Reference Guided Image Inpainting using Facial Attributes

Dongsik Yoon[1]
kevinds1106@korea.ac.kr
Jeonggi Kwak[1]
kjk8557@korea.ac.kr
Yuanming Li[1]
lym7499500@korea.ac.kr
David Han[2]
dkh42@drexel.edu
Youngsaeng Jin[1]
youngsjin@korea.ac.kr
Hanseok Ko*[1]
hsko@korea.ac.kr

[1] Korea University
Seoul, South Korea
[2] Drexel University
Philadelphia, USA

Abstract

Image inpainting is a technique of completing missing pixels such as occluded region restoration, distracting objects removal, and facial completion. Among these inpainting tasks, facial completion algorithm performs face inpainting according to the user direction. Existing approaches require delicate and well controlled input by the user, thus it is difficult for an average user to provide the guidance sufficiently accurate for the algorithm to generate desired results. To overcome this limitation, we propose an alternative user-guided inpainting architecture that manipulates facial attributes using a single reference image as the guide. Our end-to-end model consists of attribute extractors for accurate reference image attribute transfer and an inpainting model to map the attributes realistically and accurately to generated images. We customize MS-SSIM loss and learnable bidirectional attention maps in which importance structures remain intact even with irregular shaped masks. Based on our evaluation using the publicly available dataset CelebA-HQ, we demonstrate that the proposed method delivers superior performance compared to some state-of-the-art methods specialized in inpainting tasks.

1 Introduction

The recent development of Generative Adversarial Networks (GANs)[6] based image manipulation techniques led to a slew of smartphone applications in the social network scene, as users find it amusing to combine two different facial images into one or other combinations. One variant of this type of manipulation or generation is when an image is given with a mask such that the masked region needs to be filled in appropriately with feature contents

from another image. This problem is a special case of inpainting over masked regions with content information taken from another image.

Image inpainting is one of the techniques in computer vision that handles missing or damaged portions of an image by filling them in with plausible contents. As it is an old problem in computer vision, many methods have been proposed in the past. Among them, patch-based methods[1, 2, 5] complete occluded regions by applying appropriate scene segments based on contextual statistical similarities. These methods perform reasonably when the missing regions are part of backgrounds. However, the method's performance degrades significantly when the occlusions are of a part of an object such as a face. Restoring a face image with occlusion is particularly a difficult challenge for these traditional methods, as the patch used for filling in the missing region may not match well with the rest of the image. Compared to the traditional methods, GANs based inpainting are able to synthesize tractable results in comparison with non-deep learning methods, particularly in facial image restoration. With effective training, these generative models are capable of completing a plausible image while maintaining coherent consistency between the filled-in region and the rest of the image.

Unlike the traditional inpainting of letting the GANs fill in the masked region by itself, the filling process can be guided by a user. Thus, the filled-in region can be manipulated per the user guidance. There have been a good amount of research efforts in manipulating facial regions, such as adding glasses or mustache, using GANs based methods [4, 8, 9, 16, 21, 35]. These methods, however, are based on transforming a complete facial image without any missing parts. An alternate facial manipulation task is when some regions of the input facial image are masked. In such a task, inpainting is applied to those missing regions based on the user direction. Specifying these directives for useful region manipulation can be difficult. Numerous studies[13, 23, 24, 32] have been suggested to address this issue, using additional conditions such as user specified edges, colors, or landmarks. However, employing these conditions is not straightforward for an average user since accurate specifications are needed for generating user desired images.

To enable a typical user to manipulate facial images with masked regions, we propose a novel method based on using a reference image as a guide in the inpainting process. Input to our approach consists of an image with a masked region and an intact reference image. Our network first extracts attributes from the reference image so that the inpainting process will fill in the masked region according to the extracted attributes of the reference image. Thus, our architecture consists of an inpainting model and attribute extractors. Our generator adopts Learnable Bidirectional Attention Maps (LBAM)[31] in each layer to allow inpainting of any irregular shaped mask specified by the user. To ensure that the attributes of the generated image closely reflect the reference image attributes, we developed an attributes constraint loss by minimizing the difference between the extracted attributes of the reference image and the fake image. We train this architecture in an end-to-end framework by utilizing CelebA-HQ dataset with their facial attributes labels. Additionally, we adopt Multi-Scale Structure Similarity (MS-SSIM)[30] loss[20] to maintain facial structure consistency.

The main contributions of our work are as follows: (1) We propose a novel framework for an easy-to-use facial image manipulation by using a single reference image as a guide. (2) By applying random attributes, our method is capable of generating pluralistic images while maintaining reference attributes. (3) The proposed method is capable of inpainting masked regions effectively while delivering manipulated images reflecting closely to the user intentions.

2 Related Work

2.1 Traditional Inpainting

Image inpainting has been continuously studied as a method to complete missing regions. Traditional diffusion image inpainting methods used to fill only small, narrow holes such as scratches with surrounding pixels. Patch-based methods[1, 5] propagated information from the background area to the hole using patch similarity as a way to fill larger holes. Patchmatch[2], which improved the aforementioned methods, was able to synthesize more realistic textures using a fast nearest neighbor field algorithm. But these methods unable to synthesize novel objects that are not in the ground truth.

2.2 Inpainting by Deep Generative Model

Recently, learning based GANs methods used for image inpainting. CE[25] was the first inpainting model using a deep neural network, that the mask part of the center square was completed using adversarial loss and *L2* pixel-wise loss. However, this method was suffered to synthesis a plausible novel object that was not trained. GLCIC[11] used two auxiliary discriminators to solve that suffered to synthesis novel objects of existing methods. CA[33], one of the pioneer attention-based methods, trained coarse-to-fine networks for image inpainting. PConv[17] was a renowned model that improved the inpainting performance for irregular masks by using partial convolution. Partial convolution[17, 18] was effective to prevent convolution filters capture zeros when passing through the hole. More recently, LBAM[31] proposed learnable bidirectional maps that able to synthesize more realistic inpainting for irregular masks. In contrast to PConv[17], this model utilized bidirectional attention maps for re-normalization of features on U-net[26] architecture. GConv[34] proposed gated convolution that learning dynamic feature by gating mechanism for each spatial region, it further adopted sketch condition to help user modifying images.

Most image inpainting methods synthesize sole one result for each masked image, even if there more reasonable possibilities. Contrary to the existing method, PIC[37] used short+long term attention layer to produce pluralistic results. Additionally, this method addressed a probabilistic principled framework with two parallel paths called reconstructive path and generative path. Pii-GAN[3] used a novel style extractor that able to extract style features from ground truth for input into the generator. This method synthesized a variety of results coherent with the contextual semantics of the input image.

2.3 Facial Manipulate by Generative Model

There was a variety of research in manipulating facial regions, such as adding glasses or mustache, using attributes. AttGAN[8], was a novel approach to control face image by a generative model trained with facial attributes. PA-GAN[9] achieved to disentangle irrelevant attributes using a progressive attention network. While these approaches[8, 9] transform a complete facial image without any missing parts. An alternate facial manipulation task is when some regions of the input facial image are masked. EdgeConnect (EC)[23] utilized edge information for image inpainting, therefore manipulate facial features with guidance edge. LaFIn[32] transformed facial images into different face structures by utilizing landmark information. SC-FEGAN[13] proposed user-guided facial inpainting using various conditions. In this work, users inputted free-form sketch, mask, and color for user desired

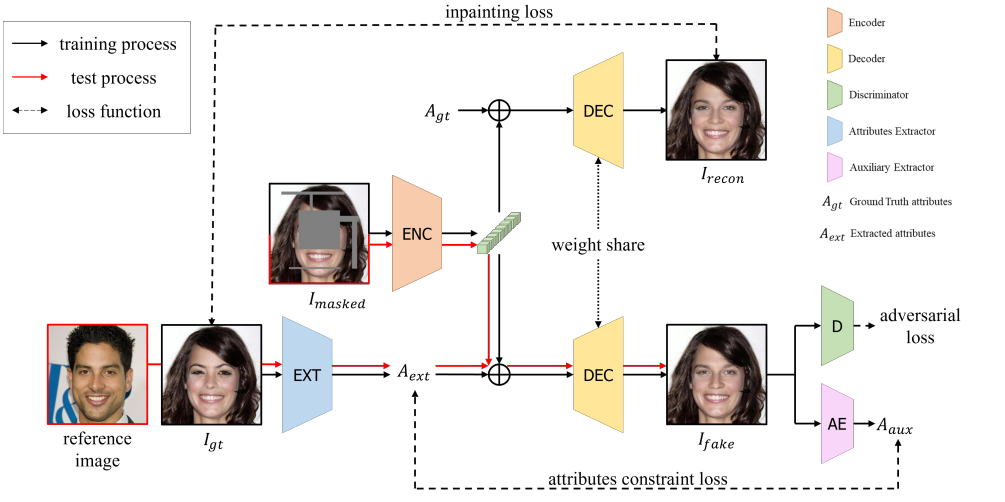

Figure 1: Summary of the proposed architecture. I_{masked} and M are the input of G, we omit M in this figure to express clearly our framework. For the test stage (red line), the user extract desired attributes using our attributes extractor to a reference image.

facial editing. These aforementioned methods[13, 23, 32] are not straightforward for an average user since accurate specifications are needed for generating user desired images.

3 Proposed Method

3.1 Model Architecture

This section introduces our user-guided inpainting architecture. As shown in Figure 1, our architecture consists of four models, generator, discriminator, attributes extractor, and auxiliary extractor. Our generator adopts LBAM[31] in each layer. Since we allow the user to modify random parts of the face, we customize bidirectional attention maps that has shown powerful performance for irregular shaped hole inpainting. Let I_{gt} be a ground truth image and its attributes be A_{gt}. We define the masked image as,

$$I_{masked} = I_{gt} \odot M, \tag{1}$$

where M is the input mask that occluded portion value as 0. We input $I_{in} = [I_{masked}, M]$ into our generator $G(\cdot)$, during the training process. We encode I_{in} into latent features and concatenate them with A_{gt} to reconstruct the original image. Thus, we decode the combined features to complete the image.

$$I_{recon} = G(I_{in}, A_{gt}) \tag{2}$$

Moreover, we create another combined feature using extracted attributes $A_{ext} = \text{Ext}(I_{gt})$, where $\text{Ext}(\cdot)$ is the attributes extractor. Then we decode another combined features for edit according to the extracted attributes.

$$I_{fake} = G(I_{in}, A_{ext}) \tag{3}$$

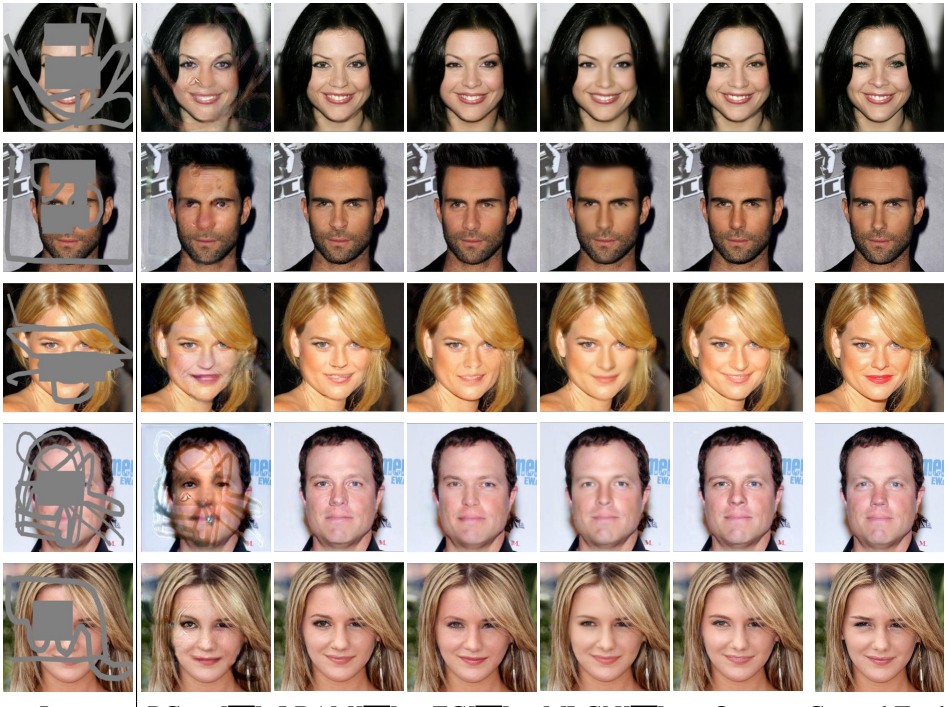

Figure 2: Qualitative comparison with PConv[17], LBAM[31], EC[23], MLGN[20], and Ours.

In the process of generating I_{recon} and I_{fake}, the decoders share their weights. We calculate only inpainting loss to I_{recon} with I_{gt}, and we encode I_{fake} into discriminator and auxiliary extractor for calculating adversarial loss and attributes constraint loss, respectively. In discriminator, we adopt Spectral Normalization[22, 34] which is fast and stable with a simple formulation in comparison to the other normalization. The attributes extractor predicts the attributes of an input facial image with VGG-16 networks[29]. It is comprised of a feature extracting layer and a fully connected layer for accurate attribute extraction. Our auxiliary extractor $\mathrm{AE}(\cdot)$ is comprised of only convolutional layers to extract attributes of fake images and learn the attributes extractor smoothly. If the attributes extractor aims to train with A_{gt} directly, it trains to depend only on fixed ground truth images. Therefore, we design to train an auxiliary extractor by using I_{gt} and its attributes before training attributes extractor with $A_{aux} = \mathrm{AE}(I_{fake})$. By doing so, it promotes the attributes extractor for delicate attribute extraction and ensures that the attributes of the fake image closely reflect the reference image attributes.

3.2 Loss function

In order to deliver manipulated images reflecting closely to the user intention while masked regions are effectively inpainted, we divided our loss into three separate losses: inpainting loss, attributes constraint loss, and adversarial loss.

Inpainting Loss. Our proposed inpainting loss is comprised of reconstruction loss, per-

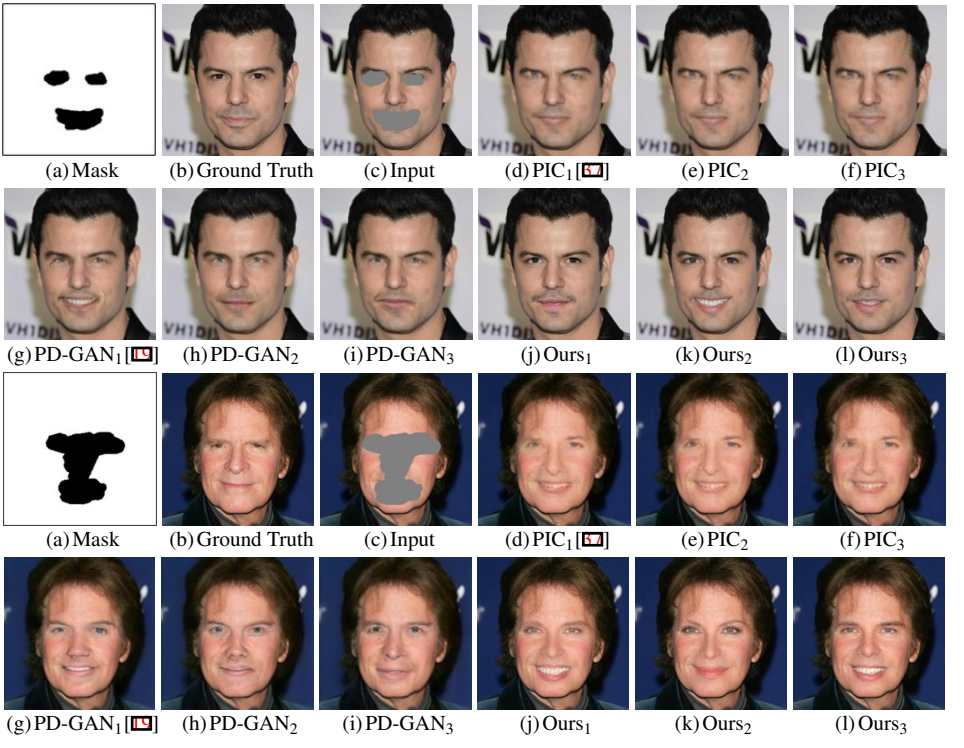

Figure 3: Pluralistic qualitative comparison with state-of-the-art methods with PIC[37], PD-GAN[19], and Ours. Images other than Ours take from the paper[19]. We excluded these images from the training process for comparison.

ceptual loss, style loss, and MS-SSIM loss, for image completion. Reconstruction loss completes erased regions using $l1$-norm error. We calculate hole region and valid region respectively by comparing the generated image with the ground truth.

$$\begin{aligned} I_{hole} &= I_{recon} \odot (1-M) \\ I_{valid} &= I_{recon} \odot M \end{aligned} \tag{4}$$

$$\begin{aligned} L_{hole} &= ||I_{hole} - I_{gt} \odot (1-M)||_1 \\ L_{valid} &= ||I_{valid} - I_{masked}||_1 \end{aligned} \tag{5}$$

Additionally, we utilize perceptual loss and style loss that most of the existing image inpainting tasks have used. We calculate these losses with the VGG-16 network pre-trained on ImageNet[27]. As the name suggests, perceptual loss measures the distance between the generated image and the feature maps of the ground truth image. It promotes capturing high-level semantics and image quality be consistent with human perception. We defined completion image as $I_{comp} = I_{masked} \oplus I_{hole}$, and defined perceptual loss as

$$L_{percep} = \frac{1}{N} \sum_{i=1}^{N} ||\phi_i(I_{gt}) - \phi_i(I_{comp})||^2, \tag{6}$$

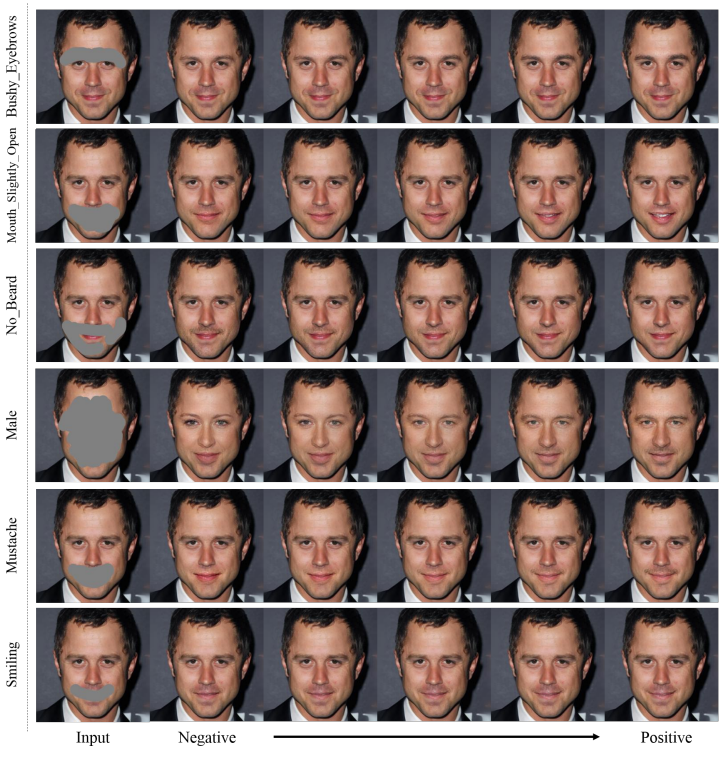

Figure 4: Illustration of attributes intensity control. From left to right are: attributes alter gradually negative to positive.

where ϕ_i is the feature maps of the $i'th$ layer of a pre-trained network and N is the number of layers in VGG-16 network. We adopt style loss, as defined by [28], which alleviates "checkerboard" artifacts due to transposed convolution layers. The style loss is defined as

$$L_{style} = \frac{1}{N}\sum_{j=1}^{N} ||G_j^{\phi}(I_{gt}) - G_j^{\phi}(I_{comp})||^2, \tag{7}$$

where G_j^{ϕ} is a gram matrix comprised from feature maps ϕ_j. Furthermore, we customize another loss using MS-SSIM[30], which is one of the methods to compare image quality. MS-SSIM loss is utilized at facial image inpainting tasks[20], and it preserves important facial structure information such as a nose or a mouth.

$$L_{\text{MS-SSIM}} = 1 - \frac{1}{N}\sum_{i=1}^{N} \text{MS-SSIM}_n \tag{8}$$

Attributes Constraint Loss. To ensure that the attributes of the generated image closely reflects the user desired attributes, we developed an attributes constraint loss. We minimize the difference between the extracted attributes of the reference image and the fake image to make the model respond more sensitively to input attributes. Moreover, we minimize the additional difference between $\text{AE}(I_{gt})$ and A_{gt} in order to train the auxiliary extractor. Unlike other works that perform facial manipulation using classifier [8, 9], our constraint

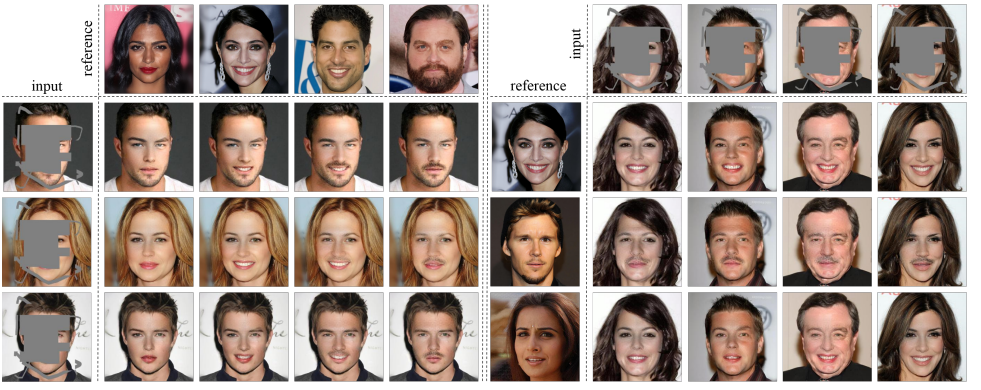

Figure 5: Overview of proposed facial image inpainting.

loss induces accurate extraction of attributes.

$$L_{attr} = \text{MSE}(A_{aux}, A_{ext}) + \text{MSE}(A_{gt}, \text{AE}(I_{gt})) \tag{9}$$

Adversarial Loss. We adopt WGAN-GP[7] that optimizes the Wasserstein distance by further employing gradient-penalty to calculate the adversarial loss. Specifically, it utilizes the Earth-Mover distance to compare synthesized and real distributions of high-dimensional data. Following this approach, our adversarial loss L_G and L_D are denoted as,

$$L_G = \mathbb{E}_{I_{in}, A_{ext}}[D(G(I_{in}, A_{ext}))], \tag{10}$$

$$L_D = \mathbb{E}_{I_{gt}}[D(I_{gt})] - \mathbb{E}_{I_{fake}}[D(I_{fake})] - \lambda_{gp}\mathbb{E}_{\hat{I}}[(||\nabla_{\hat{I}} D(\hat{I})||_2 - 1)^2]. \tag{11}$$

Overall our loss denoted as,

$$L_{all} = \lambda_{adv}L_{adv} + L_{attr} + \lambda_{ssim}L_{\text{MS-SSIM}} + \lambda_{sty}L_{style} + \lambda_{per}L_{percep} + \lambda_{hole}L_{hole} + L_{valid}, \tag{12}$$

where, λ are hyper-parameters that regulate the relative importance of the terms.

4 Experiments

4.1 Implement Details

We used the Pytorch library for implementation. Our hyper-parameters λ_{adv}, λ_{ssim}, λ_{sty}, λ_{per}, and λ_{hole} are set to 0.1, 3, 120, 0.01, and 6 respectively. These hyper-parameters were adjusted based on the qualitative performance of the empirical train process with reference studies[20, 31]. Our models were optimized using Adam optimizer[15]. We evaluate all the models in this paper using CelebA-HQ dataset[14] and utilize 28,000 selected images for training to optimize parameters and 2000 for testing. Eight attributes with visible impact are chosen in our experiments, including "Bushy_Eyebrows", "Mouth_Slightly_Open", "Big_Lips", "Male", "Mustache", "Smiling", "Wearing_Lipstick", and "No_Beard". For experiments, we trained and evaluated the proposed methods by 256×256 images with irregular holes. In addition, to adopt irregular holes on images, we utilize the Quickdraw irregular mask dataset[12] combined with 85×85 square holes at a random position. Combining square holes with Quickdraw dataset promotes the model to be more robust to irregular holes.

	Mask	Ours	LBAM	PConv	EC	MLGN
SSIM	Quickdraw	**0.833**	0.821	0.772	0.817	0.832
	10-20%	0.811	0.814	0.789	0.827	**0.839**
	20-30%	0.740	0.744	0.700	0.761	**0.777**
	30-40%	0.660	0.667	0.602	0.681	**0.706**
	40-50%	0.571	0.583	0.502	0.595	**0.624**
LPIPS	Quickdraw	**0.042**	0.047	0.088	0.047	0.063
	10-20%	0.068	0.057	0.080	**0.051**	0.065
	20-30%	0.103	0.087	0.130	**0.082**	0.103
	30-40%	0.146	**0.126**	0.214	0.128	0.148
	40-50%	0.198	**0.172**	0.309	0.183	0.201
FID	Quickdraw	**24.91**	25.79	33.69	27.49	28.45
	10-20%	28.53	27.63	31.53	**25.65**	26.73
	20-30%	35.26	36.51	61.03	**34.80**	38.34
	30-40%	**43.29**	48.47	127.2	47.14	52.54
	40-50%	**56.67**	64.40	207.3	63.75	73.07

Table 1: Quantitative comparison on CelebA-HQ. The best results of each row is boldfaced.

4.2 Qualitative Comparisons

First, we compare an image inpainting quality against four state-of-the-art methods and our baseline. Figure 2 shows images generated by the proposed method with those generated by the other methods. Images generated by PConv[17] have failed to maintain a facial structure. Our model performed better than all the others in terms of the generated image quality and plausibility. In Figure 2, we adopt the ground truth attributes to only evaluate the quality of the image inpainting task of our model. In our approach, despite the large irregular holes, facial structures are well preserved. Figure 3 compares pluralistic images generated by PIC[37], PD-GAN[19], and ours. Compared to the shown existing methods, our method accomplished more believable and diverse instances using random attributes. Figure 4 shows images generated by manipulating each attribute from the positive and to the negative extremes. As shown in Figure 4 5, our method successfully delivers plausible images according to the given mask shapes and attributes of varying degrees.

4.3 Quantitative Comparisons

We compare an image inpainting quantitatively against four existing methods and ours with different types and sizes of masks. Despite existing methods specialize in only image inpainting tasks, as shown in Table 1, our method outperforms similarly or superiorly over the three commonly used metrics SSIM, LPIPS[36], and FID[10]. We demonstrate that our method edits according to input attributes, concurrently while maintains high inpainting performance.

4.4 Ablation Study

To justify and validate the effectiveness of the proposed loss function, we conduct qualitative comparisons. Figure 6 shows the inpainting results with and without specific loss terms. By

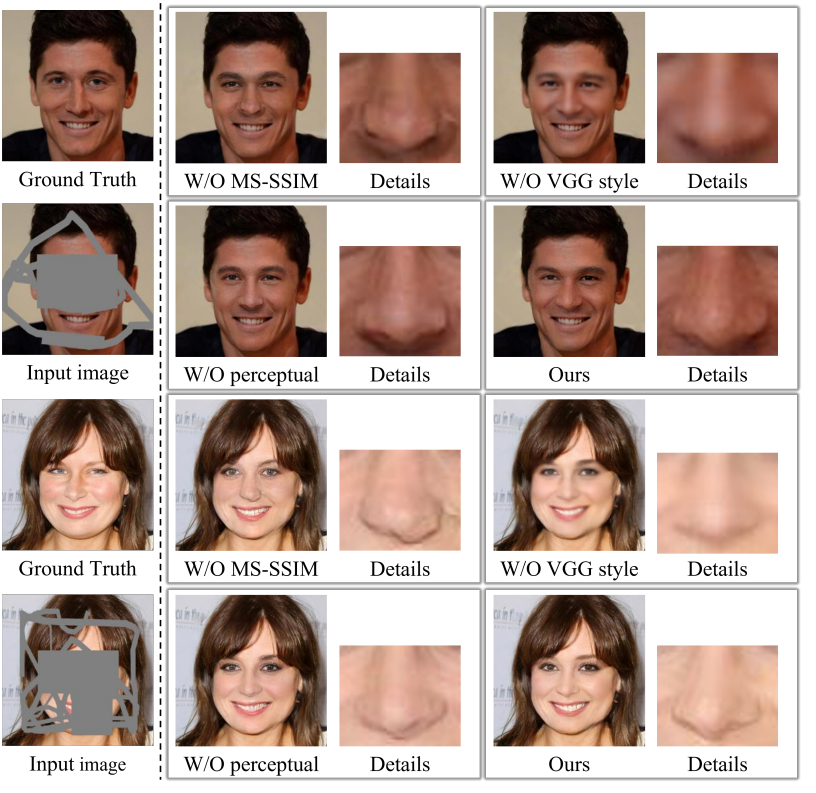

Figure 6: Qualitative comparisons with and without specific loss terms.

using perceptual loss and style loss the model learns to produce texture details in the output without any distortions or blurriness. Particularly, we demonstrate that MS-SSIM loss is a key component of our framework to maintain facial structure.

5 Conclusion

In this paper, we proposed a high-quality user-guided inpainting architecture that manipulates facial attributes of a masked image by injecting the attributes from another intact reference image. The proposed novel architecture combines LBAM with an attributes extractor to reflect the features from a reference image chosen by the user. While applying a reference image as the guide for image manipulation may not empower the user with arbitrary control, choice of the reference image is arbitrary by the user and the user guidance on image manipulation becomes much simpler. Experimental results demonstrated that our method delivers high inpainting performance while at the same time making it much easier for a user to guide the inpainting. Further, we generated multiple and diverse plausible images for a single masked input from the extracted attributes. In future work, we aim to user-guided facial inpainting for non-annotated datasets using latent features of a reference image, not facial attributes.

Acknowledgment

This research was supported by Deep Machine Lab (Q2109331).

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. IEEE transactions on image processing, 10(8):1200–1211, 2001.

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph., 28(3):24, 2009.

[3] Weiwei Cai and Zhanguo Wei. Piigan: Generative adversarial networks for pluralistic image inpainting. IEEE Access, 8:48451–48463, 2020.

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8789–8797, 2018.

[5] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In Proceedings of the seventh IEEE international conference on computer vision, volume 2, pages 1033–1038. IEEE, 1999.

[6] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014.

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028, 2017.

[8] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11):5464–5478, 2019.

[9] Zhenliang He, Meina Kan, Jichao Zhang, and Shiguang Shan. Pa-gan: Progressive attention generative adversarial network for facial attribute editing. arXiv preprint arXiv:2007.05892, 2020.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500, 2017.

[11] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics (ToG), 36(4):1–14, 2017.

[12] Karim Iskakov. Semi-parametric image inpainting. arXiv preprint arXiv:1807.02855, 2018.

[13] 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, pages 1745–1753, 2019.

[14] 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.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[16] Jeong-gi Kwak, David K Han, and Hanseok Ko. Cafe-gan: Arbitrary face attribute editing with complementary attention feature. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pages 524–540. Springer, 2020.

[17] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), pages 85–100, 2018.

[18] Guilin Liu, Kevin J Shih, Ting-Chun Wang, Fitsum A Reda, Karan Sapra, Zhiding Yu, Andrew Tao, and Bryan Catanzaro. Partial convolution based padding. arXiv preprint arXiv:1811.11718, 2018.

[19] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. arXiv preprint arXiv:2105.02201, 2021.

[20] Jie Liu and Cheolkon Jung. Facial image inpainting using multi-level generative network. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pages 1168–1173. IEEE, 2019.

[21] 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, pages 3673–3682, 2019.

[22] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957, 2018.

[23] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019.

[24] Nilesh Pandey and Andreas Savakis. Extreme face inpainting with sketch-guided conditional gan. arXiv preprint arXiv:2105.06033, 2021.

[25] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2536–2544, 2016.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115 (3):211–252, 2015.

[28] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, pages 4491–4500, 2017.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[30] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, pages 1398–1402. Ieee, 2003.

[31] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8858–8867, 2019.

[32] Yang Yang and Xiaojie Guo. Generative landmark guided face inpainting. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pages 14–26. Springer, 2020.

[33] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5505–5514, 2018.

[34] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4471–4480, 2019.

[35] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In Proceedings of the European conference on computer vision (ECCV), pages 417–432, 2018.

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

[37] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1438–1447, 2019.