A Style-Based Caricature Generator

Lamyanba Laishram[1][0000−0002−0324−214X], Muhammad Shaheryar[1][0000−0003−3992−1387], Jong Taek Lee[1][0000−0002−6962−3148] and Soon Ki Jung[1][0000−0003−0239−6785]

School of Computer Science and Engineering, Kyungpook National University, Republic of Korea
{yanbalaishram, shaheryar, jongtaeklee, skjung}@knu.ac.kr

Abstract. A facial caricature is a creation of new artistic and exaggerated faces which translates into a real image to convey sarcasm or humor while keeping the identity of the subject. In this work, we proposed a new way to create caricatures by exaggerating facial features like the eyes and mouth while keeping the facial contour intact and a realistic style. Our method can be categorized into two steps. First, the facial exaggeration process transformed faces into caricature face images while maintaining facial contours. Second, the appearance style generator is trained in unpaired using the generated caricature faces to produce a facial caricature that can change to any realistic style of our preference. Experimental results show our model produces more realistic and disentangled caricature images as compared to some of the previous methods. Our method can also generate caricature images from real images.

Keywords: Caricature · Style Generator · Generative Adversarial Network

1 Introduction

In the world of comics, animation, posters, and advertising, in particular, artistic portraits are very common in our daily lives. A caricature is a representation of a person whose distinctive features are simplified or exaggerated through sketching or artistic drawings. A facial caricature is a form of art used to convey sarcasm or humor and is used commonly in entertainment.

Applications based on computer vision have a wide range and the creation of caricatures can be done without the need for an artist. Similar to the way an artist approach creating caricatures, a method based on computer vision can also be divided into two stages: (i) identifying the distinct features and exaggerating those features, and (ii) applying styles to the deformed image according to the artist's taste. The separation of these two categories provides flexibility and disentanglement which eventually results in the generation of good-quality caricatures.

Earlier approaches for creating a facial caricature require professional skills to get good results [2]. Traditional artworks tended to emphasize exaggerating facial forms by increasing the shape representation's divergence from the average,

as in the case of 2D landmarks or 3D meshes [3, 27, 14]. With the advancement in applications of computer vision techniques, several automated caricature generations have emerged [33, 11, 4]. Moreover, automatic portrait style transfer based on image style transfer [22, 32, 24] and image-to-image translation [21] have been extensively studied. Recently with the development in the Generative adversarial networks (GANs) [13], the state-of-the-art face generator StyleGAN [19, 20] provides disentangled and high-fidelity artistic images via transfer learning.

In recent years, Deep learning techniques are very successful in performing image-to-image translation by learning from representation data examples [15, 16]. Unfortunately, paired real and caricature are not commonly found. The translation process is not feasible to be trained in a supervised manner and building such a dataset is tedious. One of the readily available caricature datasets is WebCaricature [17], which consists of 6042 caricatures and 5974 photographs from 252 different identities. In our work, we created a set of 10,000 caricature images from the FFHQ dataset [19] for automatic caricature generation which we will discuss in Section 3.

Due to the limited data availability of paired images, most of the research on image-to-image translation in this work is starting to move towards training on unpaired images [5, 16, 40] and learning from unpaired portrait and caricature [4, 35]. However, learning unpaired images can introduce highly varied exaggerations from different artists with divergent styles. Most images will have different poses and scales which might result in difficulty to distinguish facial features. In our method, we performed an unpaired learning approach using a specific caricature design of exaggerating face parts while still maintaining the facial contour of the real image.

We aim to create a method of generating new caricature faces from a real image with realistic details and obtain different stylization results. Our method first modifies a real face into a caricature face and then used that to train a generative model to produce different styles. A summary of our contribution is as follows:

- We proposed a method of generating facial caricature images with big eyes and big mouths using face patches. The method can generate different faces and can apply for multiple style transfers on a specific face.
- Our method is an unpaired learning process of creating a caricature face first from a real face image. A powerful style network is then trained using the generated caricature faces to synthesize different styles transfer.
- Our generated caricature is more realistic and high-quality as compared with the previous methods while still providing a completely disentangling style.

The remainder of our work is organized as follows: The related work in the creation of caricature and style transfer are discussed in section 2. The methodology behind the creation of our caricature and the style transfer are discussed in Section 3. Experimental results are shown and analyzed in Section 4. We finally conclude our work in Section 5.

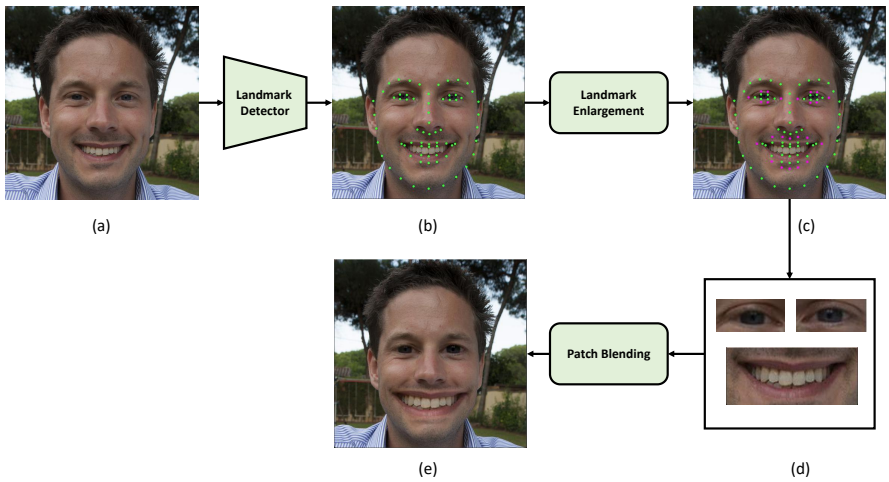

Fig. 1: Facial caricature generation pipeline: (a) input image, (b) facial landmarks using a landmark detector, (c) enlarged landmark location for eyes and mouth regions, (d) segmented and scaled eyes and mouth patches, and (d) final caricature result generated after blending the scaled face patches to the original face image.

2 Related work

In this section, we discuss some of the works related to our paper: caricature creation and style transfer.

2.1 Caricature Creation

The generation of caricature is to identify and exaggerate distinct features of a face while still maintaining the identity of the individual. The creation of caricatures can be performed in three ways: deforming facial attributes, style transfer, or methods using both.

Traditional methods perform by magnifying the deviation from the mean, either by explicitly identifying and warping landmarks [12, 25] or by utilizing data-driven approaches to estimate distinctive facial characteristics [26, 38]. With the advancement in generative networks, some image-to-image translation work [39, 23] has been done to apply transfer style. However, because these networks are unsuitable for techniques with large spatial variation, their outputs have low visual quality.

Cao et al. [4] use two CycleGANs which are trained on image and landmarks space for texture rending and geometry deformation. WarpGAN [33] can produce better visual quality and more shape exaggeration by providing flexible spatial

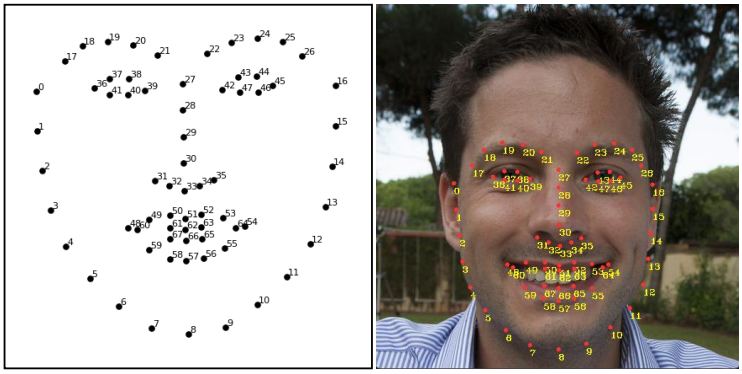

Fig. 2: Facial landmarks: (a) all 68 facial landmark annotations, and (b) test example of generating all 68 facial landmarks performed on FFHQ dataset [19].

variability on both image geometry and texture. CariGAN [4] is a GAN trained using unpaired images to learn the image-to-caricature translation. Shi et al. [33] proposed an end-to-end GAN framework that trains warping and style. Deformation fields are used by AutoToon [11] to apply the exaggerations. AutoToon is trained in a supervised manner using paired data from artist-warp photos to learn warping fields. It maps to only one domain, so it cannot produce diverse exaggerations.

2.2 Style Transfer

One type of image synthesis issue is style transfer, which seeks to create a content image with several styles. Due to the efficient ability to extract semantic features by CNNs [10], numerous style transfer networks are implemented. The initial process of rendering styles is performed by Gatys et al. [8] using hierarchical features from a VGG network [34]. The first neural style transfer approach was put out by Gatys et al. [9] and employs a CNN to transfer the style information from the style picture to the content image. The drawback is that both the style and content of pictures should be similar, which is not the case with caricatures.

A promising area of research has been the use of Generative Adversarial Networks (GANs) [13] for picture synthesis, where cutting-edge outcomes have been shown in applications like text-to-image translation [31] and image inpainting [37]. Using Generative Adversarial Networks (GANs) [13] for image synthesis has been a promising field of study, where state-of-the-art results have been demonstrated in applications ranging from text to image translation [31], image inpainting [37] and many more. Unpaired image translation is accomplished by CycleGAN [40] using a cycle consistency loss. StarGAN [5, 6] uses a single generator to learn mappings between various picture domains. To capture the

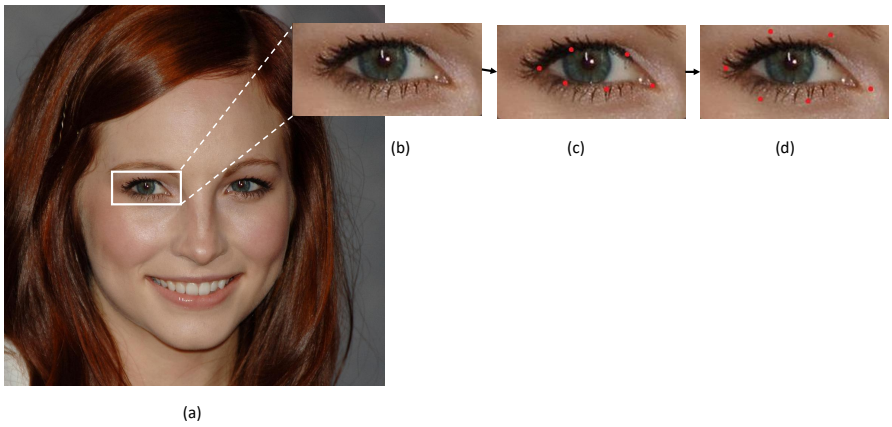

Fig. 3: Generated facial patches: (a) image from the FFHQ dataset [19], (b) visualizing the left eye from our point of view, (c) the landmark positions of the left eye, and (d) increasing the landmark indexes for improving blending results.

geometric transformation, it is challenging to learn photo-to-caricature mapping directly in an image-to-image translation approach.

StyleGAN [19, 20, 18] generates high-fidelity face images with hierarchical channel style control. StyleGAN was refined by Pinkney and Adler [30] using sparse cartoon data, and they discovered that approach was effective in producing realistic cartoon faces. DualStyleGAN [36] offers customizable management of dual styles transfer for both the expanded artistic portrait domain and the original face domain.

3 Methodology

The goal of our method is to generate a caricature face that looks realistic and train a state-of-the-art style generator with our newly generated caricature face. Our method provides completely disentangling styles for the generated caricature faces. Our whole method is sectioned into two steps: face caricature creation and face style generation. Face caricature creation focuses on the creation of exaggerated faces with enlarged eyes and mouths from real faces. Face style generation focuses on the generation of caricature faces with realistic and distinct styles.

Fig. 4: Our generated caricature images from real images.

3.1 Face Caricature Creation

Real-face images are used for the creation of a caricature face. The face images are randomly sampled from the FFHQ dataset [19] which covers diverse gender, races, ages, expressions, poses and etc. The first process is to find the facial landmark points as shown in Figure 2. The pre-trained facial landmark detector inside the dlib library [1] is used to estimate the location map of facial structures on the face. Dlib is a commonly used open-source library that can recognize 68 (x, y) coordinates of the structure of a face image. These 68 landmarks are specifically assigned for each part of the face like eyes, eyebrows, nose, mouth, and face contour.

Our implementation specifically focuses on the enlargement of the eyes and mouth region of the face. The indexes of the landmarks of the eyes can be categorized into two groups, such as the left eye and the right eye. The left eye is represented by indexes 37 to 42 whereas the right eye is by indexes 43 to 48. The mouth area consists of the upper lips and the lower lips. When we consider the mouth as a whole, we take only the top landmarks indexes of the upper lip and the bottom indexes of the bottom lips. Indexes 49 to 60 represent the mouth region. All these landmark positions are shown in Figure 2.

Using the eyes and mouth landmarks indexes, we segmented two eye regions and a mouth region patch. Before creating these patches, the corresponding landmarks are increased to make the interested area of the patches bigger as shown in Figure 3. The patches are then scaled first to a factor of 1.5 and then blend back to the original center location of the eyes and mouth part respectively. The scaled patches regions are patched back to the original image using the

Poisson image editing technique [29] which blends in seamlessly. The blending technique affects the image illumination and the texture. The Poisson seamless editing can be represented as follows:

$$v = argmin_v \sum_{i \in S, j \in N_i \cap S} ((v_i - v_j) - (s_i - s_j))^2 + \sum_{i \in S, j \in N_i \cap \neg S} ((v_i - t_j) - (s_i - t_j))^2$$

where v is the pixel values of the new image, s is the pixel values of the source image, t is the pixel values of the target image, S is the destination domain, and N_i a set of neighboring pixels of i.

Our final caricature faces are realistic as it is produced from real images. We illustrate some of our caricature datasets in Figure 4.

3.2 Style Transfer Generator

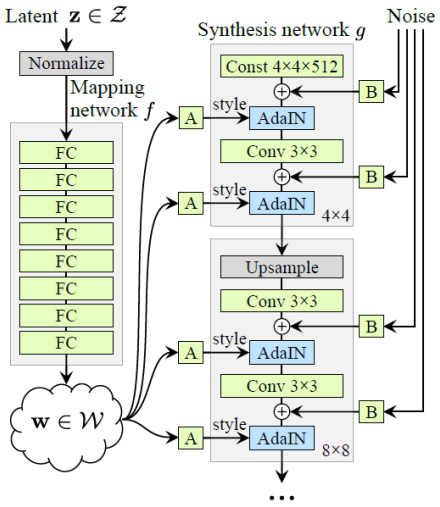

Fig. 5: StyleGAN Architecture [19, 20]

For the style transfer process, we trained a powerful style-based generator called StyleGAN [19, 20]. We trained the generator only using our newly produced face caricature images which have enlarged eye and mouth regions. The architecture of StyleGAN consists of two networks: a mapping network and a synthesis network as shown in Figure 5. A mapping network f is an 8-layer MLP that maps a given latent code $z \in Z$ to produce $w \in W$, defined as $f : Z \rightarrow W$.

Fig. 6: Four examples results of our caricature generation with style generator and each row is one caricature identity with seven different styles.

The synthesis network g are 18 convolutional layers that are controlled through adaptive instance normalization (AdaIN) [7] at each layer with the learned affine transformation "A" of latent code w. A scalable Gaussian noise input "B" is also fed in each layer of the synthesis network g.

The architecture is designed in a way that each style controls only one convolution. Random latent codes can be used to control the styles of the generated images. After we train our new caricature images, the generator can produce caricature images with different styles of facial attributes like skin tone, hair color, shapes, etc. Note that, unlike previous caricature generation, we trained our generator using only our generated caricature faces and no paired data.

3.3 Implementation

Our experiment is implemented with the diverse set of face FFHQ dataset [19]. We collected 55,000 FFHQ images for the face caricature generation and the style transfer training. The image resolution we worked on is 256 X 256. The style generator is trained with the same network architecture and other hyperparameters as the original Stylegan-ADA generator [18]. Our core algorithm is developed using PyTorch 1.7.1 [28] and CUDA 11.3. The experiment is performed using four NVIDIA TITAN Xp GPUs and a batch size of 16.

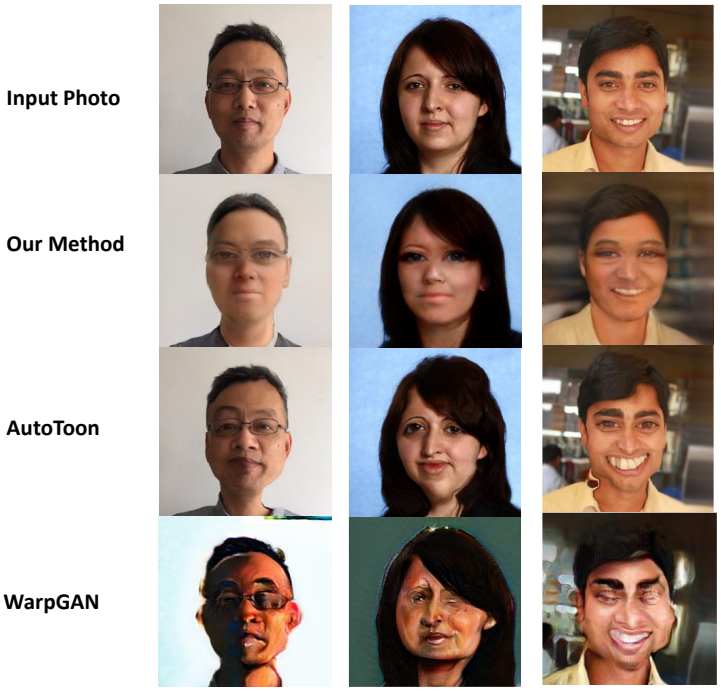

Fig. 7: Comparing our method with WarpGAN [33] and AutoToon [11].

4 Experiments

We explore various possibilities that our caricature generator can perform. Our caricature generator can generate any face type with exaggerated facial parts. Our caricature faces preserve the contour of the face shape. Face caricatures with different poses, hairstyles, face shapes, eye colors, etc can be generated. Figure 6 show different style provided for a specific caricature identity. Each row represents one identity and numerous style techniques can be applied to the generated caricature. This is possible because of the disentanglement nature of the StyleGAN generator. The latent space of StyleGAN is disentangled and we used it to our benefit. The generated images have a realistic style as we produced the caricature faces from the real images. Our caricature generator can also generate a caricature image from a real image. We demonstrate the effectiveness of our method by applying it to a different range of images gathered from publicly available content. These include images characterized by different facial expressions, poses, and illumination.

We qualitatively compare our caricature generation method with the previous caricature creation methods like AutoToon [11] and WrapGAN [33] as shown in Figure 7. We find that all three methods produced very different results. The style of WarpGAN is tightly linked to its warping, which results in irregularities

or deformation of facial features, and the quality of the caricature is degraded considerably. On the other hand, AutoToon exaggerates facial characteristics while maintaining their general quality and consistency in a way that is true to the original image, particularly with regard to specifics like the eyes, ears, and teeth. AutoToon needs paired learning method to generate these results. Our method doesn't change the face contours and exaggerates only the specific face region. The identity information and facial expression are also preserved. Our result looks more realistic as compared to other techniques, yet shows facial deformation. Since our generator is trained only on our caricature faces and not paired images, it is difficult to obtain our generator result directly from real images. We believe that we can improve our results by introducing an encoder to guide the latent space of the generator. This will be our future work.

5 Conclusion

In this paper, we proposed a framework for generating realistic unpaired caricature images. We proposed a new approach to keep the facial contours intact while exaggerating the facial parts like the eyes and mouth regions. We used a powerful style-based architecture to produce a realistic caricature from real face images. Our approach supports flexible controls to change the style of the generated caricature faces. Experimental results demonstrate that the proposed method creates caricatures that are more realistic than other state-of-the-art caricature generation methods. Although our model achieved superior results, there still exist problems that need to be tackled in caricature generation. The caricature generation is limited to all the drawbacks of the StyleGAN architecture. We will further improvements in the caricature generation process in the future.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00203, Development of 5G-based Predictive Visual Security Technology for Preemptive Threat Response) and also by the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2022-RS-2022-00156389) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

Bibliography

[1] dlib c++ library. http://dlib.net

[2] Akleman, E., Palmer, J., Logan, R.: Making extreme caricatures with a new interactive 2d deformation technique with simplicial complexes. In: Proceedings of visual. vol. 1, p. 2000. Citeseer (2000)

[3] Brennan, S.E.: Caricature generator: The dynamic exaggeration of faces by computer. Leonardo **18**(3), 170–178 (1985)

[4] Cao, K., Liao, J., Yuan, L.: Carigans: Unpaired photo-to-caricature translation. arXiv preprint arXiv:1811.00222 (2018)

[5] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)

[6] Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)

[7] Deng, Y., Tang, F., Dong, W., Sun, W., Huang, F., Xu, C.: Arbitrary style transfer via multi-adaptation network. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 2719–2727. MM '20, Association for Computing Machinery, New York, NY, USA (2020)

[8] Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. Advances in neural information processing systems **28** (2015)

[9] Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)

[10] Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E.: Controlling perceptual factors in neural style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3985–3993 (2017)

[11] Gong, J., Hold-Geoffroy, Y., Lu, J.: Autotoon: Automatic geometric warping for face cartoon generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 360–369 (2020)

[12] Gooch, B., Reinhard, E., Gooch, A.: Human facial illustrations: Creation and psychophysical evaluation. ACM Transactions on Graphics (TOG) **23**(1), 27–44 (2004)

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

[14] Han, X., Hou, K., Du, D., Qiu, Y., Cui, S., Zhou, K., Yu, Y.: Caricatureshop: Personalized and photorealistic caricature sketching. IEEE transactions on visualization and computer graphics **26**(7), 2349–2361 (2018)

[15] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. science **313**(5786), 504–507 (2006)
[16] Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
[17] Huo, J., Li, W., Shi, Y., Gao, Y., Yin, H.: Webcaricature: a benchmark for caricature recognition. arXiv preprint arXiv:1703.03230 (2017)
[18] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in Neural Information Processing Systems **33**, 12104–12114 (2020)
[19] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
[20] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
[21] Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. arXiv preprint arXiv:1907.10830 (2019)
[22] Li, C., Wand, M.: Combining markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2479–2486 (2016)
[23] Li, W., Xiong, W., Liao, H., Huo, J., Gao, Y., Luo, J.: Carigan: Caricature generation through weakly paired adversarial learning. Neural Networks **132**, 66–74 (2020)
[24] Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. arXiv preprint arXiv:1705.01088 (2017)
[25] Liao, P.Y.C.W.H., Li, T.Y.: Automatic caricature generation by analyzing facial features. In: Proceeding of 2004 Asia Conference on Computer Vision (ACCV2004), Korea. vol. 2 (2004)
[26] Liu, J., Chen, Y., Gao, W.: Mapping learning in eigenspace for harmonious caricature generation. In: Proceedings of the 14th ACM international conference on Multimedia. pp. 683–686 (2006)
[27] Mo, Z., Lewis, J.P., Neumann, U.: Improved automatic caricature by feature normalization and exaggeration. In: ACM SIGGRAPH 2004 Sketches, p. 57 (2004)
[28] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
[29] Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers, pp. 313–318 (2003)
[30] Pinkney, J.N., Adler, D.: Resolution dependent gan interpolation for controllable image synthesis between domains. arXiv preprint arXiv:2010.05334 (2020)

[31] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International conference on machine learning. pp. 1060–1069. PMLR (2016)
[32] Selim, A., Elgharib, M., Doyle, L.: Painting style transfer for head portraits using convolutional neural networks. ACM Transactions on Graphics (ToG) **35**(4), 1–18 (2016)
[33] Shi, Y., Deb, D., Jain, A.K.: Warpgan: Automatic caricature generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10762–10771 (2019)
[34] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
[35] Wu, R., Tao, X., Gu, X., Shen, X., Jia, J.: Attribute-driven spontaneous motion in unpaired image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5923–5932 (2019)
[36] Yang, S., Jiang, L., Liu, Z., Loy, C.C.: Pastiche master: Exemplar-based high-resolution portrait style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7693–7702 (2022)
[37] Yeh, R., Chen, C., Lim, T.Y., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with perceptual and contextual losses. arXiv preprint arXiv:1607.07539 **2**(3) (2016)
[38] Zhang, Y., Dong, W., Ma, C., Mei, X., Li, K., Huang, F., Hu, B.G., Deussen, O.: Data-driven synthesis of cartoon faces using different styles. IEEE Transactions on image processing **26**(1), 464–478 (2016)
[39] Zheng, Z., Wang, C., Yu, Z., Wang, N., Zheng, H., Zheng, B.: Unpaired photo-to-caricature translation on faces in the wild. Neurocomputing **355**, 71–81 (2019)
[40] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)