Neurocomputing 396 (2020) 13–23

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Face sketch-to-photo transformation with multi-scale self-attention GAN

Yingtao Lei[a], Weiwei Du[b], Qinghua Hu[a,*]

[a] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[b] Information and human science, Kyoto Institute of Technology, Kyoto 606-8585, Japan

ARTICLE INFO

Article history:
Received 23 October 2019
Revised 11 January 2020
Accepted 3 February 2020
Available online 12 February 2020

Communicated by Prof. Junwei Han

Keywords:
Image transformation
Sketch-to-photo
Divide and conquer
Multi-scale
Attention mechanism
Generative adversarial network

ABSTRACT

In this study, we investigate the sketch-to-photo problem, which currently poses a significant challenge in the field of computer vision. A large number of GAN-based encoder-decoder methods have been proposed for image transformation, inspired by the pix2pix model; however, these methods do not produce satisfactory results for photo generation, due to the fact that (1) they miss detailed information of input images because of a single-scale convolution operator in the shallow encoder layers, and (2) they fail to learn long-range dependencies in the deep encoder layers. To better handle these challenges, we present an approach that follows a "divide and conquer" strategy. Our method combines the advantages of a multi-scale convolutional neural network and an attention mechanism and applies these two modules to different encoder layers. Additionally, by optimizing a well-designed loss function, the complex correlations between the sketch and the photo can be calculated. Experimental results show that our method is able to generate high-quality photos from sketch images, and qualitative and quantitative analysis demonstrates its effectiveness and superiority over state-of-the-art models. This work paves a path to replace the traditional encoder structure with the "divide and conquer" strategy to handle image transformation tasks.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Forensic sketch artists help law enforcement officers to identify criminal suspects through facial composite sketches. It is essential to match these to photos to ensure preciseness, which the sketches alone sometimes fail to provide. Previous works [1–4] have handled the task of transforming viewed sketches into mugshot images, and [5] focused on forensic sketches. In this study, we aim to transform the viewed sketches into realistic facial photos.

In recent years, a wide variety of image transformation tasks have been solved by training deep convolutional neural networks (CNNs). The networks receive an input image and transform it into a corresponding output image; both detailed information and complex textures of the images are required in this process. Several researchers have recently tackled these individual tasks with specific mechanisms. For example, Cheng et al. [6] proposed a high-quality fully-automatic colorization method, Ledig et al. [7] achieved impressive super-resolution results, and Li et al. [8] obtained comparable texture synthesis quality by combining Markovian patches with a generative adversarial network (GAN). Image transformation is essentially a distribution transformation task.

The GAN proposed by Goodfellow et al. [9] stands out amongst the various generative models; it has become the preferred method of modeling distributions and transforming a distribution into another distribution. Since the release of the pix2pix model [10], a large number of GAN-based encoder–decoder methods [11–13] have been proposed; however, these existing works treat all the encoder layers equally, ignoring the fact that the features extracted from different encoder layers are distinguishable. For the shallow layers, the features contain more detailed information, and for the deep layers, the features reflect global information such as the dependencies across image regions. The majority of prior research does not handle this difference very well.

As one of the main modules of CNNs, the features generally affect their performance. Deeper and wider CNNs extract better features; however, for a deep CNN, the problem of propagating and maintaining the feature information extracted from shallow layers emerges. Dosovitskiy et al. [14] demonstrated that features lose fine details as we progress to deeper layers. The corresponding approach is to increase the width of CNNs. GoogleNet, proposed

* Corresponding author.
E-mail addresses: 2018216018@tju.edu.cn (Y. Lei), duweiwei@kit.ac.jp (W. Du), huqinghua@tju.edu.cn (Q. Hu).

https://doi.org/10.1016/j.neucom.2020.02.024
0925-2312/© 2020 Elsevier B.V. All rights reserved.

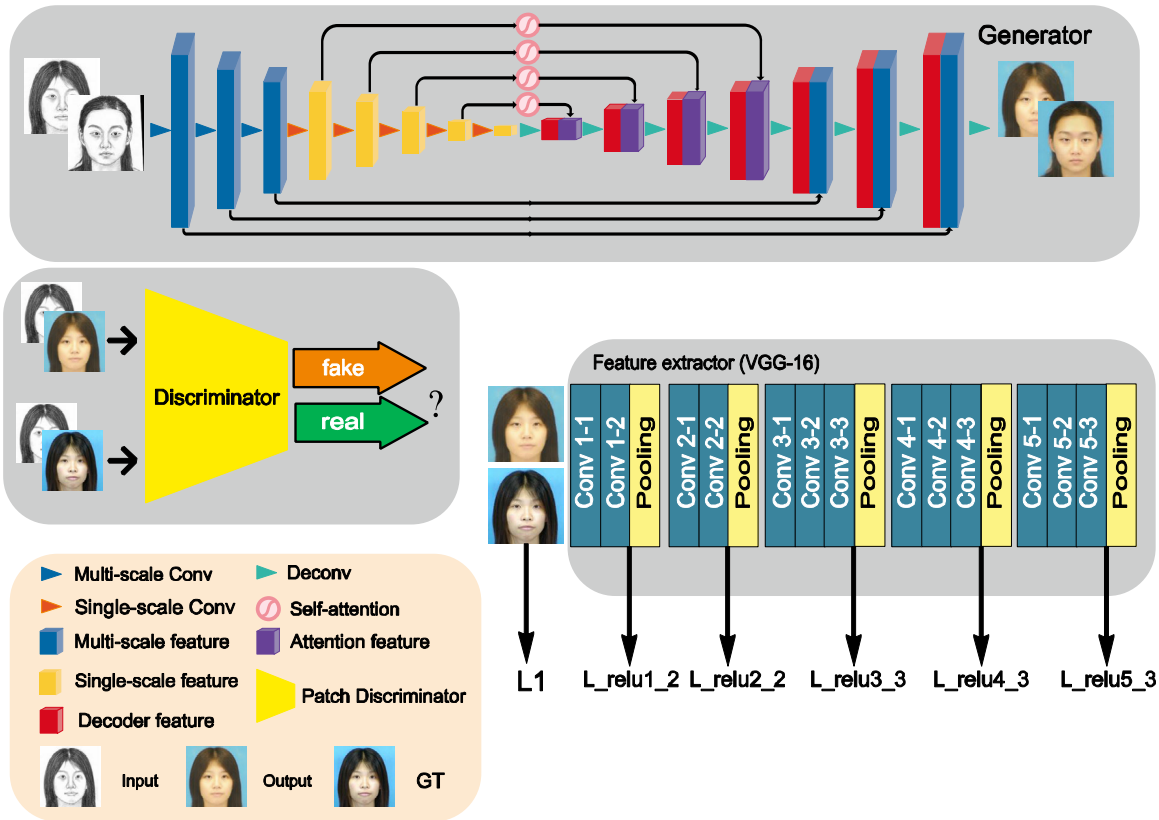


Fig. 1. Framework of our method. The encoder–decoder-based generator adopts the "divide and conquer" strategy. The multi-scale convolution operator contains three types of kernel sizes and extracts multi-scale features, which are then shared directly with the decoder layers. The features extracted from the traditional convolution operator in the deep encoder layers are processed with the self-attention module and then transferred to corresponding decoder layers; the decoder utilizes the features from the corresponding encoder layer and the previous decoder layer to generate images, and the patch discriminator attempts to distinguish between the generated and ground-truth images under the condition of input sketch images. The pre-trained VGG-16 network extracts multi-level features from the outputs and the ground truth images and makes them as similar as possible. Meanwhile, the L_1 loss of the outputs and targets is maintained.

by Szegedy et al. [15], enhanced feature extraction ability by utilizing filtering structures of different sizes at the same layers, which reduces the need for depth in networks. Based on the above findings, a wider convolution operator was chosen for our model, ensuring that we can extract features rich in information in the shallow encoder layers.

For image transformation tasks such as sketch-to-photo, focus should be on the facial region because the quality of this region, rather than the background regions, determines the performance of the reconstructed image. Thus, when synthesizing a location on the face region, there should be a higher dependence on the remaining face region than the background regions. Because of the local receptive field of the convolution operation, the shallow convolutional layers cannot model long-range dependencies across different image regions. The dependencies could only be processed in deep layers after passing through several convolutional layers. It could fail to learn the long-range dependencies for several reasons: the model could be too small to describe the dependencies, or the parameter values calculated may not be able to model the dependencies among multiple layers and multiple channels. Therefore, features extracted from the deep encoder layers cannot focus more on the critical regions. While increasing the convolution kernel size in the deep encoder layers could capture the long-range dependencies in advance, this reduces computational efficiency. Meanwhile, the number of parameters could be too large to be optimized. The self-attention mechanism shows a better balance between learning long-range dependencies and preserving the computational capability; it calculates the response at a position with the weighted sum of the features at all positions, where the weights are calculated in the different feature spaces.

Considering all of the above properties, we design a novel model which adopts different approaches to deal with different encoder layers and produces high quality results by optimizing a well-designed loss function. The contributions of this work are summarized as follows:

- The "divide and conquer" strategy is adopted. For the shallow encoder layers, a multi-scale convolutional network architecture is used to extract more abundant multi-scale details that complement each other. For the deep encoder layers, the features are not shared directly with the shallow decoder layers; rather, they are processed with the self-attention module, and the dependencies are recalculated across image regions. This procedure redistributes the attention of different regions, after which the reconstructed features are shared with the corresponding decoder layers.
- Full use is made of the pre-trained VGG-16 networks [16] as a feature extractor. The multi-level features extracted from the reconstructed images and the target images are matched as closely as possible. The overall framework of our method is shown in Fig. 1.
- Extensive experiments demonstrate the effectiveness of the proposed method and strategy. Our method produces more visually realistic photos and outperforms baselines on multiple datasets, in terms of several evaluation metrics.

2. Related work

2.1. Sketch-to-photo

The process of matching facial sketches to photos has received significant attention in recent years. Liu et al. [1] adopts a method based on the local linear preserving of geometry between the photo and sketch images, to generate pseudo-sketches. Li et al. [4] combines the features generated from face images processed via several filters and explores the interaction relationships between neighboring image patches, using Markov networks. Local feature-based analysis methods were adopted by [3,5] to handle forensic sketches. With the wide application of deep learning, a large number of CNN-based methods, such as [17,18], have been designed and achieve outstanding image transformation results. These methods make full use of the ability of CNNs to extract multi-level features and use these features to reconstruct images. The networks minimize the distance between the target images and the reconstructed images; however, they could not generate visually realistic photos. Our method extracts multi-level features of a sketch image through a multi-layer convolution operation, and then reconstructs a photo with them.

2.2. Generative adversarial network (GAN)

As one of the most productive deep generative models, GAN has been vigorously followed and applied in assorted tasks, especially for image and audio processing tasks, owing to its unusual structured learning ability. Typically, a GAN consists of two networks: the generator and discriminator. The generator tries to produce a more realistic sample to deceive the discriminator while the discriminator tries to distinguish the fake sample from the real sample, which is called an adversarial game. Following the germinal work [9], conditional GAN (cGAN) [19] was developed to generate images on the condition of class labels, after which a series of GAN-family models were proposed to solve diverse problems. These models generate images based on conditions such as text [20], attributes [21], and images [22]. Our model is based on a cGAN on the condition of sketch images.

2.3. U-Net Architectures

The U-net architecture has been used as a basic framework in several image processing problems. Since the pix2pix model was proposed by Isola et al. [10], it has been widely applied in various problems, thereby demonstrating its capacity and capability. Pix2pix picks out the U-Net structure with skip-connections as the architecture of the generator, i.e., an encoder-decoder that shares feature information between the mirrored layers in the encoder and decoder stacks. The framework based on the U-Net-like encoder-decoder has been adopted by [23,24] for special image-to-image transformation tasks. These variants adhere to the traditional pattern where the features are obtained via a single-scale convolution operation in the shallow encoder layers. They can only extract single scale features and details of the input images. For deep encoder layers, the features are directly shared with the corresponding decoder layers. However, these features could not capture long-range dependencies. Our model adopts the "divide and conquer" strategy to address these limitations. We replace the single-scale convolution operation with a multi-scale structure in the shallow encoder layers. For the deep encoder layers, we recalculate the dependencies across different image regions of the features and then share them with the decoder layers.

2.4. Attention models

Recently, the trainable attention mechanism has benefited from advances in deep learning. It mimics the visual attention mechanism to focus on informative regions in image processing. This mechanism falls under two main categories: hard and soft. The soft-attention method produces non-rigid attention maps, which are suitable for natural objects in real images. In particular, self-attention mechanisms calculate the response at a position with the feature at all positions. Vaswani et al. [25] achieved impressive results in machine translation by using the self-attention model; Zhang et al. [26] explored it in the context of GANs and considered long-range relationships; Wang et al. [27] adopted it as a non-local operation to model the spatial and temporal dependencies in video processing, and Wang et al. [28] stacked attention modules in a residual network to generate attention-aware features which were then used for image classification. Our model follows [26] to redistribute the attention of different image regions. However, our method applies the self-attention mechanism to image transformation, rather than just image generation.

3. Method

3.1. Multi-scale network structure

For several image transformation tasks, particularly matching a facial sketch to a photo, a mass of detailed low-level information is needed, which is then shared between the stacks of the U-Net-like encoder-decoder. Ref. [14] shows that some information and details are lost as the features flow through deeper convolutional layers; thus, it is crucial to extract multi-scale features containing fine details from the shallow encoder layers and share them with the deep decoder layers.

Previous works [29–34] have applied the multi-scale filtering structure for image processing and demonstrated its effectiveness. Our multi-scale encoder layer E shown in Fig. 2 extracts multi-scale features from input m as follows:

$$E : m \rightarrow \left\{\phi_{j_i}\right\}, \phi_{j_i} \in R^{H^{(j)} \times W^{(j)} \times C^{(j)}} \tag{1}$$

where ϕ_{j_i} is the ith $(i = 1, 2, 3)$ scale feature of the jth $(j = 1, \ldots, N)$ encoder layer and $H^{(j)}$, $W^{(j)}$, $C^{(j)}$ represent the size of the feature maps from the jth layer. N is equal to three because our model applies the multi-scale convolution operation to the first three encoder layers.

3.2. Attention mechanism

Most U-Net networks with the skip-connections encoder–decoder models for image transformation directly share the features extracted from the encoder layers with the corresponding decoder layers. Because of the local neighborhood operation of convolution in processing information, only the deep convolutional layers could model the long-range dependencies across image regions. However, this is likely to fail on account of the previously mentioned reasons, and such failed dependencies lead to unsuitable attention distribution. Consequently, the generator would not be able to focus on the informative regions when generating images. Therefore, we adopt the self-attention module (see Fig. 3) to recalculate the dependencies across image regions and then redistribute the attention.

The feature maps $m \in \mathbb{R}^{C \times N}$ from previous layers are transformed into two feature spaces f and g to calculate the dependencies across different image regions, where $f(m) = W_f m$, $g(m) = W_g m$. C is the number of channels, and N is the number of

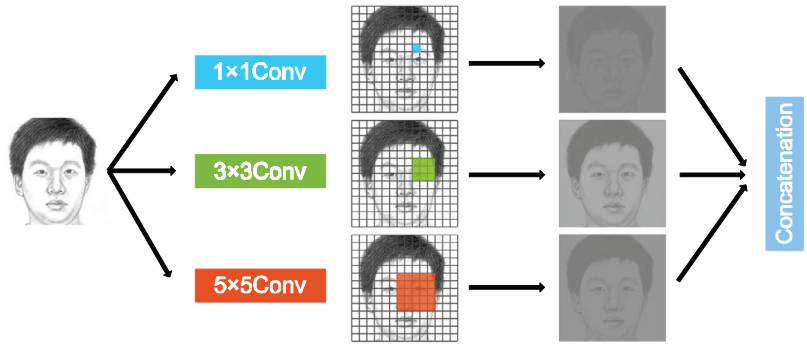


Fig. 2. Structure of the multi-scale convolution operator for extracting multi-scale features that are rich in information and fine details. The blue, green, and orange boxes denote different filter sizes. The concatenation operation is used to concatenate these three scale features.

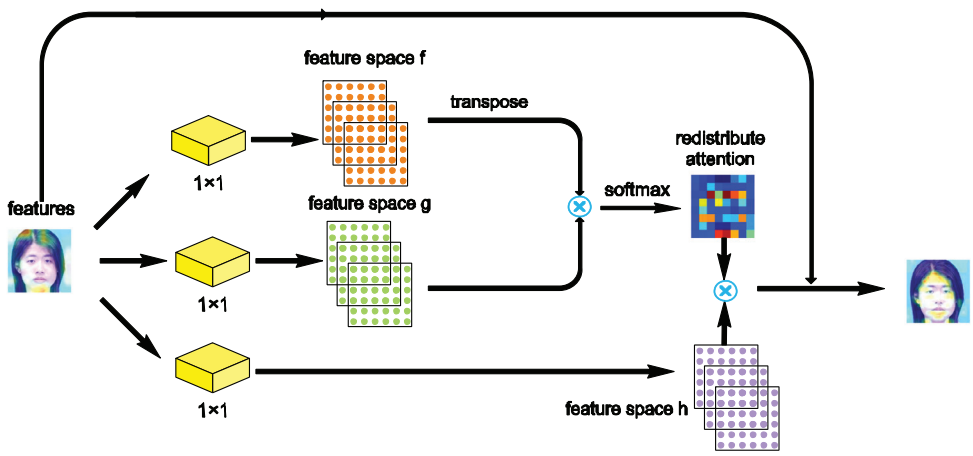


Fig. 3. Structure of the self-attention module. The softmax operation calculates the new attention distribution after matrix multiplication of the feature space f and g. The output is calculated by input features and the features with redistributed attention.

feature locations. We transpose $f(m)$ and perform matrix multiplication to calculate $s_{i,j} = f(m_i)^T g(m_j)$. After the softmax operation, we obtain $\beta_{i,j}$ as

$$\beta_{i,j} = \frac{\exp(s_{i,j})}{\sum_{i=1}^{N} \exp(s_{i,j})} \tag{2}$$

which indicates the degree of attention to the ith position when the model synthesizes the jth region. The attention of the input features is redistributed as $att = \{att_j \in \mathbb{R}^C\}, j = 1, 2, \ldots, N$, where

$$att_j = \sum_{i=1}^{N} \beta_{i,j} h(m_i), h(m_i) = W_h m_i. \tag{3}$$

In order to decrease the complexity of the model, we reduce the number of output channel of the 1×1 convolution as $C' = C/8$. Hence, the learnable weights $W_f \in \mathbb{R}^{C' \times C}$, $W_g \in \mathbb{R}^{C' \times C}$, $W_h \in \mathbb{R}^{C' \times C}$. The final features are calculated as

$$n_i = \gamma att_i + m_i \tag{4}$$

where γ is a learnable scalar initialized as 0. The value of γ increases as the training process progresses; the better the attention maps, the heavier the weight.

3.3. Loss function

The vast majority of prior research [10,11,28,35,36] adopts the L_1 loss to measure the generated image x_f and the target image x pixel-wise. Johnson et al. [37] and Chen et al. [38] used the pre-trained VGG network as a feature extractor and attempted to minimize the distance between the multi-level features of x and x_f. However, they missed the features of the highest and lowest level, respectively. In our model, full advantage is taken of the VGG-16 network pre-trained on the ImageNet dataset. As Fig. 1 shows, we measure the reconstruction loss as

$$\begin{aligned} L_{re(x,x_f)} &= \lambda_{1_2} L_\text{relu}1_2 + \lambda_{2_2} L_\text{relu}2_2 + \lambda_{3_3} L_\text{relu}3_3 \\ &\quad + \lambda_{4_3} L_\text{relu}4_3 + \lambda_{5_3} L_\text{relu}5_3, \end{aligned} \tag{5}$$

where

$$L_\text{relu}_l = ||\Phi^l(x) - \Phi^l(x_f)||_2^2. \tag{6}$$

$\Phi^l, l \in \{1_2, 2_2, 3_3, 4_3, 5_3\}$ represents the features from different feature extractor layers, and λ_l represents the weight. During the training process, the feature extractor based on pre-trained VGG-16 network is not updated, and the reconstruction loss calculated from the feature extractor penalizes the generator when the generated image deviates from the target image. We feed the output image of the pix2pix and the target image into the feature extractor and calculate the L_2 distance. The result is shown in Table 1. To balance the effects of features at different levels, the weights are set according to the distances; three combinations of the settings are compared, and the results are shown in Fig. 4. It is evident that Setting_1 and Setting_2 output visually more virtual images, while Setting_3 produces more realistic details such as hair.

With the condition information: sketch y, the loss function of cGAN can be expressed as

$$L_{GAN_G} = E_{x_f \sim p_{data}(x_f)}[D(x_f|y)] \tag{7}$$

$$L_{GAN_D} = E_{x_f \sim p_{data}(x_f)}[\log(1 - D(x_f|y))] + E_{x \sim p_{data}(x)}[\log D(x|y)] \tag{8}$$

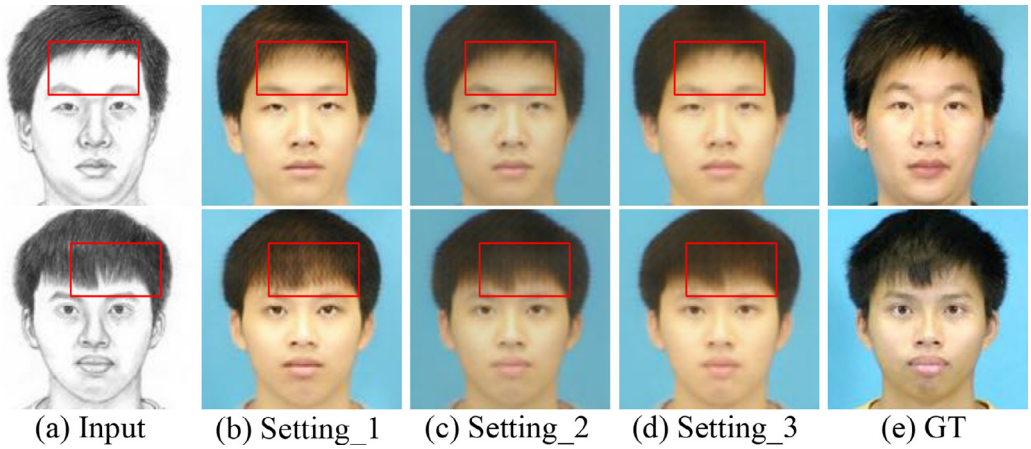


Fig. 4. Results of different combinations of settings. Setting_1 denotes the weights of five layers as 1500-150-50-150-500, Setting_2 is the combination of 3e3-3e2-1e2-3e2-1e3, and Setting_3 is the combination of 3e4-3e3-1e3-3e3-1e4.

Table 1
L_2 distance of the different level features of the output image of the pix2pix and the target image.

Layer	L_2 distance
1_2	0.0552
2_2	0.5849
3_3	1.5472
4_3	0.4143
5_3	0.1085

We follow previous work and retain the L_1 norm. Additionally, to slightly smoothen the output image, we add the total variation loss [39]:

$$L_{tv}(x_f) = \sum_{i,j} ((x_{f_{i,j+1}} - x_{f_{ij}})^2 + (x_{f_{i+1,j}} - x_{f_{ij}})^2)^{\frac{B}{2}} \tag{9}$$

where we set $B = 1$. Therefore, our final loss function can be formulated as

$$L_G = \lambda_1 L_1(x, x_f) + L_{GAN_G} + L_{re} + \lambda_{tv} L_{tv} \tag{10}$$

$$L_D = L_{GAN_D}, \tag{11}$$

where λ_1 and λ_{tv} are the weights used to control the importance of each term. The generator tries to minimize the L_G, and the discriminator tries to maximize the L_D. In this function, L_1 ensures that the generated image contains pixel level details. L_{GAN_G} attempts to exactly match the overall spatial structure of the output image with the target image. L_{re} encourages the output image to be perceptually similar to the target image, including in terms of color, texture, and exact shape. By optimizing this loss function with gradient descent, the complex correlations between the input sketch image and the target image can be deduced.

4. Experiments

4.1. Datasets

The CUHK Face Sketch (CUFS) and CUHK Face Sketch FERET (CUFSF) databases are designed for research on face sketch synthesis and recognition. The CUFS database contains 188 faces from the Chinese University of Hong Kong (CUHK) student database, 123 faces from the AR database [40], and 295 faces from the XM2VTS database [41]. There are 606 faces in total. For each face, there is a sketch drawn by an artist based on a photo taken in a frontal pose, with a neutral expression, and under the normal lighting condition. The CUFSF database includes 1194 faces from the FERET database [42]. For each face, there is a photo with lighting variation and a sketch with shape exaggeration drawn by an artist when viewing the photo. All images were resized to 256 × 256 to match the settings of our method.

4.2. Training

Our method was implemented using the PyTorch language. A mini-batch SGD and Adam optimizer were employed and set to a learning rate of 0.0002 and momentum parameter of 0.5. D was trained first, followed by G, alternately. The batch size and epoch were set to 1 and 200, respectively; the learning rate followed the StepLR policy and was decreased every 50 iterations. All of our experiments were conducted on an NVIDIA GeForce GTX TITAN Xp GPU.

4.3. Evaluation metrics

Evaluating the quality of generated images is a challenging problem. Traditional metrics such as per-pixel mean-squared error do not evaluate the structural loss; they are difficult to implement to evaluate the capability of generative models. Inception Score (IS) [43] based on Inception-V3 pre-trained on ImageNet is widely adopted. The IS will be low if the synthetic images are dissimilar to ImageNet because it considers a diverse set of generated images and ImageNet. Fréchet Inception Distance (FID) [44] utilizes Inception-V3 pre-trained on ImageNet as a feature extractor and measures the distance between features of the generated and target images, which is a more feasible approach. There is also the structural similarity index measure (SSIM), which is designed to measure the similarity between two images. The SSIM index is a full reference metric; essentially, its measurement or prediction of image quality is based on an initial uncompressed or distortion-free reference image. SSIM was designed to improve upon traditional methods such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE); it ranges from −1 to 1, where 1 indicates that the two measured images are completely identical.

4.4. Comparison with state-of-the-art models

We compare our method with state-of-the-art models, including pix2pix [10], Auto-painter [45], BicycleGAN [46], DiscoGAN [11], DualGAN [47] and CycleGAN [35], which are designed to perform image transformation tasks. Pix2pix, Auto-painter, and BicycleGAN trained on paired data are supervised learning processes. The other

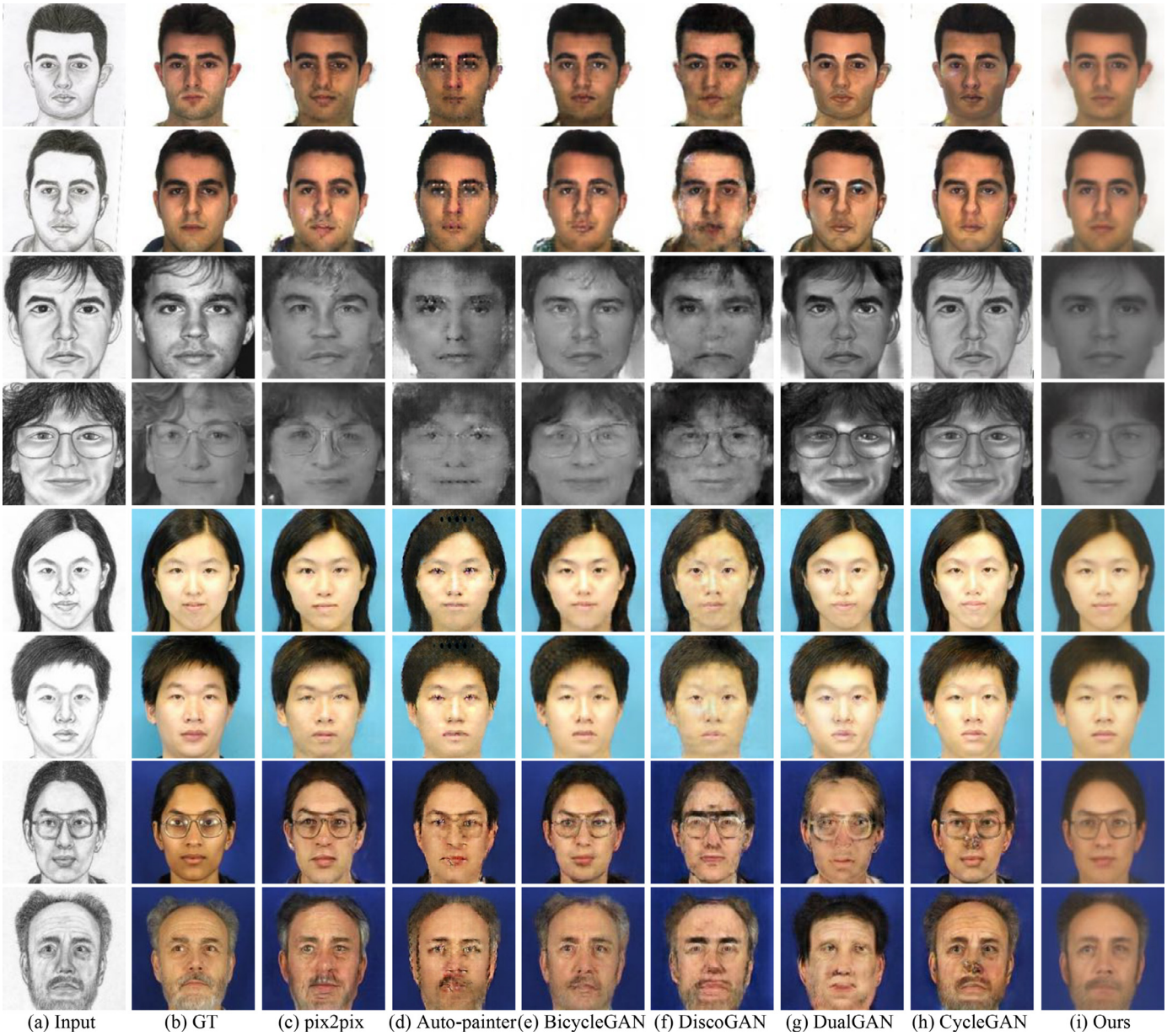


Fig. 5. Results of comparison with state-of-the-art models. We sample two translation results from each dataset. From top to bottom: AR, CUFSF, CUHK, XM2VTS.

Table 2
Quantitative results of comparison with baselines.

	pix2pix		Auto-painter		BicycleGAN		DiscoGAN		DualGAN		CycleGAN		Ours	
	FID	SSIM	FID	SSIM	FID	SSIM	FID	SSIM	FID	SSIM	FID	SSIM	FID	SSIM
AR	75.455	0.699	119.671	**0.748**	69.009	0.728	91.951	0.666	61.883	0.701	63.045	0.702	**61.309**	0.725
CUFSF	93.164	0.312	169.381	**0.688**	100.132	0.534	120.259	0.549	**66.812**	0.475	99.767	0.443	83.265	0.345
CUHK	90.424	0.631	119.201	0.705	91.125	0.697	101.117	0.634	58.534	0.660	68.005	0.647	**53.627**	**0.708**
XM2VTS	67.975	0.569	95.558	0.605	65.024	0.598	101.448	0.502	67.201	0.447	62.801	0.574	**60.043**	**0.606**

three are unsupervised learning models, designed to translate images in the absence of paired examples. For fair comparison, we train 200 epochs for these methods. The qualitative results are presented in Fig. 5, which clearly demonstrate that our method produces more visually realistic images in terms of both detail and overall quality on all four datasets. Pix2pix, Auto-painter, and BicycleGAN present fuzzy details. DiscoGAN produces unexpected results on all datasets, and DualGAN produces irrational results on the XM2VTS dataset. Moreover, CycleGAN displays visually virtual images on the first three datasets and does not handle the details effectively. As shown in Table 2, our method achieves the lowest FID score on the AR, CUHK, and XM2VTS datasets, and the highest SSIM score on the CUHK and XM2VTS datasets. DualGAN and DiscoGAN present better results (lower FID score and higher SSIM score, respectively) than our method on the CUFSF dataset; however, we observe that the images generated by DualGAN are visually similar to cartoon images. DiscoGAN generates more blurred images than our method. These images are nothing close to

Table 3
Ablation study: results of contribution of each module.

Module	FID	Module	FID
pix2pix (no module)	90.424	pix2pix+L_{re}+ms	64.391
pix2pix+L_{re}	73.784	pix2pix+L_{re}+att	62.581
pix2pix+ms	82.231	pix2pix+ms+att	65.696
pix2pix+att	79.662	pix2pix+all modules (ours)	**53.627**

Table 4
Ablation study: results of different variants of loss function.

Layer	FID
3	61.348
2-3-4	56.544
1-2-3-4-5 (ours)	53.627

natural photos. Auto-painter achieves the highest SSIM score on the AR and CUFSF datasets and a comparable SSIM score on the CUHK and XM2VTS datasets. This is because Auto-painter adopts the Wasserstein GAN [48], which is more effective for structure learning than the original GAN. Our method enhances this ability by optimizing the loss function containing different levels of the feature extractor.

4.5. Ablation study

To prove the effectiveness and rationality of the loss function, multi-scale architecture, and attention mechanism, we perform ablation studies on the CUHK dataset. The performance metric is the FID score alone.

4.5.1. Contribution of each module

To clarify the contribution of each module to the whole network, we separately add the reconstruction loss (L_{re}), multi-scale architecture (ms), and attention mechanism (att) to the pix2pix model. The quantitative results are listed in Table 3. From the table, it is evident that each module and the combinations of every two modules contribute to an improvement in the performance. The reconstruction loss function is more significant than the other two modules, and our method produces the best result by adding all of these modules.

4.5.2. Loss function

Multiple sets of experiments were conducted to demonstrate the effectiveness of the loss function design, and the results are shown in Table 4. The 3, $2-3-4$, $1-2-3-4-5$ indicate that we add the 3_3 layer alone, the 2_2, 3_3, 4_3 layers, and all the five layers of the feature extractor to the loss function. The quantitative results show that the first and last layers contribute significantly to the performance improvement. The FID score drops from 56.544 to 53.627, which indicates that these two layers should not be ignored. From Fig. 6, it can be seen that adding only the 3_3 layer to the loss function produces visually more virtual details such as hair and adding all five layers of the feature extractor to the loss function generates lighter and more visually realistic outputs.

Table 5
Ablation study: results of different designs of multi-scale architecture.

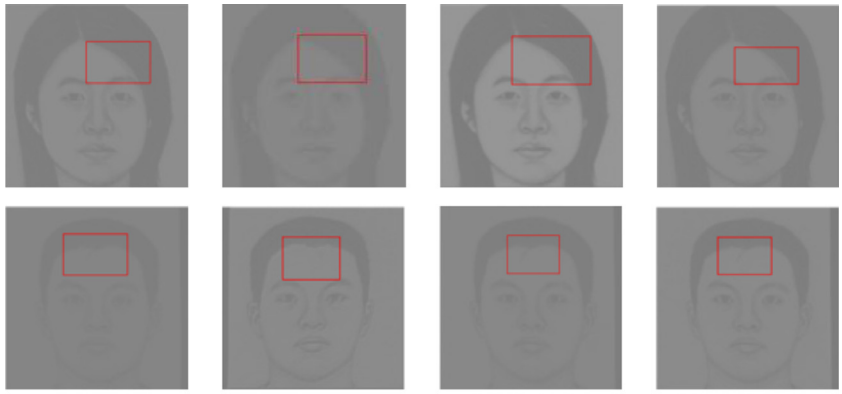
Kernel size	FID	Layer	FID
4	62.581	no ms	62.581
1–3	65.069	ms_1	58.114
1-3-5 (ours)	58.114	ms_12	54.864
1-3-5-7	57.448	ms_123 (ours)	53.627

4.5.3. Multi-scale architecture

To evaluate the effect of the multi-scale module, we replace the signal-scale convolution operation with a multi-scale filtering structure in the shallow encoder layers. We compare several different design factors, including kernel size and the number of layers to be updated. For the kernel size, we perform experiments on the first encoder layer, and the results are shown in the first two columns of Table 5. The $1-3$, $1-3-5$ and $1-3-5-7$ correspond to the combinations of 1×1, 3×3, 1×1, 3×3, 5×5 and 1×1, 3×3, 5×5, 7×7 kernel size, respectively. The 4 stands for the single scale 4×4 kernel size. We can see that the combinations of 1–3–5 and 1–3–5–7 produce a similar FID score. The combination of 1–3 shows the worst results because of the deficiency of the local receptive field. Fig. 7 is consistent with the quantitative results above. The combinations of 1–3–5 and 1–3–5–7 capture more comprehensive and precise details than the single-scale operation, and the 1–3 misses some information in the first layer. For the simplicity of the network, we adopt the combination of 1–3–5, which has a comparable FID score but fewer parameters.

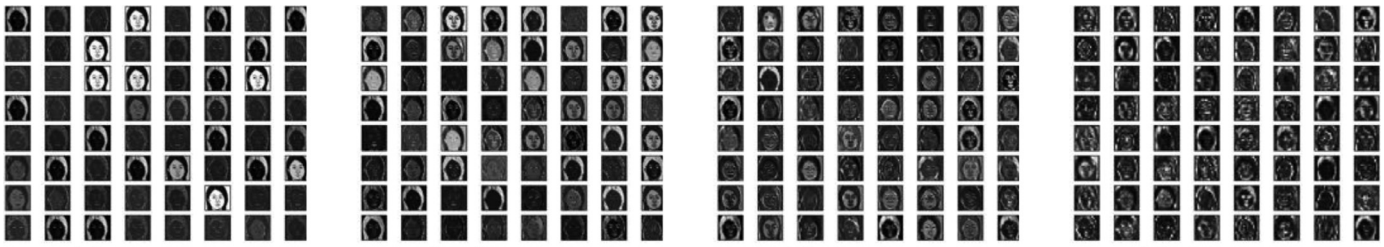
As for the number of layers that require modification, we display the visualization of the middle layers of our method without the multi-scale structure. Fig. 8 shows that the first three layers contain more detailed low-level information, while the fourth layer reflects more essential features and global information. Because of this, we substitute the multi-scale structure for the

Fig. 6. Results of adding different feature extractor layers to the loss function.



(a) 4 (b) 1-3 (c) 1-3-5 (d) 1-3-5-7

Fig. 7. Visual results of different types of kernel size combinations of the first encoder layer.



(a) layer_1 (b) layer_2 (c) layer_3 (d) layer_4

Fig. 8. Feature maps from different encoder layers.

(a) no ms (b) ms_1 (c) ms_12 (d) ms_123

Fig. 9. Results of middle layers. Those above are from the second layer and those below are from the third layer.

first three encoder layers. Layer-by-layer updation is performed, and the results are shown in the last two columns of Table 5. It can be seen that the multi-scale structure is more efficient for the first two layers than for the third layer in terms of improving performance. The FID score drops by about 5 and 3 after replacing the first and second layers, while for the third layer, it drops by about 1. This is because information is lost after flowing through the previous layers, and consequently, the third layer contains less detailed information than the first two layers. The middle features are shown in Fig. 9. We observe that the multi-scale structure not only extracts detailed features but also affects the quality of features in the next layers. For example, when the multi-scale structure is adopted in the first layer alone (ms_1), the second and third layers extract more prominent features than an implementation of our method without any multi-scale structure.

4.5.4. Self-attention mechanism

Zhang et al. [26] verifies that the self-attention mechanism produces better performance at the middle-to-high level features than at the low-level features. Accordingly, we add the self-attention module layer by layer from back to front. The FID score is shown in Table 6. As a result, adding the attention module at the fifth

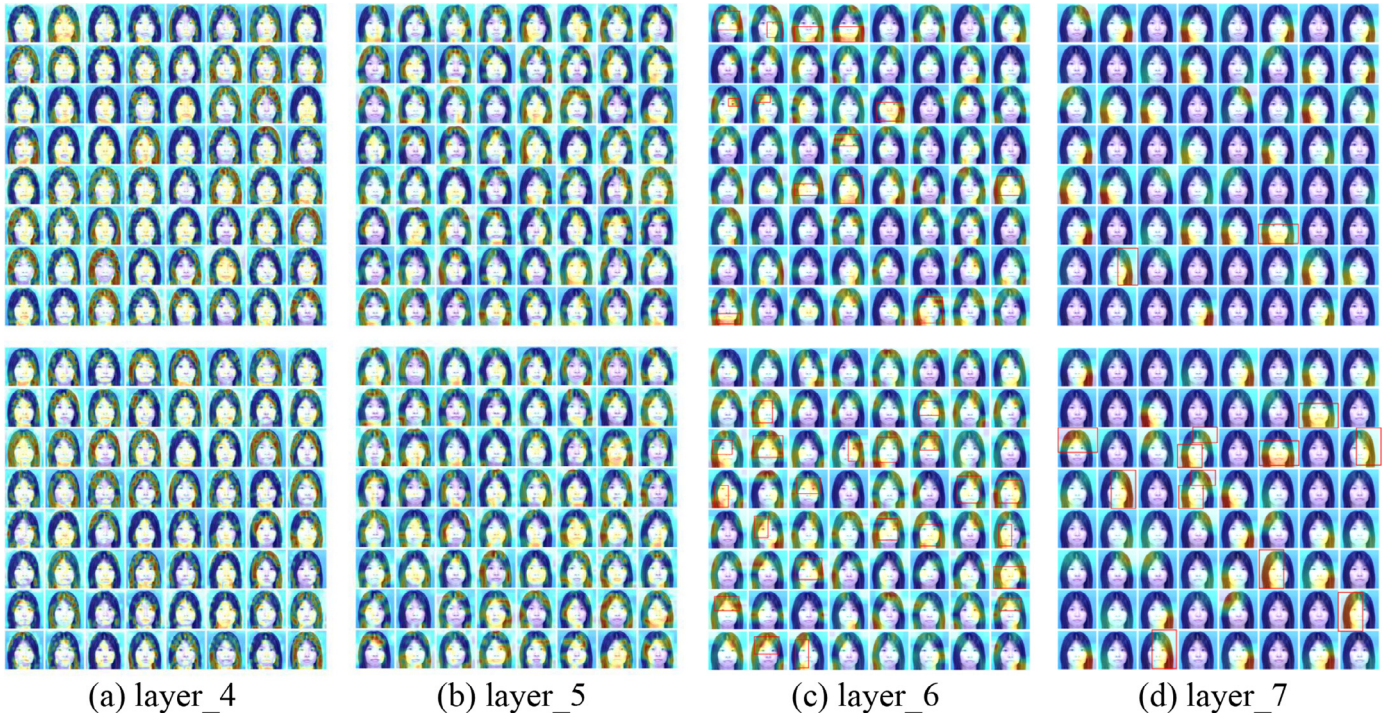


Fig. 10. Visualization of attention maps. Those above are from our method without the self-attention mechanism and those below are from our method with the attention mechanism at the fourth, fifth, sixth, and seventh layers.

Table 6
Ablation study: results of adding attention mechanism to different layers.

Layer	FID
no attention	64.391
7	59.638
67	56.865
567	54.964
4567 (ours)	53.627

and sixth layers decreases the FID score more prominently than the fourth layer, but less than the seventh layer. The FID score drops by slightly more than 1 on adding the attention module to the fourth layer; conversely, it drops by around 5 on adding the attention module to the seventh layer. We visualize the attention maps from the fourth, fifth, sixth, and seventh layers. From Fig. 10, it can be seen that the feature maps from the last layer (bottom right corner) redistribute more attention on the facial region after being processed by the attention module. The effect of attention redistribution is more apparent on the fifth and sixth layers than on the fourth layer, but not as apparent as on the seventh layer, which validates the quantitative results above.

5. Conclusion

In this paper, an advanced framework that follows the "divide and conquer" strategy is proposed. The multi-scale structure and the self-attention mechanism are combined, and fine details are extracted and maintained in the shallow encoder layers, while attention is redistributed in the deep encoder layers. Full use is made of the pre-trained VGG-16 network as the feature extractor, and a reasonable loss function is designed. Our ablation studies verify the validity and rationality of the various modules in our framework. It is demonstrated that our method shows a significant performance advantage as compared to other state-of-the-art approaches when handling the face sketch-to-photo transformation. Further, the performance of our method on different datasets proves its capacity and capability. This work provides a framework for the replacement of the traditional encoder structure by the "divide and conquer" strategy for GAN-based encoder–decoder methods and provides these methods with practical uses not limited to image transformation. Even though the proposed method achieves promising results, it still suffers from the defect of the GAN, similar to other GAN-based models, such as model collapse, which complicates the training process.

Although our model can achieve excellent translation results compared as to state-of-the-art ones, one limitation is that our results show a compromise in terms of brightness; from section *Loss function*, it can be seen that the brightness is influenced by the weights of the feature extractor. In future work, we will further explore the reasons for this consequence; additionally, we will apply our method to other images processing tasks and investigate different methods to more efficiently handle these tasks.

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.

CRediT authorship contribution statement

Yingtao Lei: Methodology, Software, Writing - review & editing, Writing - original draft, Visualization, Validation, Formal analysis, Investigation. **Weiwei Du:** Writing - review & editing. **Qinghua Hu:** Conceptualization, Writing - review & editing, Supervision, Resources.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 61925602, 61732011 and 61432011.

References

[1] Q. Liu, X. Tang, H. Jin, H. Lu, S. Ma, A nonlinear approach for face sketch synthesis and recognition, in: Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, I, 2005, pp. 1005–1010.

[2] D. Lin, X. Tang, Inter-modality face recognition, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006, pp. 13–26, doi:10.1007/11744085_2. 3954 LNCS

[3] H. Han, B.F. Klare, K. Bonnen, A.K. Jain, Matching composite sketches to face photos: a component-based approach, IEEE Trans. Inf. Forensics Secur. 8 (1) (2013) 191–204, doi:10.1109/TIFS.2012.2228856.

[4] L. Jie, X. Yu, C. Peng, N. Wang, Adaptive representation-based face sketch-photo synthesis, Neurocomputing. doi:10.1016/j.neucom.2016.10.095.

[5] B. Klare, Z. Li, A.K. Jain, Matching forensic sketches to mugshot photos, IEEE Micro 12 (1) (1992) 42–56, doi:10.1109/40.124381.

[6] Z. Cheng, Q. Yang, B. Sheng, Deep colorization, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 415–423, doi:10.1109/ICCV.2015.55. 2015

[7] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 105–114, doi:10.1109/CVPR.2017.19. 2017-January

[8] C. Li, M. Wand, Precomputed real-time texture synthesis with Markovian generative adversarial networks, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, pp. 702–716, doi:10.1007/978-3-319-46487-9_43. 9907 LNCS

[9] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Adv. Neural Inf. Process. Syst. 27 (2014) 2672–2680. arXiv: 1406.2661.

[10] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 5967–5976, doi:10.1109/CVPR.2017.632. 2017

[11] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: 34th International Conference on Machine Learning, ICML 2017, 4, 2017, pp. 2941–2949.

[12] W.-Z. Shao, J.-J. Xu, L. Chen, Q. Ge, L.-Q. Wang, B.-K. Bao, H.-B. Li, On potentials of regularized wasserstein generative adversarial networks for realistic hallucination of tiny faces, Neurocomputing (2019), doi:10.1016/j.neucom.2019.07.046.

[13] J. Guo, Y. Liu, Image completion using structure and texture GAN network, Neurocomputing (2019), doi:10.1016/j.neucom.2019.06.010.

[14] A. Dosovitskiy, T. Brox, Inverting visual representations with convolutional networks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 4829–4837, doi:10.1109/CVPR.2016.522. 2016

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9, doi:10.1109/CVPR.2015.7298594. 07-12-June

[16] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014, arXiv:1409.1556.

[17] H. Kazemi, S. Soleymani, A. Dabouei, M. Iranmanesh, N.M. Nasrabadi, Attribute-centered loss for soft-biometrics guided face sketch-photo recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 612–620, doi:10.1109/CVPRW.2018.00091. 2018-June

[18] Y. Lu, S. Wu, Y.W. Tai, C.K. Tang, Image generation from sketch constraint using contextual GAN, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11220 LNCS, 2018, pp. 213–228, doi:10.1007/978-3-030-01270-0_13.

[19] M. Mirza, S. Osindero, Conditional generative adversarial nets, CoRR (2014) CoRR arXiv:1411.1784.

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: 33rd International Conference on Machine Learning, ICML 2016, 3, 2016, pp. 1681–1690.

[21] G. Perarnau, J. van de Weijer, B. Raducanu, J.M. Álvarez, Invertible Conditional GANs for Image Editing, 2016, arXiv:1611.06355.

[22] M. Wang, H. Li, F. Li, Generative adversarial network based on resnet for conditional image restoration, IWACIII 2017 - 5th International Workshop on Advanced Computational Intelligence and Intelligent Informatics, 2017.

[23] L. Liu, H. Zhang, Y. Ji, Q.M. Jonathan Wu, Toward AI fashion design: an attribute-GAN model for clothing match, Neurocomputing (2019), doi:10.1016/j.neucom.2019.03.011.

[24] M. Sela, E. Richardson, R. Kimmel, Unrestricted facial geometry reconstruction using image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1585–1594, doi:10.1109/ICCV.2017.175. 2017-October

[25] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer, in: 35th International Conference on Machine Learning, ICML 2018, 9, 2018, pp. 6453–6462.

[26] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-Attention Generative Adversarial Networks, 2018, arXiv:1805.08318.

[27] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803, doi:10.1109/CVPR.2018.00813.

[28] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 6450–6458, doi:10.1109/CVPR.2017.683. 2017-January

[29] J. Han, G. Cheng, Z. Li, D. Zhang, A unified metric learning-based framework for co-saliency detection, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2018) 2473–2483, doi:10.1109/TCSVT.2017.2706264.

[30] G. Cheng, Z. Li, J. Han, X. Yao, L. Guo, Exploring hierarchical convolutional features for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 56 (11) (2018) 6712–6722, doi:10.1109/TGRS.2018.2841823.

[31] N. Nematzadeh, D.M. Powers, T.W. Lewis, Bioplausible multiscale filtering in retino-cortical processing as a mechanism in perceptual grouping, Brain Inform. 4 (4) (2017) 271–293, doi:10.1007/s40708-017-0072-8.

[32] C.N. Ochotorena, Y. Yamashita, Multi-scale structure-preserving image filtering, 2017 IEEE 19th International Workshop on Multimedia Signal Processing, MMSP 2017, 2017, doi:10.1109/MMSP.2017.8122290.

[33] X. Qian, Y. Fu, Y.G. Jiang, T. Xiang, X. Xue, Multi-scale deep learning architectures for person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5409–5418, doi:10.1109/ICCV.2017.577. 2017-October

[34] K.M. Nakanishi, S. ichi Maeda, T. Miyato, D. Okanohara, Neural multi-scale image compression, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, pp. 718–732, doi:10.1007/978-3-030-20876-9_45. 11366 LNCS

[35] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2242–2251, doi:10.1109/ICCV.2017.244. 2017-October

[36] H. Tang, D. Xu, N. Sebe, Y. Wang, J.J. Corso, Y. Yan, Multi-Channel Attention Selection GAN with Cascaded Semantic Guidance for Cross-View Image Translation, 2019, arXiv:1904.06807.

[37] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, pp. 694–711, doi:10.1007/978-3-319-46475-6_43. 9906 LNCS

[38] C. Chen, W. Liu, X. Tan, K.Y.K. Wong, Semi-supervised Learning for Face Sketch Synthesis in the Wild, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, pp. 216–231, doi:10.1007/978-3-030-20887-5_14. 11361 LNCS

[39] H.A. Aly, E. Dubois, Image up-sampling using total-variation regularization with a new observation model, IEEE Trans. Image Process. 14 (10) (2005) 1647–1659, doi:10.1109/TIP.2005.851684.

[40] A. Martinez, R. Benavente, The AR Face Database CVC Tech. Report #24 (1998), 1998.

[41] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre, XM2VTSDB: the extended M2VTS database, in: Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99), 1999, pp. 1–6. doi:10.1.1.35.468

[42] P. Jonathon Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104, doi:10.1109/34.879790.

[43] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.

[44] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, in: Advances in Neural Information Processing Systems, 2017, pp. 6627–6638. 2017-December

[45] Y. Liu, Z. Qin, T. Wan, Z. Luo, Auto-painter: cartoon image generation from sketch by using conditional wasserstein generative adversarial networks, Neurocomputing (2018), doi:10.1016/j.neucom.2018.05.045.

[46] J.Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A.A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, in: Advances in Neural Information Processing Systems, 2017, pp. 466–477., arXiv:1711.11586. 2017-December

[47] Z. Yi, H. Zhang, P. Tan, M. Gong, DualGAN: unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2868–2876, doi:10.1109/ICCV.2017.310. 2017-October

[48] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: 34th International Conference on Machine Learning, ICML 2017, 2017.

Yingtao Lei received his B.S. degree from Hunan University, China, in 2018. He is currently a postgraduate at the College of Intelligence and Computing, Tianjin University, China. His research interests include image processing, deep learning, and representation learning.

Weiwei Du received her B.S. degree from Tianjin Institute of Urban Construction, China, in 2002. She received her Master's degree from University of Aizu in 2005 and her Ph.D. degree from the University of Kyushu in 2008, which are both situated in Japan. She joined the Kyoto Institute of Technology as an assistant professor in 2008. Since 2009, she has been working as a full-time associate professor in the Department of Information Science at the Kyoto Institute of Technology, Japan. Her research interests are pattern recognition, machine learning, neural networks, image processing, and computer vision.

Qinghua Hu received his B.S., M.S., and Ph.D degrees from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively. He was a Post-Doctoral Fellow with the Department of Computing, Hong Kong Polytechnic University, from 2009 to 2011. He is currently the Dean of the School of Artificial Intelligence, the Vice Chairman of the Tianjin Branch of China Computer Federation, the Vice Director of the SIG Granular Computing and Knowledge Discovery, and the Chinese Association of Artificial Intelligence. He is currently supported by the Key Program, National Natural Science Foundation of China. He has published over 200 peer-reviewed papers. His current research is focused on uncertainty modeling in big data, machine learning with multi-modality data, and intelligent unmanned systems. He is an Associate Editor of the IEEE Transactions on Fuzzy Systems, Acta Automatica Sinica, and Energies.