2009) and the CUFSF dataset (Phillips, Moon, Rizvi, & Rauss, 2000). In particular, the CUFS dataset consists of three datasets including the Chinese University of Hong Kong (CUHK) (188 persons), the AR dataset (Martinez & Benavente, 2018) (123 persons), and the XM2VTS dataset (Messer et al., 2000) (295 persons). The CUFSF dataset contains 1194 subjects from the FERET dataset (Phillips et al., 2000). For each dataset, there is a sketch image drawn by the artist and a face image taken in a frontal pose, under normal or variety lighting condition, and with a neutral or exaggerated expression. Some samples are shown in Fig. 4. In particular, most faces in the CUHK dataset are Asian faces, and most faces in the AR, CUFSF and XM2VTS are Western faces.

In our experiment, we train GAF on the CUHK dataset and the AR dataset respectively to obtain two trained GAF models of generating the face image from a sketch input for Asian and Western faces. Especially, in the CUHK database, 88 sketches are selected for training, 100 sketches are selected for testing; in AR dataset, 123 sketches for training, 43 sketches for testing. Moreover, the data augmentation method is utilized to expand the training dataset. Additionally, 1194 sketches of the CUFSF dataset and 295 sketches of the XM2VTS dataset are utilized to test the generalization ability of the proposed GAF.

4.2. The visualization results of generating face images from sketches

To compare with the existing methods, we follow the official training and testing assignment for the CUHK dataset, i.e., 88 sketches are selected for training, 100 sketches are selected for testing. In our experiment, the size of the input sketch of GAF is

[1] http://mmlab.ie.cuhk.edu.hk/facesketch.html

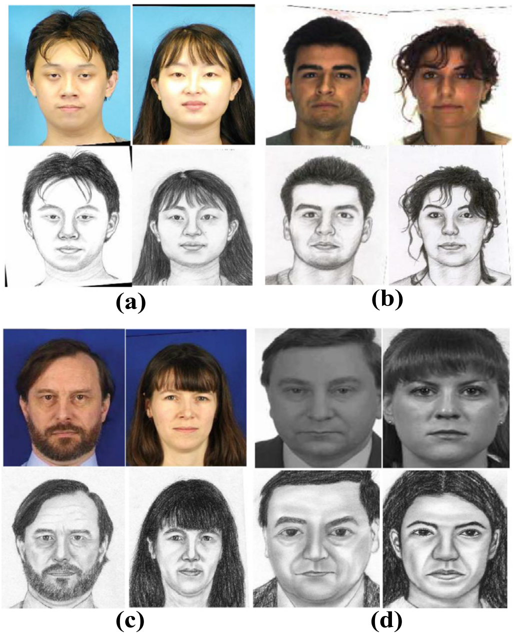

Fig. 4. Examples of face image-sketch pairs from (a) CUHK student dataset, (b) AR dataset, (c) XM2VTS dataset, and (d) CUFSF dataset.

equal to the size of the official cropped sketch, i.e., 200 × 250. For the parameters of GAF model, we use minibatch SGD and Adam solver (Kingma & Ba, 2015) with learning rate being 0.0002, and the momentum parameters $\beta_1 = 0.5$. We empirically set parameter $\lambda = 100$ for the overall function in Eq. (5).

The synthesized face image results are shown in Fig. 5 for the CUHK dataset. Herein, we compare our method with the state-of-the-art face photo-sketch synthesis methods such as MrFSPS (Peng et al., 2016), pix2pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017), DualGAN (Yi, Zhang, Tan, & Gong, 2017) and SCA-GAN (Yu et al., 2020). The MrFSPS method combines multiple features from sketch images by using multiple filters and deploying Markov networks to exploit the interacting relationships between the neighboring sketch patches (Peng et al., 2016). The pix2pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017) are the state-of-the-art methods to solve the problem of style transfer of paired and unpaired images, in which the generators all are based on U-Net. The DualGAN (Yi et al., 2017) is a unsupervised learning method for general-purpose image-to-image translation, which only relies on unlabeled image data. Moreover, the CA-GAN method employs the facial composition information to the face image synthesis, particularly employing a perceptual loss function to enhance the similarity of the synthesized face images and the real face images. In Fig. 5, it can be observed that the synthesized face images of the MrFSPS method are slightly blurred. Furthermore, the results of the pix2pix, DualGAN and CycleGAN can overcome the blurry effect by designing special loss functions. However, they tend to produce undesirable artifacts, specially the CycleGAN resulting in color distortion outputs. The CA-GAN can produce high-resolution synthesis face images. However, a few of the results still have color distortion. In contrast, the proposed GAF model is able to preserve high-frequency details and avoid color distortion simultaneously. In particular, GAF uses the illumination distribution layer to enhance the light distribution over face area to make the synthesized face images vivid. Especially, the proposed illumination distribution function can adjust the contrast of the generated face images to make the light distribution uniform.

Result analysis The traditional MrFSPS method tends to generate the face image with blur in the neck area (Wang et al., 2018). The pix2pix, DualGAN and CycleGAN can overcome the blurry effect due to using the adversarial loss and L-1 loss function. However, they tend to produce undesirable artifacts due to the instability during the training process. Also, applying CycleGAN to photo synthesis results in color distortion due to the lack of L-1 loss when training the network. In contrast, the proposed GAF model applies the L-1 reconstruction errors of the initial synthesized image and the final synthesized image against the ground truth to training. Moreover, the proposed GAF first uses a U-Net generator to extract the global features of the input sketch to generate the initial face image. Then, using a U-Net with attention gating mechanism to suppress unrelated features. This can effectively maintain the identity consistency of the synthesized face image and the corresponding ground truth. Here, we also carry on the experiment to verify the performance of the GAF that only uses one U-Net with attention gating mechanism. The synthesized results on the CUHK and AR dataset are shown in Fig. 6. It is easy to find that the image synthesized by the GAF that only uses one U-Net with attention gating mechanism has severe blurring. This result further illustrates the rationality and necessity of the GAF composed of two generators to obtain good synthesis result. Moreover, the illumination distribution function in the illumination distribution layer not only enhances the light distribution of the face area but also adjusts the contrast of the synthesized images. The detailed analysis is presented in the Section 4.6.

4.3. Identity verification

To verify the reliability and authenticity of face photos generated using GAF, we quantitatively evaluate whether the generated face photo and the ground truth face photo are the same person by using the well-known face recognition system, FaceNet (Schroff et al., 2015). For the FaceNet, it first employ the MTCNN network (Cai, Fan, Feris, & Vasconcelos, 2016) to preprocess the test face photos. Then, using the pre-trained network model to extract the features of each test face photo. Moreover, we can obtain the face similarity by calculating the Euclidean distance between the features of the generated face photo and the ground truth face photo. Here, we use the pre-trained Inception ResNet v1 model (He, Zhang, Ren, & Sun, 2016) based on the VGGFace2 (Parkhi, Vedaldi, & Zisserman, 2015) dataset to extract the features of the generated face photo and the ground truth face photo. Specifically, the identity validation process is shown in Fig. 7.

We measure the face similarity between the generated face image and the ground truth face photo on the AR dataset (Martinez & Benavente, 2018). To generate face images from the sketches, we randomly select 123 sketches as the training dataset to train GAF model, and 43 sketches for testing. For the parameters of GAF model, we also use minibatch SGD and Adam solver (Kingma & Ba, 2015) with learning rate being 0.0002 and the momentum parameters $\beta_1 = 0.5$. The parameter λ is empirically set to 100 for the overall function Eq. (5). Then, we can obtain 43 generated face images from the 43 test sketches. Furthermore, we randomly select 4 generated face images and the corresponding ground truth face images as the input of the FaceNet (Schroff et al., 2015).

The output Euclidean distance matrix between any two face images is shown in Table 1 where the GAF-image0, GAF-image1, GAF-image2, GAF-image3 represent the face images of 4 individuals generated by the GAF model; GT-image0, GT-image1, GT-image2 and GT-image3 correspond to the individual ground truth face photos. We set the similarity threshold is $d = 1$ at

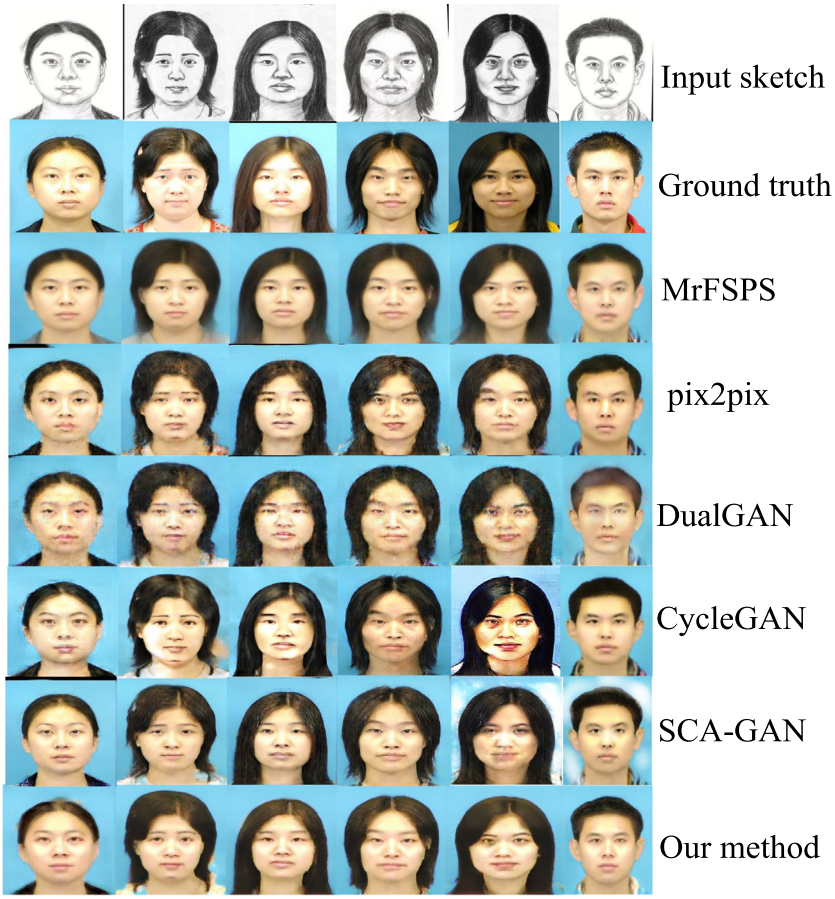

Fig. 5. Comparison of sketch to photo synthesis results on the CUHK student dataset. From top to bottom: Input sketch, Ground truth, MrFSPS, pix2pix, DualGAN, CycleGAN, SCA-GAN and Our method.

Table 1
Inter-class distance matrix between any two face images where GAF-image0, GAF-image1, GAF-image2, GAF-image3 represent the generated face image of different individuals by GAF model. GT-image0, GT-image1, GT-image2 and GT-image2 represent the corresponding ground truth face image, respectively. The similarity threshold is 1 at $FAR = 0.1\%$. That is, the FaceNet system determines that the two face images are not the same person when the Euclidean distance between the two face images is greater than 1. Otherwise, the two face images are the same person when the Euclidean distance between the two face images is less than 1. The Euclidean distance of the same person face images is 0.

Euclidean distance	GT-image0	GAF-image0	GT-image1	GAF-image1	GT-image2	GAF-image2	GT-image3	GAF-image3
GT-image0	**0.0000**	**0.8995**	1.2375	1.2403	1.3744	1.3075	1.3344	1.3781
GAF-image0	**0.8995**	**0.0000**	1.2536	1.0368	1.2648	1.1146	1.4329	1.3074
GT-image1	1.2375	1.2536	**0.0000**	**0.9091**	1.3163	1.3330	1.3517	1.4093
GAF-image1	1.2403	1.0368	**0.9091**	**0.0000**	1.2606	1.1814	1.3040	1.3078
GT-image2	1.3744	1.2648	1.3163	1.2606	**0.0000**	**0.7942**	1.1918	1.06723
GAF-image2	1.3075	1.1146	1.3330	1.1814	**0.7942**	**0.0000**	1.2738	1.0204
GT-image3	1.3344	1.4329	1.3517	1.3040	1.1918	1.2738	**0.0000**	**0.9809**
GAF-image3	1.3781	1.3074	1.4093	1.3078	1.0672	1.0204	**0.9809**	**0.0000**

$FAR = 0.1\%$ in the face verification system, that is, the face evaluation scheme determines that the two face images are not the same person when the Euclidean distance between the two face images is greater than 1. Otherwise, the two face images are the same person. It is desired that the Euclidean distance of the same person's face images is 0. From the Table 1, it can be noted that the face images generated by GAF are consistent in keeping the identity such as the distances tend to be 0. In particular, facial images from different identity sketches well maintain everyone's unique characteristics.

Table 1 focuses on measuring the similarity of face images between different people, i.e., inter-class distance. To measure the similarity of face images from the same people, we still use this face verification system to measure the intra-class distance of face images. Fig. 8 shows that a person usually has multiple photos with different expressions or different lighting in practice, i.e. face photos GT-00 to GT-05. The GAF-image0 to GAF-image6 represent the face images generated by GAF from the facial sketches of six persons, i.e. image0 to image6. Herein, we also set the similarity threshold is $d = 1$ at FAR $= 0.1\%$ of the face verification system. Table 2 shows the intra-class distance matrix for the face similarity. From the Table 2, it can be noted that the Euclidean distance between any two images is less than 1. Although there are a few of distances to close to 1, it is reasonable since there is a big difference between neutral and large expressions. Therefore, we can conclude that the face images generated by GAF can maintain the identity consistency. In particular, the proposed face

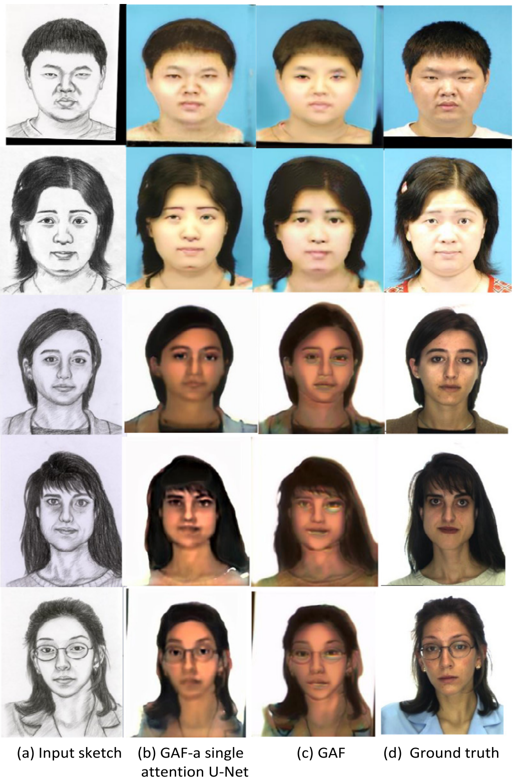

Fig. 6. Comparison of sketch to photo synthesis results on the CUHK and AR dataset. From left to right: Input sketch, GAF that only uses one U-Net with attention mechanism, the proposed GAF, ground truth.

Table 2
Intra-class distance matrix between any two face images. The two face images are the same person when the Euclidean distance between the two face images is less than 1. (GT-*ij* denotes the *j*th ground truth face photo of the *i*th person).

Euclidean distance	GT-00	GT-01	GT-02	GT-03	GT-04	GT-05
GAF-image0	0.9601	0.4207	0.4324	0.3304	0.6136	0.6038
	GT-10	GT-11	GT-12	GT-13	GT-14	GT-15
GAF-image1	0.9760	0.5900	0.9984	0.9717	0.7803	0.8697
	GT-20	GT-21	GT-22	GT-23	GT-24	GT-25
GAF-image2	0.8964	0.6643	0.5510	0.4599	0.7030	0.6122
	GT-30	GT-31	GT-32	GT-33	GT-34	GT-35
GAF-image3	0.9012	0.3984	0.5310	0.5411	0.7056	0.5134
	GT-40	GT-41	GT-42	GT-43	GT-44	GT-45
GAF-image4	0.8803	0.7615	0.4599	0.5791	0.4945	0.6582
	GT-50	GT-51	GT-52	GT-53	GT-54	GT-55
GAF-image5	0.8995	0.8203	0.7030	0.7056	0.4055	0.8043
	GT-60	GT-61	GT-62	GT-63	GT-64	GT-65
GAF-image6	0.9886	0.5907	0.6122	0.5134	0.6582	0.8146

evaluation scheme can accurately assess the identity consistency.

4.4. The quantitative analysis

In this subsection, we present the quantitative results by comparison the state-of-the-art face image-sketch synthesis methods such as MrFSPS (Peng et al., 2016), pix2pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017), DualGAN (Yi et al., 2017) and CA-GAN (Yu et al., 2020) on CUHK dataset.

The synthesized face image quality assessment: We adopt the Feature Similarity Index Metric (FSIM) (Zhang, Zhang, Mou, & Zhang, 2011) to objectively assess the quality of the synthesized face images. The FSIM obtains the quality index by measuring the feature similarity between a synthesized image and the ground truth, in which the features include the phase consistency (PC) and the image gradient magnitude (GM). The comparison results are shown in Table 3. It can be noted that the MrFSPS method (non-deep learning method) is better than the others (deep learning methods). However, their FSIMs are still comparable.

The face recognition assessment: We use the existing face recognition method, Null-space Linear Discriminant Analysis (NLDA) (Chen, Liao, Ko, Lin, & Yu, 2000), to statistically evaluate the face recognition accuracy based on the synthesized face images. Herein, we use the CUHK dataset, and employ the ground-truth photos as the training data and the synthesized images as the test data in NLDA. Performing on the face image data generated by our GAF, the NLDA algorithm reaches the highest accuracy. Moreover, Fig. 9 also shows the recognition accuracy against variations of the number of the principal feature vectors employed by NLDA. It can be noted that around first 40 principal feature vectors, the NLDA algorithm achieves the highest recognition accuracy.

4.5. Generalization verification

Generalization is used to describe a model's ability to react to new data. That is, a model can digest new data and make accurate predictions after being trained on a training set. A model's generalization ability is the key to the success of a model (Bishop, 2006). Therefore, we employ the existing public dataset (XM2VTS and CUFSF) based on the trained GAF model to verify the generalization. Moreover, we conducted the same experiment on the trained pix2pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017).

For the XM2VTS dataset, all face sketches are used, i.e., 295 sketches are selected for testing. For the CUFSF dataset, all face sketches are used to test, i.e., 1194 sketches are selected for testing. Fig. 10 shows the face generation results for the XM2VTS datasets on pix2pix, CycleGAN, and the proposed GAF model. From Fig. 10, it is not difficult to find that the proposed GAF model can accurately and clearly capture the facial expression, eye direction and facial beard. Moreover, GAF obtains a more clear face image than the other two models. In addition, Fig. 11 shows the face generation results for the CUFSF datasets. These results are also obtained from the trained pix2pix, CycleGAN, and the proposed GAF model. From Fig. 11, it is obvious to find that GAF has a better generalization ability than the other models for the face generation task. We can find that the generated photos using GAF have artifacts. The reason for the artifacts is the size of the original sketch image is too small, i.e., the size of the original sketch image is 64 × 80. However, the trained GAF model is formed by training GAF model on AR dataset. The input size of GAF model is 200 × 250. Therefore, when we input the size of the CUFSF's sketch is 64 × 80 to the trained GAF, the generated photos have artifacts. Even so, the trained GAF can well maintain identity consistency. From all the results, the proposed GAF model has good generalization ability.

4.6. Effectiveness of illumination distribution layer

To validate the illumination distribution layer, we remove it from the GAF model, that is, the two GAF models with and without the illumination layer are trained on the CUHK dataset respectively. The results are shown in Fig. 12. It can be noted that there are distinct highlights on the face areas in the second column due to the illumination distribution. Moreover, it can also be noted that the contrast of the generated face images is adjustable and the illumination is evenly distributed over the face area. The illumination distribution layer makes the resulting face images more realistic and vivid.

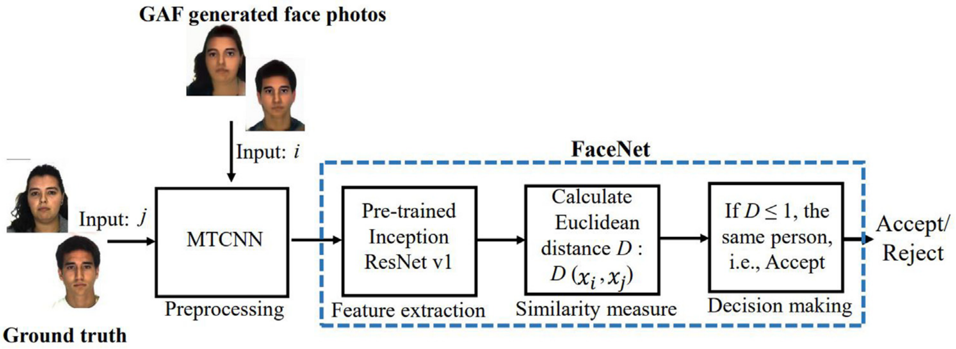
GAF generated face photos
Input: i
FaceNet
Input: j
MTCNN
Preprocessing
Pre-trained Inception ResNet v1
Feature extraction
Calculate Euclidean distance D : $D\,(x_i, x_j)$
Similarity measure
If $D \leq 1$, the same person, i.e., Accept
Decision making
Accept/ Reject
Ground truth

Fig. 7. The identity verification scheme based on the FaceNet.

Table 3
The assess results of the synthesized face image quality (FSIM) and the face recognition (NLDA) by comparison of photo-sketch synthesis methods based on CUHK dataset.

Criterion	MrFSPS (Peng et al., 2016)	pix2pix (Isola et al., 2017)	CycleGAN (Zhu et al., 2017)	DualGAN (Yi et al., 2017)	SCA-GAN (Yu et al., 2020)	Ours
FSIM↑	**0.8031**	0.6997	0.7826	0.7939	0.7950	0.7963
NLDA↑	96.7	93.8	96.07	97.4	98.5	**99.02**

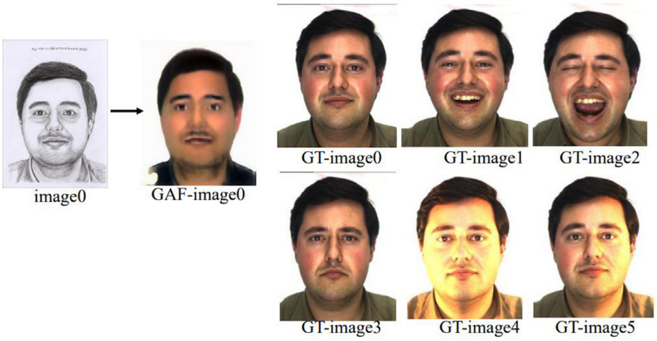

Fig. 8. An example of multiple face photos of the same person. The GAF-image0 represents the face image from an input sketch image0 generated by the GAF model. GT-00 to GT-05 represent the multiple ground truth face photos with different expressions and lighting.

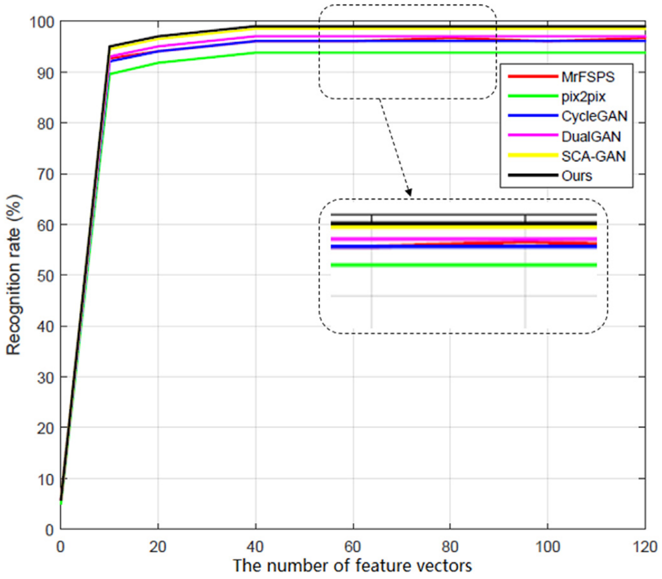

Fig. 9. The recognition accuracy against variations of the number of principal feature vectors in NLDA on the CUHK dataset.

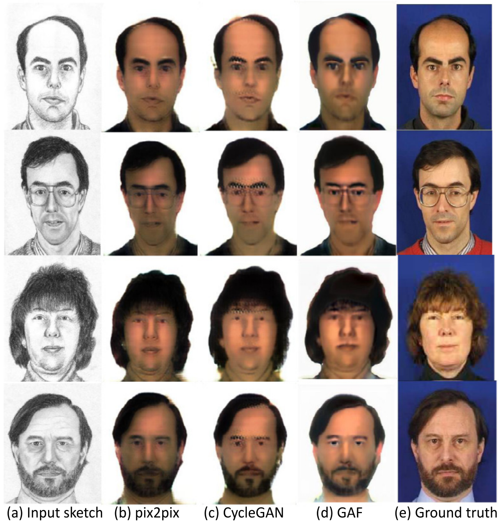

Fig. 10. The face images generated on the XM2VTS dataset using the trained pix2pix, CycleGAN, and the proposed GAF model. From left to right: input sketch, the face image generated using the trained pix2pix, the face image generated using the trained CycleGAN, the face image generated using the trained GAF, ground truth.

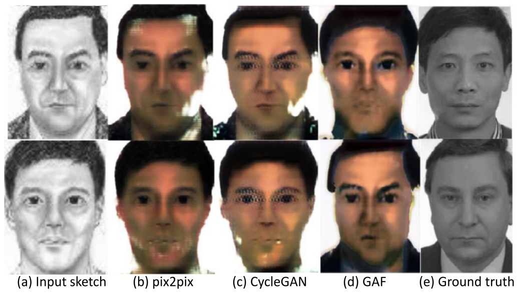

Fig. 11. The face images are generated on the CUFSF dataset using the trained pix2pix, CycleGAN, and the proposed GAF model. From left to right: input sketch, the face image generated using the trained pix2pix, the face image generated using the trained CycleGAN, the face image generated using the trained GAF, ground truth.

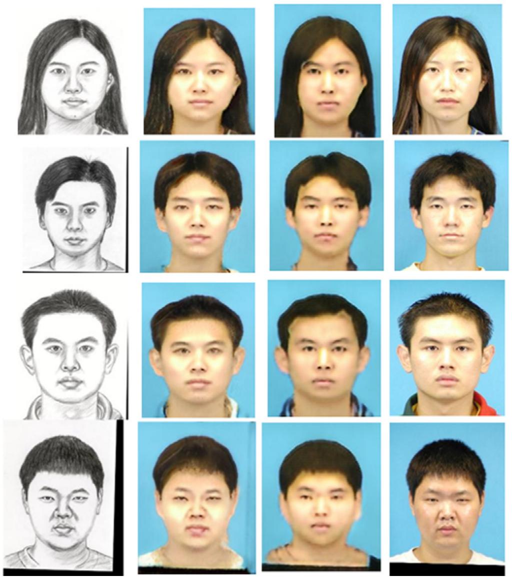

Fig. 12. Comparison of GAF model trained without/with the illumination distribution layer on the CUHK dataset. From left to right: input sketch, the generated photo with the illumination distribution layer, the generated photo without the illumination distribution layer and the ground truth.

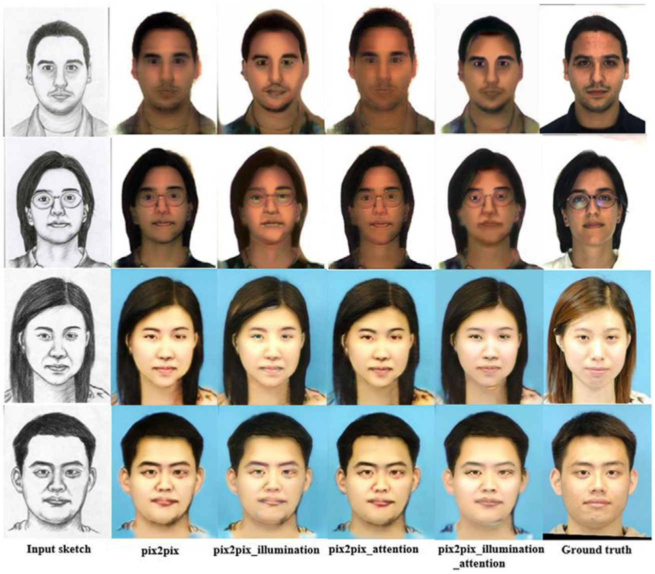

Fig. 13. The performance validation experiment for the attention mechanism and illumination distribution layer on the pix2pix model. From left to right: input sketch, the face images are generated using the pix2pix model, the face images are generated using the pix2pix model plus the illumination distribution layer, the face images are generated using the pix2pix model plus attention mechanism, the face images are generated using the pix2pix model plus the illumination distribution layer and attention mechanism and the ground truth.

To further verify the effectiveness of the illumination distribution layer and attention mechanism, we apply them to the existing pix2pix models (Isola et al., 2017). The results are shown in Fig. 13. It is easy to observe that the face images generated by pix2pix using illumination distribution layer and the pix2pix using attention mechanism are better than that of the pix2pixe only. Moreover, the pix2pix using the illumination distribution layer and attention mechanism together shows the best performance. This test furtherly verifies the effectiveness and practicability of the illumination distribution layer and attention mechanism.

5. Computational complexity

The traditional methods of generating face images from sketches are used to employ sub-networks for learning the different face feature regions, which usually leads to a high cost of computation and storage requirement such as MrFSPS (Peng et al., 2016). Particularly, the MrFSPS method uses the nearest neighbor searching (NNS) that is very time-consuming. The computational complexity of MrFSPS is up to $O(cp^2MN)$ (Zhang et al., 2018), where c denotes the number of all candidates of the search region, p denotes the image patch size, M denotes the number of image patches on each image, and N denotes the number of training data. For the proposed GAF and SCA-GAN (Yu et al., 2020), since there is no NNS part, the computational complexity is only $O(1)$ when the size of the training dataset is fixed.

The SCA-GAN method (Yu et al., 2020) performed training on a single Pascal Titan Xp GPU using a training set of 500 samples. It took six hours to reach the best prediction state. Our GAF performed training on a single GeForce GTX 1080 GPU using the same number of training samples. It took three and a half hours to reach the desired prediction state.

6. Conclusion

In this paper, we study the problem of generating face image from a sketch, propose a novel end-to-end generative adversarial fusion network model, i.e. GAF. GAF fuses two U-Net generators and a discriminator by joining learning the content and adversarial loss function. In particular, a parametric tanh activation function is proposed to learn and control illumination highlight distribution over faces, which is integrated between two U-Net generators by an illumination distribution layer. It is making that learning and controlling the illumination highlight distribution is not a difficult problem in the convolution network. Additionally, we fuse the attention mechanism to the second U-Net generator that can not only maintain the identity consistent, but also refine the generated facial details. Moreover, the experimental results show that the proposed GAF can achieve promising progress in the face image quality results, the recognition accuracy of synthesized face image and the attack face recognition system results.

Declaration of competing interest

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.

Acknowledgments

This work was supported by the EUH2020 project-AniAge (691215), the National Key Research and Development Program of China under Grant No. 2018AAA0100400, the Joint Fund of the Equipments Pre-Research and Ministry of Education of China under Grant No. 6141A020337, the Science and Technology Program of Qingdao under Grant No. 21-1-4-ny-19-nsh, the Natural Science Foundation of Shandong Province, China under Grant No. ZR2020MF131.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE conference on computer vision and pattern recognition*.

Bishop, C. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision, Vol. 9908* (pp. 354–370).

Chen, L., Liao, H. M., Ko, M., Lin, J., & Yu, G. (2000). A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, *33*(10), 1713–1726.

Chen, W., & Shi, K. (2021). Multi-scale attention convolutional neural network for time series classification. *Neural Networks*, *136*, 126–140.

Di, X., & Patel, V. M. (2018). Face synthesis from visual attributes via sketch using conditional VAEs and GANs. *Arxiv:1801.00077*.

Duan, S., Chen, Z., Wu, Q. M. J., Cai, L., & Lu, D. (2021). Multi-scale gradients self-attention residual learning for face photo-sketch transformation. *IEEE Transactions on Information Forensics Security*, *16*, 1218–1230.

Galea, C., & Farrugia, R. A. (2018). Matching software-generated sketches to face photographs with a very deep CNN, morphed faces, and transfer learning. *IEEE Transactions on Information Forensics and Security*, *13*(6), 1421–1431.

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).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *IEEE conference on computer vision and pattern recognition* (pp. 5967–5976).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations ICLR*.

Klare, B., Li, Z., & Jain, A. K. (2011). Matching forensic sketches to mug shot photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(3), 639–646.

Martinez, A., & Benavente, R. (2018). *The AR face database*: *Technical report 24, CVC*.

Messer, K., Matas, J., Kittler, J., Jonsson, K., Luettin, J., & Maitre, G. (2000). Xm2vtsdb: The extended m2vts database. In *Proc. of audio- and video-based person authentication* (pp. 965–966).

Nefian, A., & Hayes, M. (1999). Face recognition using an embedded HMM. In *Proceedings of international conference on audio-and video-based biometric personauthentication* (pp. 19–24).

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M. C. H., Heinrich, M. P., Misawa, K., et al. (2018). Attention U-net: Learning where to look for the pancreas. arXiv:1804.03999.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *British machine vision conference* (pp. 1–12).

Peng, C., Gao, X., Wang, N., Tao, D., Li, X., & Li, J. (2016). Multiple representations-based face sketch-photo synthesis. *IEEE Transaction Neural Networks and Learning Systems*, *27*(11), 2201–2215.

Peng, C., Wang, N., Li, J., & Gao, X. (2020). Face sketch synthesis in the wild via deep patch representation-based probabilistic graphical model. *IEEE Transactions on Information Forensics Security*, *15*, 172–183.

Peng, C., Wang, N., Li, J., & Gao, X. (2021). Soft semantic representation for cross-domain face recognition. *IEEE Transactions on Information Forensics Security*, *16*, 346–360.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis Machine Intelligence*, *22*(10), 1090–1104.

Saez Trigueros, D., Meng, L., & Hartnett, M. (2021). Generating photo-realistic training data to improve face recognition accuracy. *Neural Networks*, *134*, 86–94.

Sangkloy, P., Lu, J., Fang, C., Yu, F., & Hays, J. (2017). Scribbler: Controlling deep image synthesis with sketch and color. In *IEEE conference on computer vision and pattern recognition* (pp. 6836–6845).

Sannidhan, M. S., Prabhu, G. A., Robbins, D. E., & Shasky, C. (2019). Evaluating the performance of face sketch generation using generative adversarial networks. *Pattern Recognition Letter*, *128*, 452–458.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified embedding for face recognition and clustering. In *IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 6000–6010).

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., et al. (2017). Residual attention network for image classification. In *Conference on computer vision and pattern recognition* (pp. 6450–6458).

Wang, L., Sindagi, V., & Patel, V. M. (2018). High-quality facial photo-sketch synthesis using multi-adversarial networks. In *IEEE international conference on automatic face and gesture recognition* (pp. 83–90).

Wang, X., & Tang, X. (2009). Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(11), 1955–1967.

Wang, N., Tao, D., Gao, X., Li, X., & Li, J. (2014). A comprehensive survey to face hallucination. *International Journal of Computer Vision*, *106*(1), 9–30.

Xiao, B., Gao, X., Tao, D., & Li, X. (2009). A new approach for face recognition by sketches in photos. *Signal Process*, *89*(8), 1576–1588.

Yi, R., Liu, Y.-J., Lai, Y.-K., & Rosin, P. L. (2019). APDrawingGAN: GEnerating artistic portrait drawings from face photos with hierarchical GANs. In *IEEE conference on computer vision and pattern recognition* (pp. 10743–10752).

Yi, Z., Zhang, H. R., Tan, P., & Gong, M. (2017). DualGAN: Unsupervised dual learning for image-to-image translation. In *IEEE international conference on computer vision* (pp. 2868–2876).

Yu, J., Xu, X., Gao, F., Shi, S., Wang, M., Tao, D., et al. (2020). Towards realistic face photo-sketch synthesis via composition-aided GANs. *IEEE Transactions on Cybernatics*, *51*(9), 4350–4362.

Zhang, S., Ji, R., Hu, J., Gao, Y., & Lin, C.-W. (2018). Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid. In *International joint conferences on artificial intelligence (IJCAI)* (pp. 1163–1169).

Zhang, H., Xu, T., & Li, H. (2017). StackGAN: TExt to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE international conference on computer vision* (pp. 5908–5916).

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Tranactions on Image Processing*, *20*(8), 2378–2386.

Zhu, M., Li, J., Wang, N., & Gao, X. (2019). A deep collaborative framework for face photo-sketch synthesis. *IEEE Transactions Neural Networks Learning Systems*, *30*(10), 3096–3108.

Zhu, Y., Li, R., Yang, Y., & Ye, N. (2020). Learning cascade attention for fine-grained image classification. *Neural Networks*, *122*, 174–182.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE international conference on computer vision* (pp. 2242–2251).