Tech Science Press

Human Face Sketch to RGB Image with Edge Optimization and Generative Adversarial Networks

Feng Zhang[1], Huihuang Zhao[1,2,*], Wang Ying[1,2], Qingyun Liu[1,2], Alex Noel Joseph Raj[3] and Bin Fu[4]

[1]College of Computer Science and Technology, Hengyang Normal University, Hengyang, 421002, China
[2]Hunan Provincial Key Laboratory of Intelligent Information Processing and Application, Hengyang, 421002, China
[3]Key Laboratory of Digital Signal and Image Processing of Guangdong, Shantou, 515063, China
[4]Department of Computer Science, University of Texas Rio Grande Valley, Edinburg, TX, USA
*Corresponding Author: Huihuang Zhao. Email: happyzhh@hynu.edu.cn

Abstract: Generating an RGB image from a sketch is a challenging and interesting topic. This paper proposes a method to transform a face sketch into a color image based on generation confrontation network and edge optimization. A neural network model based on Generative Adversarial Networks for transferring sketch to RGB image is designed. The face sketch and its RGB image is taken as the training data set. The human face sketch is transformed into an RGB image by the training method of generative adversarial networks confrontation. Aiming to generate a better result especially in edge, an improved loss function based on edge optimization is proposed. The experimental results show that the clarity of the output image, the maintenance of facial features, and the color processing of the image are enhanced best by the image translation model based on the generative adversarial network. Finally, the results are compared with other existing methods. Analyzing the experimental results shows that the color face image generated by our method is closer to the target image, and has achieved a better performance in term of Structural Similarity (SSIM).

Keywords: Human face sketch; generative adversarial networks; RGB image; edge optimization

1 Introduction

The translation of a sketch into an RGB image is an important research topic in the field of computer graphics and computer vision. This translation is the conversion of an image into a corresponding output image. In general, image translation is based on the end-to-end architecture of deep learning. Conditional Generative Adversarial Nets (CGANS) [1] can realize the unified processing of many kinds of images and accelerate the development of image transformation. Nowadays, the process of image translation is much simpler, but the ability to do it is much greater. This paper mainly studies the application of Generative Adversarial Nets (GAN) [2] in the transformation of face sketch into an RGB image.

Gatys et al. [3] proposed a neural network based image style transfer algorithm, which computes the image style features using a Gram matrix and then uses the graphs to generate the image style features and later complete the content and texture modeling of the image. The gradient descent algorithm is used to train the input image in the way of error back-propagation, thus the image is styled and final result with style transfered image is obtained. However, this method belongs to one-to-one mapping, which requires alarge number of paired images, involving various types of images, a lot of time and effort. Unfortunately, resolution of the output image is low, and the interference of early generation is relatively large, and thus can be used in practice in a smaller range. Li et al. [4] proposed a combination algorithm of Markov random field model and deep convolutional neural network, referred to as the CNNMRF model. The principle is to use the most similar regions in the style to match each input neural region by reducing the loss of image features. It makes the generated image more similar to the input image. But this method does not handle the style of the image well. Phillip et al. [5] proposed a general mode of image translation, which is based on the CGAN target loss function and added an ℓ_1 loss to the CGAN target loss function. The U-Net [6] structure is used in the generation, and the PatchGAN [7] structure is used the discriminator. In this paper, the effects of several important architectural choices are analyzed. The pix2pix image is clearer and more natural than the CGAN image, but the experiment needs supervised learning and the data set is difficult to obtain. Huaming et al. [8] proposed a GAN-based face image translation, based on the dual-generated adversarial network. This model was formed by adding two loss functions to its objective function. The processing effect of the face sketch is still good, but the processed image is still a gray image. The above models have good experimental results in their respective data sets but do not process the human face sketch image as an RGB image [9].

This paper presents a novel method for transforming the human face sketch image into an RGB image. The main contributions of the paper are as follows:

—A method of transforming a human face sketch image into an RGB image is proposed, which is based on generating the adversarial neural network model

—An improved optimized target loss function is proposed to enhance the image boundary, resulting in a higher resolution output image.

2 Principles

2.1 Generative Adversarial Networks

The generative adversarial network consists of two parts, generator G and discriminator D [10], and its training process can be considered as a zero-sum game between generator and discriminator. Among them, to learn the distribution of real samples, the generator G can generate fake images and continuously output fake images to fool the discriminator. The role of D is used to detect whether the given image is true or false, and present a classification score. In the training process, the goal of generating network G is to try to generate real pictures to deceive and discriminate network D. The goal of D is to try to separate the pictures generated by G from the real pictures. After several iterations, both of them are optimized [11]. When the discriminator cannot distinguish the generated picture from the input picture, it indicates that both networks have been optimized and a balance point has been reached. The performance of the GAN network is mainly evaluated by comparing the obvious difference between the output image and the target output image. If the difference is not obvious, it indicates that the GAN has better performance. The training process of the network model is shown in Fig. 1.
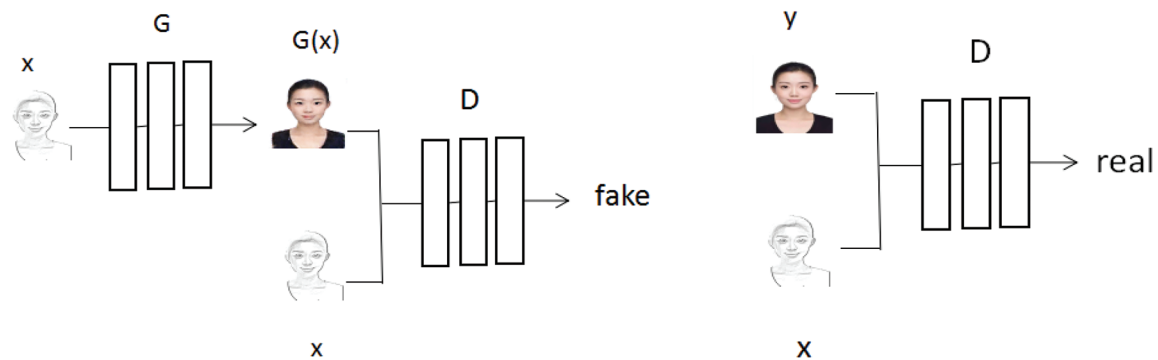
Figure 1: Generated adversarial networks architecture

2.1.1 Generator

During the proposed method, the Generator architecture is based on the "U-Net" architecture [12], and we use the U-Net structure, where skip connections are introduced between the encoder and the decoder to increase the connection between the image layer [13]. In the generator network, the input image is preprocessed by a network composed of two convolution layers, a batch normalization layer and a linear activation unit at each layer of the network [14]. The U-Net network structure is shown in Fig. 2. It is added skip connections between layers in the encoder-decoder.
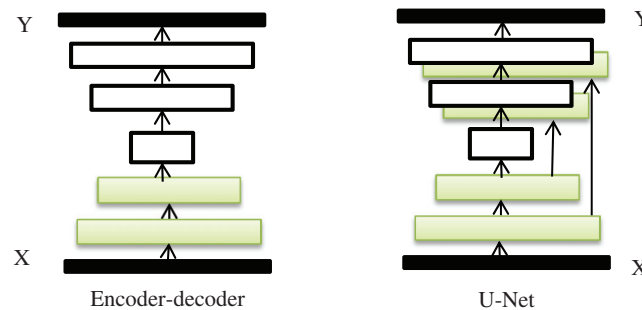
Figure 2: Two choices for the architecture of the generator

2.1.2 Discriminator

For the discriminant, we employ the reference architecture in [15], and theclassifier in PatchGAN is used to classifygtrue and false. During our method, we modify it with a different patch size. The size of the convolution kernel in the generator network we used was 4 × 4, with a step size of 2. Its output is activated by *relu* function and later by Tanh [16]. The final result was mapped out by (0, 1), where the output of 0 is false, 1 is true.

2.2 Image Style Transformation

With the development of deep learning, image-to-image [16,17] style transformation or image-to-image translation has become a challenging topic in research. The original image style transform was proposed by Gatys et al. [18]. Its principle is to use VGG16 model to extract image features from two images. Goodfellow et al. [19] proposed the method of the generative adversarial network, combined with deep learning, and introduced the generative adversarial network. This generative adversarial learning method learns from the original data in the data set and generates the data distribution that can act as the real data. Zhu et al. [16] proposed the CycleGAN method, to convert pictures of one style into pictures of another style. The

architecture represents a general structure for processing image style transformation, and the implementation requires the use of a generative antagonistic network for unpaired data sets.

There are two types of image style transforming namely original image style translated algorithm and the fast style translate algorithm. The input images are an original image and a style image. The loss function is constructed by using the loss value of an initial image, and the generated graph by feedback and redrawing it [20]. However, this algorithm requires network training every time a picture is generated, which is complex and not suitable for larger data sets. Johnson et al. [21] proposed a fast style transform algorithm, whose main idea is to train a network that can translate any picture into the corresponding style of the network. The algorithm includes an image transform network and a loss network [22]. During training, a large number of pictures are used in the training stage to get the model, and then apply the model to get the resulting output to get the generated picture.

In this paper, GAN network structure is adopted for style transformation [23,24]. The unpaired RGB image data set is preprocessed to obtain corresponding sketch images, and these two images are saved as training sets and verification sets for deep learning. An improved generative adversarial network which can realize the supervised learning of two data sets is used to generate a training model [25]. And we can randomly select a face sketch image which is from a common dataset and feed it to the trained network, then obtain the final RGB image through hue mapping [26,27].

2.3 Network Architectures

We adjusted our network architecture with a BRc(Convolution-BatchNorm-ReLU layer and c filters) [28]. The Convolution—BatchNormDropout—ReLU layer with a 50% loss rate is represented by BRc. The convolutions used in the experiments are 4 × 4 spatial filters with a step size of 2 [29]. The convolution is down-sampled at 2 times in the encoder and discriminator and up-sampled at 2 times in the decoder [30].

2.3.1 Structure of the Generator

The codec architecture includes:

Encoder: B64–B128–B256–B256–B256–B512–B512–B512

Decoder: BR512–BR512–BR512–BR256–BR256–B128–B64

After the last layer of the decoder, we use convolution to map to several output channels (the image being shaded is 2), and finally, use the Tanh function. With U-Net architecture, the number of channels between the i-th and n-i layers changes so the number of channels in the decoder also changes.

U-Net decoder: BR512–BR1024–BR1024–B1024–B512–B512–B256–B128

2.3.2 Structure of the Discriminator

The 128 × 128 discriminator architecture is: B64–B128–B256–B256–B512–B512

After the last layer, the convolution map is applied to one dimension for output and then processed using the Sigmoid function. All leaky ReLUs have a slope of 0.2.

3 Methods

We adopt the method of generative adversarial networks, G as the generation model, from the random noise vector X to the output image Y:G:X→Y learning mapping. While CGAN is learning a mapping from observed image Z and random noise vector X to output image Y:G: {Z,Y}→Y.

The objective function of GAN enforced by the condition can be expressed as

$$L_{cGAN}(G, D) = E_{x,y}[\log D(x, y)] + E_{x,z}[\log(1 - D(x, G(x, z)))], \tag{1}$$

where E is the mathematical expectation. When generator G is trained to minimize this goal, discriminator D maximizes this goal, i.e., $G^* = \arg\min_{G}\max_{D} L_{cGAN}(G, D) + \lambda_1 L_T(G)$, where $L_T(G)$ is defined as

$$L_T(G) = E_{x,y,z}[||y - G(x, z)||_1], \tag{2}$$

To test the importance of the adjustment discriminator, we also compared an unconditional variable in which the discriminator did not observe x:

$$L_{GAN}(G, D) = E_y[\log D(y)] + E_{x,z}[\log(1 - D(G(x, z)))], \tag{3}$$

Previous approaches have found it beneficial to mix the GAN objective with a more traditional loss, such as ℓ_2–norm [6]. The discriminator's job remains unchanged, but the generator is tasked to not only fool the discriminator but also to be near the ground-truth output in ℓ_2–norm. We further make use of the improved loss proposed by pix2pix [8], where the discriminator does not observe X and ℓ_1–norm is mixed with gan target to encourage the reduction of blurring [19].

Our total loss function is defined as

$$G^* = \arg\min_{G}\max_{D} L_{cGAN}(G, D) + \lambda_1 L_T(G) + \lambda_2 L_{Edge}(G), \tag{4}$$

where λ_1 and λ_2 control the importance of two items, L_{Edge} is the loss of edge in the result image. Our network learning is consistent with pix2pix, and only uses the form of noise reduction in the partial convolution layer of the generator, that is to say, network learning only maps from X to Y without Z. The image edge information constraint added to the objective is used to guide the generator to generate clearer and realistic portrait.

4 Experiments

4.1 Experimental Platform and Data

In this paper, 64-bit Ubuntu16.04 operating system and GPU based Tensorflow1.4.1 deep learning framework are used in the experiment [4]. The CelebFaces Attributes Dataset (CelebA) is used as the training data set to train our model [13]. CelebA dataset covers large pose variations and background clutter. And it has large diversities, large quantities. The face images were used to convert between RGB images and face sketch images, and there 200 face selecting images which down load from internet are used as test data set to evaluate the generalization of the network model. The selected face samples are mainly clear-faced, well-organized images and the background in the image includs white, blue, and light gray colors. Experiments have shown that images with white and light gray backgrounds work better.

4.2 Details of the Implementation of the Network Model

This paper studies the transformation of the face sketch image into an RGB image, which follows the basic framework of the generative adversarial network [15,17]. Modules in the form of convolution-BatchNormReLu [14,19] are used by both generators and discriminators. The generator architecture uses U-Net to skip connections to break through the bottleneck reached by sampling. The discriminator uses PatchGAN, which only works on the structure and patches and classifies the authenticity of each N × N patch. We tested different N values and found that N is much smaller than the size of the original image and has little impact on the output of the image, so there is no requirement on the size of the input image, so this network model can operate on any size of the image.

We tested PatchGANs of different sizes. After 232 iterations, we found that the image output using the PatchGANs patch of 64 × 64 is clearer. The output of images using other sizes of PatchGANs is not good. By adjusting the size of the 256 × 256 input image to randomly dither the image to 286 × 286, then randomly cut to the size of 256 × 256, 128 × 128, 64 × 64, 32 × 32, 16 × 16 output image, and then choose the best experimental results. In this paper, all the networks are randomly trained. The weights are initialized from a Gaussian distribution with an average of 0 and a standard deviation of 0.02. All the experiments were performed in batches, with the activated bottleneck cleared by the batchnorm operation to effectively skip the innermost layer. Leave the experiment as it is by removing the innermost batch.

4.3 Transform Face Sketch to RGB Image

To evaluate the transformation performance of the model in this paper from sketch face images to RGB images, the experimental data were composed of 600 random face sketched images as the source data and RGB face images as the target output of the network model. Experimental results are shown in Fig. 3. In the image transformation, the model guarantees the similarity between the output image and the target image, the image noise can be ignored, and the generated RGB image is realistic.

One can easily find that our method can achieve a better result than methods of Gatys and CNNMRF. The result images are more close to the target images and the edge is more clear.

4.4 Comparing

To show more intuitively the superiority of the generation adversarial network model with edge enhancement in human face sketch image transformation, we compare some of the images with Gatys method [18] and CNNMRF method [4]. The comparison is both quantitative (visual) and quantitative.

4.4.1 Visual Comparison

In this paper, some experimental results are compared intuitively. Some results based on Gatys method and CNNMRF method are selected and compared. We found that Gatys method produces the worst image effect, we could see the basic outline, but the color processing effect is obviously not good, facial features are also relatively fuzzy. The CNNMRF method is better than Gatys method in image color and diffent emotions, but image contours are not blurred. The image generated in this experiment is closer to the target image in terms of facial contour, facial features, and image color. Also, it maintains the facial features of the input image well. In general, in the experiment of transforming the human face sketch image into an RGB image, the experimental result of this model is better than that of other models.

To demonstrate the superiority of the experiment in a better way, we tested face images with complex wearing colors. The results show that our method is not only better than Gatys method and CNNMRF in the effect of face feature transformation but also better than these two methods in the transformation of clothing color. The experimental results are shown in Fig. 4.

4.4.2 Quantitative Comparison

To further illustrate the reliability of the experimental results, we conducted a quantitative comparison. We invited 30 students to perceive the effects of several techniques on image processing, including GAN, Neural style [18] and CNNMRF [4]. Participants were asked to rate the generated images. The scores ranged from 1 to 5, representing poor, pass, medium, good, and excellent. Each method has 20 images from a total of 60 images. Fig. 5 summarizes the average scores of the three methods, reflecting the similarity between the generated image and the target image. It can be seen from the figure that the image generated by the GAN method has the highest score, which also reflects the contribution of this experiment.

Figure 3: Results of face sketch image transformation into RGB image. (a) Input, (b) target, (c) gatys, (d) CNNMRF, (e) our method

*4.4.3 SSIM Method for Image Similarity Comparison*

To make the explanation more convincing, we used the SSIM method (Structural similarity) to analyze the similarity of the images obtained by the three methods. SSIM is an index to measure the similarity of two images, which consists of brightness contrast, contrast, and structure contrast [9]. The two images are more similar, where the value of SSIM (x, y) is closer to 1. The following is a comparison of the SSIM (x, y) values of some pictures we intercepted, as shown in Tab. 1.
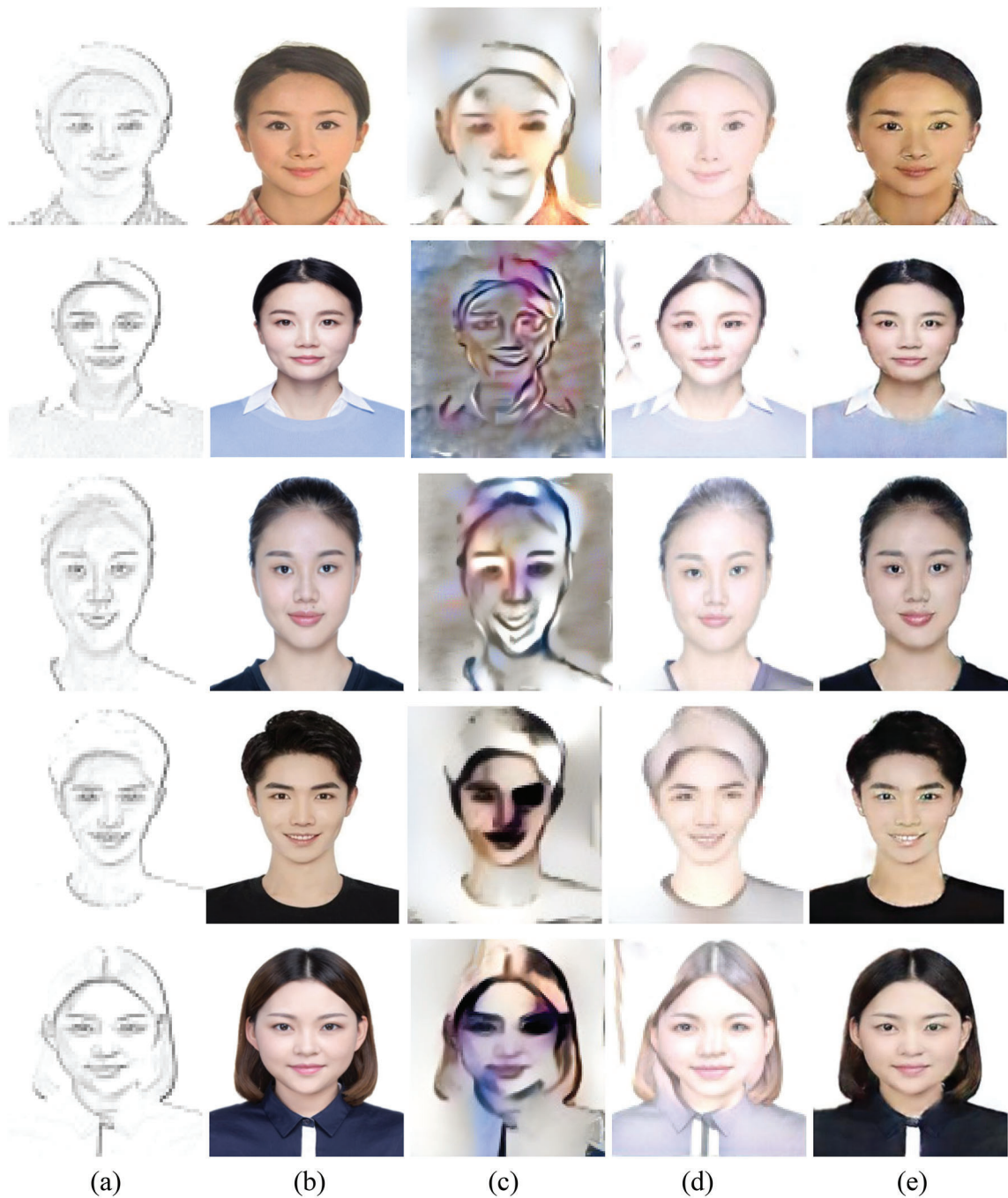
Figure 4: Result of conversion of face images in complex color clothes. (a) Input, (b) target, (c) gatys, (d) CNNMRF, (e) our method

It can be seen from the above data that the image processed by the image edge enhancement method has the highest similarity to the target image. The similarity of the images processed by Gatys method and CNNMRF method is relatively low. This also reflects the feasibility and superiority of the method proposed in this article.
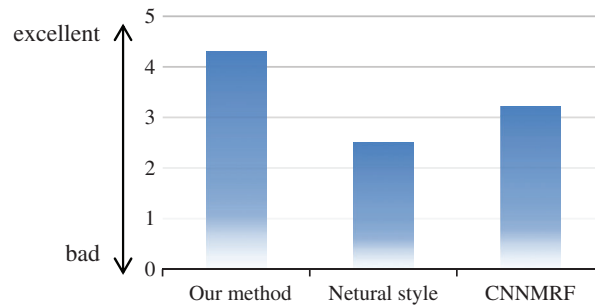
Figure 5: User study results proving that our method better reflects the style of the RGB image

Table 1: Comparison of image similarity generated by three methods (Intercept part)

Image number	Gatys	CNNMRF	Our methods
1	0.5359	0.7857	0.8850
2	0.4690	0.6252	0.8475
3	0.6307	0.7008	0.8631
4	0.6536	0.6705	0.8540
5	0.6727	0.8071	0.8813
6	0.5877	0.6989	0.8651
7	0.5382	0.7696	0.8713
8	0.5233	0.5768	0.8591
9	0.6541	0.7119	0.9038
10	0.3410	0.5124	0.8301

5 Conclusion

In this paper, an improved model by adding an edge optimization and Generative Adversarial Networks is proposed. Through a series of experiments, it is shown that the model proposed in this paper is superior to the traditional method in terms of the definition and feature preservation of face sketch image transformation. The disadvantage is that the data set used in this paper is small, and some of the image transformation details are not good enough. Besides, our model is only better than other models in single-channel human face sketch image transformation, but not the best in multi-channel processing sketch image transformation. The next step is to refine the model to best handle any complex image and try to experiment with multi-channel images.

Conflicts of Interest: We declare that they have no conflicts of interest to report regarding the present study.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, W. F. David *et al.,* "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, no. 2, pp. 2672–2680, 2014.

[2] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *IEEE Int. Conf. on Computer Vision*, Honolulu, Hawaii, USA, pp. 3730–3738, 2017.

[3] C. Li, R. V. Sánchez, G. Zurita, M. Cerrada and D. Cabrera, "Fault diagnosis for rotating machinery using vibration measurement deep statistical feature learning," *Sensors*, vol. 66, no. 6, pp. 895, 2016.

[4] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," in *IEEE Int. Conf. on Computer Vision*, Las Vegas, NV, USA, pp. 1063–6919, 2016.

[5] D. J. Zeng, Y. Dai, F. Li, J. Wang and A. K. Sangaiah, "Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 3971–3980, 2019.

[6] Z. Zhang, Q. Liu and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 1749–1753, 2018.

[7] Y. J. Luo, J. H. Qin, X. Y. Xiang, Y. Tan, Q. Liu *et al.,* "Coverless real-time image information hiding based on image block matching and dense convolutional network," *Journal of Real-Time Image Processing*, vol. 17, no. 1, pp. 125–135, 2020.

[8] N. Wang, W. Zha, J. Li and X. Gao, "Back projection: An effective postprocessing method for gan-based face sketch synthesis," *Pattern Recognition Letters*, vol. 107, no. 5, pp. 59–65, 2018.

[9] H. Zhao, P. L. Rosin and Y. Lai, "Block compressive sensing for solder joint images with wavelet packet thresholding," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 6, pp. 1190–1199, 2019.

[10] W. Zhang, G. Li, H. Ma and Y. Yu, "Automatic color sketch generation using deep style transfer," *IEEE Computer Graphics and Applications*, vol. 39, no. 2, pp. 26–37, 2019.

[11] A. V. Fonin, S. K. Asiya, I. A. Gagarskaia, E. I. Kostyleva and M. Karasev *et al.,* "Intrinsically disordered proteins ph-induced structural transitions in overcrowded milieu," *Biophysical Journal*, vol. 114, no. 3, 591a, 2018.

[12] Y. B. Yoon, M. S. Kim and H. C. Choi, "End-to-end learning for arbitrary image style transfer," *Electronics Letters*, vol. 54, no. 22, pp. 1276–1278, 2018.

[13] A. G. Mullissa, C. Persello and A. Stein, "PolSARNet: A deep fully convolutional network for polarimetric SAR image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 12, pp. 5300–5309, 2019.

[14] D. Hu, S. Peng, X. Liu and J. Du, "Motion style transfer via deep autoencoder and spatio-temporal feature constraint," *Journal of Computer-Aided Design and Computer Graphics*, vol. 30, no. 5, pp. 946–956, 2018.

[15] D. Holden, I. Habibie and I. Kusajima, "Fast neural style transfer for motion data," *IEEE Computer Graphics & Applications*, vol. 37, no. 4, pp. 42–49, 2017.

[16] S. R. Zhou, M. L. Ke and P. Luo, "Multi-camera transfer GAN for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 59, no. 1, pp. 393–400, 2019.

[17] Z. L. Wang, Z. Z. Chen and F. Wu, "Thermal to visible facial image translation using generative adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1161–1165, 2018.

[18] L. A. Gatys, A. S. Ecker and M. Bethge, "Image style transfer using convolutional neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 2414–2423, 2016.

[19] C. Ruan, D. Chen and H. Hu, "Multimodal supervised image translation," *Electronics Letters*, vol. 55, no. 4, pp. 190–192, 2019.

[20] L. Zhao, H. Bai, J. Liang, B. Zeng, A. Wang et al., "Simultaneous color-depth super-resolution with conditional generative adversarial networks," *Pattern Recognition*, vol. 88, no. 4, pp. 356–369, 2019.

[21] J. Johnson, A. Alexandre and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 694–711, 2016.

[22] B. Huang, W. Chen, X. Wu, C. Lin and P. N. Suganthan, "High-quality face image generated with conditional boundary equilibrium generative adversarial networks," *Pattern Recognition Letters*, vol. 111, no. 8, pp. 72–79, 2018.

[23] X. Mao, S. Wang, L. Zheng and Q. Huang, "Semantic invariant cross-domain image generation with generative adversarial networks," *Neurocomputing*, vol. 293, no. 7, pp. 55–63, 2018.

[24] J. Liu, C. K. Gu, J. Wang, G. Youn and J. U. Kim, "Multi-scale multi-class conditional generative adversarial network for handwritten character generation," *Journal of Supercomputing*, vol. 75, no. 4, pp. 1922–1940, 2019.

[25] N. Akai, L. Y. M. Saiki, T. Hirayama and H. Murase, "Misalignment recognition using markov random fields with fully connected latent variables for detecting localization failures," *IEEE Robotics & Automation Letters*, vol. 4, no. 4, pp. 3955–3962, 2019.

[26] T. Portenier, Q. Hu, A. Szabo, S. A. Bigdeli and P. Favaro et al., "Faceshop: Deep sketch-based face image editing," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 99.1–99.113, 2018.

[27] S. Li, F. Liu, J. Liang, Z. Cai and Z. Liang, "Optimization of face recognition system based on azure IoT edge," *Computers, Materials & Continua*, vol. 61, no. 3, pp. 1377–1389, 2019.

[28] J. H. Wu, F. Huang, W. J. Hu, W. He, B. Tu et al., "Face detection method for public safety surveillance based on convex grouping," *Computer Systems Science and Engineering*, vol. 33, no. 5, pp. 327–334, 2018.

[29] J. Liu, Y. H. Yang, S. Q. Lv, J. Wang and H. Chen, "Attention-based BiGRU-CNN for Chinese question classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 6, pp. 1–12, 2019.

[30] J. M. Zhang, X. K. Jin, J. Sun, J. Wang and A. K. Sangaiah, "Spatial and semantic convolutional features for robust visual object tracking," *Multimedia Tools and Applications*, vol. 79, no. 7, pp. 15095–15115, 2020.