ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747428

HIGH-FIDELITY PORTRAIT EDITING VIA EXPLORING DIFFERENTIABLE GUIDED SKETCHES FROM THE LATENT SPACE

Chengrong Wang[⋆] *Chenjie Cao*[†] *Yanwei Fu*[†] *Xiangyang Xue*[†⋆]

[⋆] School of Computer Science, Fudan University, Shanghai, China
[†] School of Data Science, Fudan University, Shanghai, China

ABSTRACT

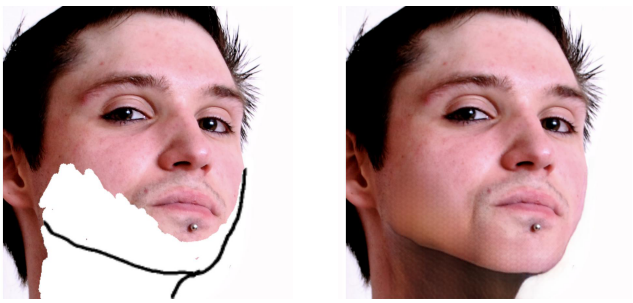
This paper studies the task of sketch-guided high-fidelity portrait editing. Advanced unconditional generators, such as StyleGAN, can generate a high-quality portrait image with great diversity. In previous researches, StyleGAN has successfully been utilized for color-guided image editing through latent vector optimization. Nonetheless, passing sketch information to the generating model directly is non-trivial. To this end, we present an algorithm that addresses the problem of well controlling the generation process via differentiable guided sketches from latent space. Specifically, we re-purpose the classic operator – eXtended difference-of-Gaussians (XDoG) that derives differentiable sketches from images. We also propose a multi-scale sketch loss assisted with which can finally guide the model follow the guidance sketch to generate. Extensive experiments validate the efficacy of our model in sketch-guided editing. We show that the quality of produced images is better than that of competitors.

1. INTRODUCTION

This paper studies the task of high-fidelity sketch-guided portrait editing. Typically, human users mask the unwanted region and add some sketches given a source input image, as in Fig. 1(a). Then the image editing algorithm should generate a new image in terms of the masks and sketches expressed by the annotators. Of course, in real-world applications, it is vital to guarantee the nice quality of edited images, which should have the photo-realistic synthesized quality with the coherent visual and texture details. However, it is not easy work to generate high-quality images conditioned on input sketches.

Previous works [1, 2, 3, 4, 5, 6, 7, 8, 9] take the compact and concise guidance to image editing: users only need to operate a few, and the network can produce the edited results automatically. In contrast, it is relatively less studied of directly utilizing sketches to guide image editing, with few exceptions of [1, 2, 4]. Critically, previous efforts are either focused on the images editing of low resolution 256×256, or

This work was supported in part by NSFC under Grant (No. 62076067), and STCSM Project (19ZR1471800).



(a) Input (b) Result by SC-FEGAN (c) Result by DeepPS (d) Result by our method

Fig. 1. (a) The input is consists of raw image, mask and guiding sketch. (b) Once zoomed in, the repeating patterns in the masked region is easily noticed. (c) The resolution is 256×256, thus it is blurrier than others. (d) The result of our method, resolution 1024×1024, without artifact.

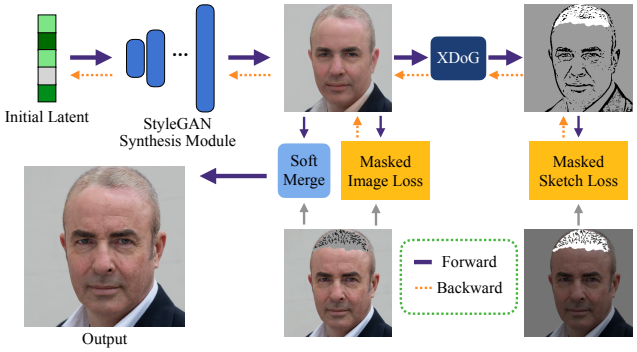


Fig. 2. Overall architecture of the proposed method. We have sketch loss and image loss for the optimization.

potentially introducing some artifact patterns [1, 4], as qualitatively compared in Fig. 1.

The generation quality also highly relies on the quality of the input sketch from the annotators. The input sketches that are quite different from those in the training set might make the model generate a low-quality image. In the other word, the 'good' sketches are not easy to be drawn by non-professional users, while the 'bad' sketches may frequently degrade the quality of edited images produced by the algorithm. Moreover, the image editing model should be robust to distorted stroke and ignore the meaningless stroke to derive high-fidelity image editing results.

To this end, we present a novel algorithm for high-fidelity image editing by addressing the problems above. Our model extends the high-fidelity unconditional face generator StyleGAN [10, 11]. To make the unconditional generator conditioning on our sketch, we propose a novel strategy of passing the difference information between the current sketch and the target sketch to the latent space of the synthesis module in StyleGAN.

To overcome the problem of sensitivity to tiny distortion in the sketches, we take the strategy of searching the generation space of the trained generator rather than directly training a mapping from sketch to its corresponding image. Empirically, we find that this strategy will produce a better quality of edited images. Furthermore, we facilitate the latent space in image editing. Notably, we compare the input sketch with the extracted edges from the image generated by the trained generator. The distance of these two sketches are passed to the latent space of the generator.

To make our latent searching algorithm feasible, our framework integrates a differentiable edge extractor and a reasonable metric for comparing the sketches. Specifically, we re-purpose a traditional edge detector, eXtended Difference-of-Gaussian (XDoG), for the differentiable edge extraction, and a multi-scale sketch loss is proposed to measure the distance between sketches and edges.

We conduct several experiments on various types of editing. The results show that our algorithm consistently produces higher editing results in most cases and is robust to strange sketch input.

Contributions. We make several contributions here. (1) The latent space in StyleGAN, has been re-purposed to help solve our sketch-guided image editing task. Such a latent space can make our algorithm robust to tiny distortion and produce consistently good images. (2) A XDoG operator is introduced to image editing. This differentiable operator enables our network updated in an end-to-end manner. (3) A multi-scale sketch loss is proposed for measuring the difference between sketches.

2. METHODOLOGY

Overview Our goal is to generate a plausible face image whose masked part is consistent with the guiding sketch; the unmasked part is consistent with the raw image. Particularly, we use the trained high-fidelity face generator StyleGAN as the backbone for the generation. We combine the latent optimization algorithm and a classical differentiable edge detection algorithm XDoG for searching the corresponding sketch from the latent space of the generator. The same searching method is employed to keep the generated result's unmasked region close to the source image. Furthermore, we especially designed a sketch loss for such a sketch-guided editing task.

2.1. Basic Modules

StyleGAN Generator StyleGAN [10, 11] has two vital components. One is a mapping network denoted as $\mathbf{G_{map}}$, which maps a random vector to the disentangled latent space. Within this disentangled space, our algorithm will learn to search the corresponding latent vector by measuring the difference between the guided sketches and the sketch corresponding to the current point in latent space. Another one is a synthesis network, represented as $\mathbf{G_{syn}}$, which projects the disentangled latent to a photo-real image. We keep the synthesis network in our generation method.

Edge Extractor Module XDoG [12] is an extended version of the Difference-of-Gaussian algorithm for edge detection. We choose the classical edge detection algorithms for their efficiency and stability. Typically, the DoG algorithm first conducts Gaussian blur on the gray-scale input image using two different standard deviations, then weighted these two blurred images and minus one blurred result from the other. The subtracted result will finally be passed to a ramp function to get the edge results. The XDoG algorithm uses $tanh$ as a ramp function, and the conv2d operator can perform Gaussian blurring with the Gaussian kernel. Therefore the conv2d and tanh operators are employed to compose a classical edge extractor. We denote the edge extractor module as $\mathbf{E}$.

Sketch Loss Module The general image loss, i.e., L1/L2 distance, can only tell the difference of each pixel. They cannot tell the difference between two strokes. Suppose we have two sketch canvases of the exact resolution. We have one line on each canvas. If the length and width of lines are fixed, then the L1 distance and L2 distance are both fixed as long as the two lines are disjoint. We propose a position-aware sketch loss which calculates the L1 distance in different scales of sketches.

2.2. The Pipeline of Image Editing

In the image editing task, a source image $\mathbf{I_{raw}}$, a binary mask $\mathbf{M}$, and a guidance sketch $\mathbf{S_{guide}}$ is given.

	SC-FEGAN	DeepPS	Ours
Score	0.207	0.241	0.552

Table 1. Average scores for user studies, which are collected from volunteers who select the best one from shuffled results.

We use the average latent $\mathbf{w}$ derived in the training of stylegan as the starting point,

$$\mathbf{I_0} = \mathbf{G_{syn}}(\mathbf{w}) \tag{1}$$

Next, our algorithm performs latent space searching to refine the generated results. We define two loss here. One is image loss, which is combine by L_1 distance and image perceptual loss $\mathcal{L}_{perc}$. We give the definition of perceptual loss, as

$$\mathcal{L}_{perc}(\mathbf{I_1}, \mathbf{I_2}) = \| \sum_{j=1}^{5} \frac{\lambda_j}{N_j} (\mathbf{F}_j(\mathbf{I_1}) - \mathbf{F}_j(\mathbf{I_2})) \|_2^2 \tag{2}$$

Where $\mathbf{F}$ is the VGG16 pretrained on ImageNet, $\mathbf{F_j}$ is the j^{th} conv layer of $\mathbf{F}$, N_j is the total number of entries in the corresponding feature map, and λ_j is the weight for the corresponding feature layer.

Another is sketch loss which is formally defined as the following:

$$\mathcal{D}_{sketch}(\mathbf{S_1}, \mathbf{S_2}) = \sum_{j} \|(\mathbf{P}_j(\mathbf{S_1}) - \mathbf{P}_j(\mathbf{S_2}))\|_1 \tag{3}$$

where $\mathbf{P_j}$ is maxpooling operation with stride j, each j is $1, 2, 4, 8$.

The final loss for us to explore in latent space searching is defined as

$$\begin{aligned} \mathcal{L} = \mathcal{D}_{sketch}(\mathbf{S_{guide}} \odot \mathbf{M}, \mathbf{E}(\mathbf{G_{syn}}(\mathbf{w})) \odot \mathbf{M}) + \\ \mathcal{L}_{perc}(\mathbf{I_{raw}} \odot (1 - \mathbf{M}), \mathbf{G_{syn}}(\mathbf{w}) \odot (1 - \mathbf{M})) + \\ \|\mathbf{I_{raw}} \odot (1 - \mathbf{M}) - \mathbf{G_{syn}}(\mathbf{w}) \odot (1 - \mathbf{M})\|_1 \end{aligned} \tag{4}$$

We use this loss to updating the latent vector $\mathbf{w}$, the generated $\mathbf{I_0}$ updates consequently, then we approach the target result iteratively and will get a converged result. We do a soft merge of the generated image and the original image according to the mask to get the final result.

3. EXPERIMENTS

Competitors and Data. We compare our method with two state-of-the-art sketch guided face editing methods SC-FEGAN [1] and Deep Plastic Surgery [4] (DeepPS). These two methods are both trained on CelebA-HQ [13]. For fair comprison, we also use the StyleGAN[11] trained on CelebA-HQ. And the testing samples are from high-fidelity face dataset FFHQ [10].

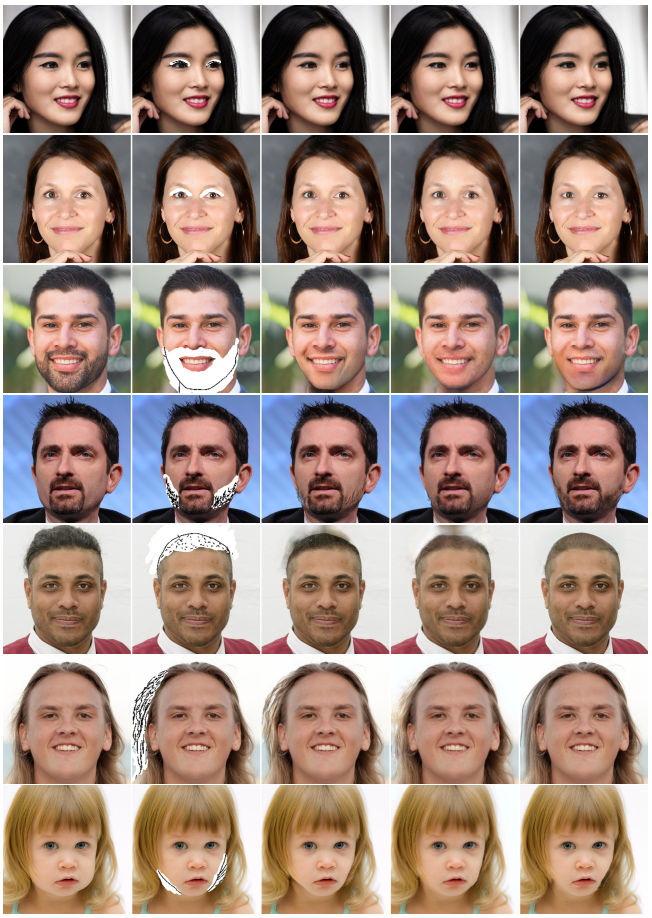


Fig. 3. The results of editing. (left to right) source image, source image with guided sketch, SC-FEGAN results, DeepPS results, ours.

We use the exact same source image, mask and guidance sketch except for resolution, for all three models. We conduct experiments on editing hair style, beard, eyes, face shape, etc. Our presenting results preserve the unmasked region of source images.

3.1. The Results of Subtle Editing

It would be of great interest to just give some subtle editing to an image. This will facilitate various useful applications. For example, one just want to edit the eyes, beard or hairs. However, editing such subtle parts would be great challenging. The proposed model should in principle better produce the results of edited images.

Eyes Editing. Double-fold eyelid editing enjoys great values in real-world problems. Therefore, we offer the results of Double-fold eyelid editing compared with other approaches in this section, as shown in Fig. 3. Specifically, an Asian girl with a single-fold eyelid is edited to a pretty one with a double-fold eyelid as shown in the first row. Our edit result

is much plausible than other methods. Differences between the three editing algorithms can be distinguished after zooming in. Particularly, our method enjoys better eyelids results with more natural and elegant style, while SC-FEGAN generates rigid results and DeepPS even fails to achieve double-fold eyelids. Besides, the double-fold eyelids can also be edited to single-fold eyelids as shown in the second row of Fig. 3. And our methods can achieve competitive results compared with other state-of-the-art methods.

Beard and Mustache Editing. Our method can remove the beard and mustache easily. We only need to mask the beard region and add some sketches for the facial shape to complete the outlines of target faces, as shown in the third row of Fig. 3. The experiments show that our method also supports adding beard as SC-FEGAN. However, the results from SC-FEGAN seem sparse and long which causes some artifacts. And the results from DeepPS fail to generate any pieces of bread. By contrast, our results are more natural and reliable. For the beard removal task, SC-FEGAN suffers from the limitation of serious checkerboard artifacts. Similar phenomena appear frequently when SC-FEGAN tries to generate clean skin on a large-scale. These checkerboard artifacts might be caused by the perceptual loss as mentioned in [14]. In contrast, benefited from the powerful generation capability of StyleGAN, our method can deal with the large-scale removal tasks easily and achieve great performance.

Hair-Style Editing. We also provide some challenging hair-style editing instances in Fig. 3. As the hair-style editing results shown in the sixth row, our method can generate more reliable images while other methods fail to generate the hair-style without artifacts. For the hair cut in the fifth row in Fig. 3, our method enjoys more stable results, while the other two competitors suffer from blurry and confused borderlines.

3.2. Face Shape Editing

Face thinning applications are very popular among smartphone users who like taking selfies. Benefits from the sketch editing, our editing algorithm is more powerful than the traditional face thinning application since users can draw the facial edge with the exact shapes. We try to make the baby to have thinner faces as shown in the last row of Fig. 3.

All of the compared methods succeed in thinning the baby's face, while SC-FEGAN and ours have better shape controlling. Specifically, our method devotes to finding latent spaces whose decoded images are close to the sketch targets after the differentiable XDoG line detecting. At the same time, our results enjoy great robustness, and generalization benefited from StyleGAN.

3.3. Multi-Scale Sketch Loss

We also validate the efficiency of the proposed multi-scale sketch loss. As shown in Fig. 4, the proposed loss outperforms the L1 loss in a variety of cases. In the first column, the L1 loss fails to generate hair; in the third column, the L1 loss fails to generate beard, while our multi-scale sketch loss succeeds in both cases. In the fourth row, the L1 loss has a blurry result, while the multi-scale loss has a sharp face outline. The second, fifth, and sixth columns show the fine controlling that our multi-scale loss can achieve. Such excellent controlling corresponds to the particular property called position-awareness–our proposed multi-scale loss can tell the difference caused by the distance between two sketch strokes. The L1/L2 loss will omit such distance since they are purely pixel-wise loss.

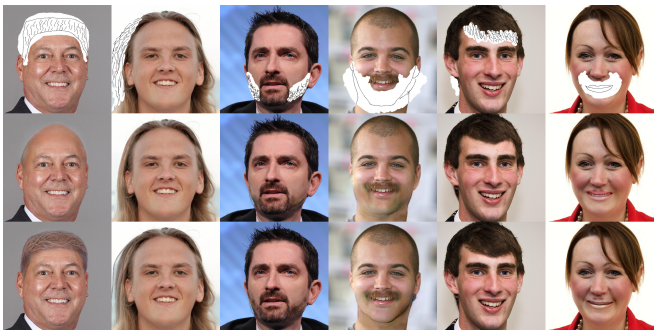


Fig. 4. Compare the vanilla L1 loss and our multi-scale loss. From up to bottom, inputs, L1 loss, multi-scale sketch loss.

3.4. User Study

Quantitative metrics are not able to evaluate the editing quality since there is no ground truth. Therefore, we provide user studies to ensure the effectiveness of our proposed method. Specifically, we invite 10 volunteers to judge the quality of the edited results. We compare the three methods mentioned above. Each volunteer is asked to select the best one from three results considering of both sketch faithfulness and image verisimilitude. As shown in Table 1, our method can achieve the best result in human perception.

4. CONCLUSION

We present a high-fidelity sketch-guided face editing algorithm, which re-purposes the high-quality image generator StyleGAN to generate great-quality images which follow the sketch guidance. We exploit the idea of searching latent and the classical efficient yet robust edge extractor to process input sketches robustly. We conduct experiments on multiple kinds of face editing and comprehensively compare our method and others. The results of experiments show that our method generates edited images of much higher quality.

5. REFERENCES

[1] Youngjoo Jo and Jongyoul Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[2] Tiziano Portenier, Qiyang Hu, Attila Szabó, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker, "Faceshop: Deep sketch-based face image editing," *ACM Trans. Graph.*, vol. 37, no. 4, July 2018.

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka, "Image2stylegan++: How to edit the embedded images?," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo, "Deep plastic surgery: Robust and controllable image editing with human-drawn sketches," in *European Conference on Computer Vision*, 2020.

[5] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen, "Cascade ef-gan: Progressive facial expression editing with local focuses," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou, "Interpreting the latent space of gans for semantic face editing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan, "Mask-guided portrait editing with conditional gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk, "Editing in style: Uncovering the local semantics of gans," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[10] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] Holger Winnemöller, Jan Eric Kyprianidis, and Sven C Olsen, "Xdog: an extended difference-of-gaussians compendium including advanced image stylization," *Computers & Graphics*, vol. 36, no. 6, pp. 740–753, 2012.

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.