SAC-GAN : Face Image Inpainting with Spatial-aware Attribute Controllable GAN

Dongmin Cha[1 [0000−0002−8327−145X]], Taehun Kim[1], Joonyeong Lee[1], and Dajin Kim[1]

Dept. of CSE, Pohang University of Science and Technology (POSTECH), Korea
{cardongmin, taehoon1018, joonyeonglee, dkim}@postech.ac.kr
https://github.com/easternCar/spatial_attribute_control_inpaint.git

Abstract. The objective of image inpainting is refilling the masked area with semantically appropriate pixels and producing visually realistic images as an output. After the introduction of generative adversarial networks (GAN), many inpainting approaches are showing promising development. Several attempts have been recently made to control reconstructed output with the desired attribute on face images using exemplar images and style vectors. Nevertheless, conventional style vector has the limitation that to project style attribute representation onto linear vector without preserving dimensional information. We introduce spatial-aware attribute controllable GAN (SAC-GAN) for face image inpainting, which is effective for reconstructing masked images with desired controllable facial attributes with advantage of utilizing style tensors as spatial forms. Various experiments to control over facial characteristics demonstrate the superiority of our method compared with previous image inpainting methods.

1 Introduction

Image inpainting is a task about image generation which has long been studied and dealt with in computer vision. Given masked images with missing regions, the main objective of image inpainting is understanding the masked images through neural networks and refilling the hole pixels with appropriate contents to produce the final reconstructed output. Image inpainting has been mainly applied to restore damaged images or refill appropriate content after removing some specific object in a photo.

Although conventional inpainting works [1,2] have demonstrated considerable reconstruction ability, image inpainting has achieved notable development since the introduction of generative adversarial networks (GAN) [3], which facilitate image synthesis ensuring plausible quality. Inpainting approaches with GAN-based methods [4,5,6] were able to show more visually realistic results than traditional approaches due to competitive training driven by adversarial networks. Recently, many efforts are focusing on user-controllable image inpainting by refilling masked region with desired contents beyond the limitation of traditional deterministic image inpainting. For example, providing an exemplar image for

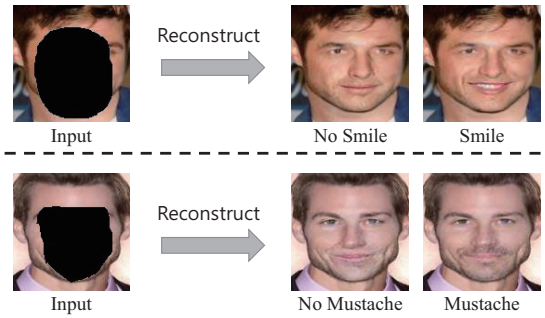

Fig. 1. Examples of face image inpainting with attribute manipulation ('smile', 'mustache') using our proposed model. We can reconstruct intended image by giving a condition, which is considered a domain of attribute.

the masked region [7], giving a guideline based on edge image [8,9], or facial landmarks [10].

After StarGAN [11] demonstrated GAN-based image translation beyond multiple domains, StyleGAN [12] showed the potential of image generation with user preference by mapping styles into features. Based on these translation methods, COMOD-GAN [13] adopted StyleGAN based style codes for controllable image inpainting with conditional input. Additionally, [14] made a trial to reconstruct images with style codes using AdaIN [15,12] or weight modulation [16].

Nevertheless, StyleMapGAN [17] pointed out that existing style-aware image generation or reconstruction projects style attributes into a linear vector with ignorance of spatial information for facial attributes. This means that facial attributes have information regarding shapes and dimensions, and projecting those style attributes onto linear vectors without preserving spatial information may reduce image generation quality. They proposed handling style codes as spatial tensors, called style maps.

Motivated by these previous works, we further propose spatial-aware attribute controllable GAN for image inpainting (SAC-GAN). Given input masked image, the target attribute is converted to spatial style map through convolutional mapping network $\mathcal{M}$. Then, we adopted a cross attention module to style maps for enhancement of contextual consistency in feature space to achieve long-range dependency between image feature and spatial style map. Finally, obtained style maps go through upsampling networks to produce multi-scale style maps, which are modulated to each layer of the decoder to reconstruct masked areas with proper contents with target attributes. To confirm the advantages of the proposed model, we conducted comparative experiments with other approaches and ablation studies with and without various loss conditions. The main contributions of this study are summarized as follows: (1) Attribute-controllable inpainting model with user-guided condition input and high-quality image reconstruction based on GAN. (2) Convolutional mapper network based on modulation which preserves dimensional information for spatial style maps with cross

attention module for global consistency between feature from masked image and spatial style maps. (3) Multi-scale spatial style maps obtained using an upsampling network are applied to the decoder using modulation to ensure a higher quality than conventional linear style vectors.

2 Related Works

2.1 Image Inpainting

Conventional image inpainting is categorized into diffusion-based and patch-based approaches. Patch-based approaches [1] usually relied on dividing image as small patches and refilling occluded areas with outer patches by computing scores, such as cosine similarity.

However, since deep-learning became a trend in computer vision, the capability of generative adversarial networks (GAN)[3] has shown remarkable performance in image generation. Many recent researches about image inpainting have been based on GAN and showed high quality in the reconstructed image [4,5,6,18]. Some approaches attempted to imitate the traditional approach of patch-based inpainting with the GAN-based model [19]. Other methods [20,21] tried to utilize binary masks to emphasize pixels to be reconstructed.

Because of the deterministic property of conventional image inpainting approaches, several researches focus on diverse image restoration to produce multiple possible predictions for damaged images. For example, considering distributions [22,23], or adopting visual transformers for the prediction of prior possibilities [24]. COMOD-GAN [13] showed visually excellent inpainting outputs by using stochastic learning and style codes [12] for large scale image restoration. It showed the possibility of applying modulation in image inpainting tasks. Inspired by attention-based [25] works like Self-Attention GAN [26], UCT-GAN [7] applied feature-cross attention map for image inpainting from two features from different images to ensure high consistency between pixels from the reconstructed image by combining two features semantically.

2.2 Facial Attribute Manipulation

StarGAN [11] demonstrated face image translation for multiple domains. The improved StarGAN v2 [27] showed the possibility of generating diverse results within one domain using the additional mapping network and feature-level style encoder. Styles for image generation are usually utilized for editing facial attributes or mixing several facial images with high quality. Unlike pixel-wise style translation [28,29], AdaIN [15] presented style editing with the generative model using instance normalization and showed remarkable possibility in manipulating a specific style. PA-GAN [30] focused on feature disentanglement using pairs of generators and discriminators for progressive image generation with attributes. StyleGAN [12] showed how a latent-space aware network facilitates multi-domain style transfer with AdaIN. StyleGAN v2 [16] proposed weight modulation which demonstrates better image generation quality than AdaIN.

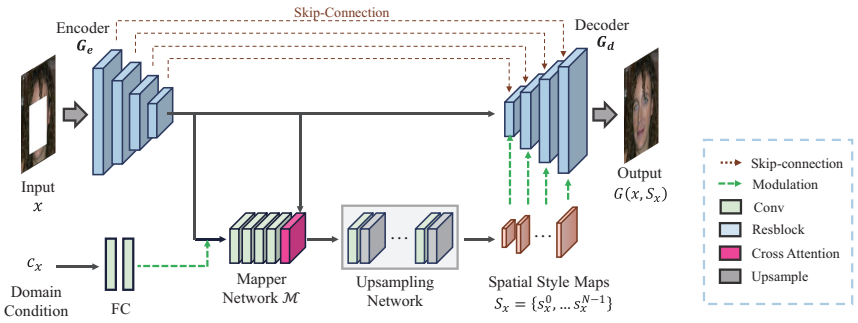

Fig. 2. Overall architecture of SAC-GAN consisting of the encoder G_e and the decoder G_d as a generative network. The mapper $\mathcal{M}$ extracts spatial style maps [17] from feature from the encoder and mapper. The style maps S_x are applied to decoder's layer by weight modulation [16].

3 Proposed Methods

As shown in Fig. 2, overall network is based on an encoder-decoder based generative architecture with weight modulation [16] to handle facial attribute features as style codes. However, beyond previous approaches dealing with styles [12,13], our model handles style codes as spatial tensors form containing spatial dimension information like StyleMapGAN [17] .

3.1 Facial Attributes as domains

Similar to L2M-GAN [31], we define a set representing a particular facial attribute as a domain, which can be considered as male/female, wearing glasses, mustache, or any other possible attributes. Then, we can consider a style code S_k for a specific domain $c_k \rightarrow \mathcal{K}$. In implementation, we consider domain $\mathcal{K}$ as a condition value c_k. Then synthesized output from input image x will be described as $G(x, S_k)$, where G denotes the encoder-decoder structure generator $G = \{G_e, G_d\}$. The proposed network for extracting a style maps S_k from condition input c_k and its detailed process are explained in the following sections.

3.2 Convolutional Mapper Network

Because of projection style information to linear vectors, some significant dimensional information could be lost during the modulation process. StyleMapGAN [17] suggests handling style code as a spatial form called style maps. With this motivation, we propose a conditional mapper based on co-modulation [13]. Instead of linear vector and fully-connected layers, our mapper network $\mathcal{M}$ produces a spatial style map s_k as output.

Additionally, in mapper network $\mathcal{M}$, to ensure dependency between far pixels in style maps and features from masked input image x, we adopted a cross attention module in similar way as [26,7]. Cross attention module generates attention

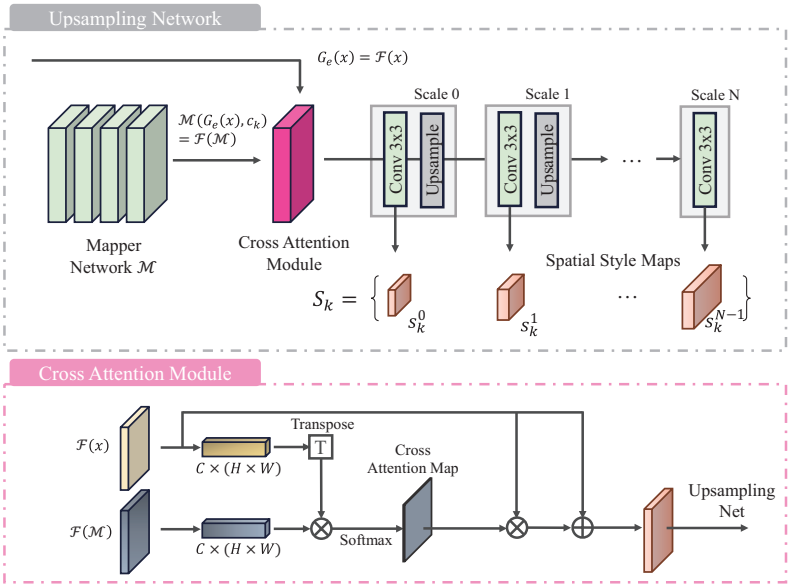

Fig. 3. Illustrations of conditional mapper network $\mathcal{M}$, upsampling layers, and cross attention module. The initial style map is generated from $\mathcal{M}$ and cross attention module gradually upsampled to produce pyramid-like spatial style maps [17].

map $A \in \mathcal{R}^{H \times W}$ from the encoder's feature map $\mathcal{F}(x) = G_e(x) \in \mathcal{R}^{C \times H \times W}$ and output of mapper network $\mathcal{M}(G_e(x), c_k) \in \mathcal{R}^{C \times H \times W}$. At first they are reshaped through 1×1 convolution and the cross attention module produces the initial style map s_k as:

$$A = \sigma(tr(\mathcal{F}(x)) \cdot \mathcal{F}(\mathcal{M})), \qquad s_k = (A \cdot \mathcal{F}(x)) \oplus \mathcal{F}(x) \tag{1}$$

Then, the refined style map s_k is sent to the upsampling network to produce style maps with various scales. Where input masked image x is given, style maps for a specific attribute k are denoted as:

$$S_k = \{s_k^0, \ldots, s_k^{N-1}\} = \mathcal{M}(G_e(x), c_k) \tag{2}$$

Where c_k is condition value for attribute domain $\mathcal{K}$ and $G_e(x)$ is the output feature from encoder G_e from input masked image x. The pyramid-like set S_k includes style maps with various scales. As shown in Fig. 3, the condition value c_k is converted to a one-hot vector and concatenated to the feature $G_e(x)$. The condition value is converted to a one-hot vector and passed through the mapping network. The mapped vector is reshaped to the same size as the feature $G_e(x)$ and the concatenation of the vector and $G_e(x)$ is passed through spatial mapping layers.

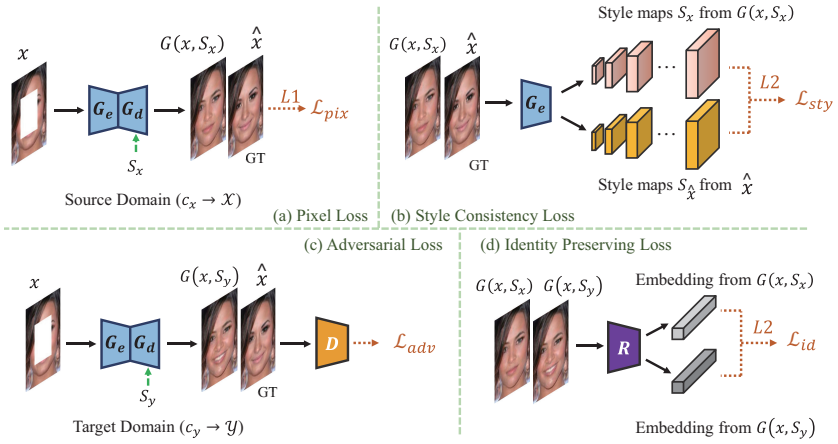

Fig. 4. Overview of our four loss functions for training our proposed models. First pixel loss ($\mathcal{L}_{pix}$) and style consistency loss ($\mathcal{L}_{sty}$) are computed from $G(x, S_x)$ (source domain $\mathcal{X}$). Then adversarial loss ($\mathcal{L}_{adv}$) and identity preserving loss ($\mathcal{L}_{id}$) are computed from $G(x, S_y)$ (target domain $\mathcal{Y}$).

3.3 Generator with Co-Modulation

The encoder G_e takes images as input to represent the image as a feature map. We denoted masks pixels by concatenating binary masks to the input image. The encoder output feature $G_e(x)$ obtained from input image x and mask is reconstructed in the decoder G_d containing skip-connection [32] and style modulation [16] from the set of style codes S_i. The modulated convolutional from each decoder layer applies the intended style (S_i) to the feature map during the reconstruction process.

3.4 Training Objectives

Unlike previous style-aware image translation methods that use complete images as inputs, we have to perform style modulation and image reconstruction simultaneously from masked input images that lack visual information. To achieve these objectives, we reconstruct input image x with style maps S_x from domain attribute $c_x \rightarrow \mathcal{X}$ which is in the same domain as input image $x \in \mathcal{X}$. After optimizing pixel reconstruction loss and style consistency loss in source-to-source inpainting, then we optimize adversarial loss and face identity loss in source-to-target inpainting. The training scheme is operated by four main loss functions including pixel-wise loss ($\mathcal{L}_{pix}$), adversarial loss ($\mathcal{L}_{adv}$), style consistency loss ($\mathcal{L}_{sty}$), and identity preserving loss ($\mathcal{L}_{id}$).

Pixel-wise Reconstruction Loss. Basically, a generative model for reconstructing images requires a pixel-wise loss to refill approximate content in occluded regions. Because L2-norm loss has a drawback of making the reconstructed image blurry, L1 loss is adopted. Our generator G consists of the encoder

Algorithm 1 Training Procedure

Prepare dataset for domain $\mathcal{X}$ and domain $\mathcal{Y}$
Fix recognizer encoder R
while G_e, G_d, D, $\mathcal{M}$ is not converged **do**
    Sample batch $\hat{x}$ and its domain condition $c_x \rightarrow \mathcal{X}$ from dataset
    Generate masks m for $\hat{x}$ and construct input x by $x \leftarrow \hat{x} \odot M$
Phase 1 : Reconstruct with source domain $\mathcal{X}$
    Extract feature $G_e(x)$ from G_e
    Extract S_x from $\mathcal{M}$ with $G_e(x)$ and condition c_x
    Reconstruct outputs $G(x, S_x) \leftarrow G_e(x)$ from G_d
    Compute losses $\mathcal{L}_{re}$ and $\mathcal{L}_{sty}$
    Update G_e, G_d and $\mathcal{M}$ with $\mathcal{L}_{re}$ and $\mathcal{L}_{sty}$
Phase 2 : Reconstruct with target domain $\mathcal{Y}$
    Pick target domain condition $c_y \rightarrow \mathcal{Y}$
    Extract S_y from $\mathcal{M}$ with $G_e(x)$ and condition c_y
    Reconstruct outputs $G(x, S_y) \leftarrow G_e(x)$ from G_d
    Compute losses $\mathcal{L}_{adv}$ and $\mathcal{L}_{id}$
    Update G_e, G_d and $\mathcal{M}$ with $\mathcal{L}_{id}$ and $\mathcal{L}_{adv}$
    Update D with $\mathcal{L}_{adv}$
end while

and the decoder, which refills missing area from input image x with appropriate pixels to produce reconstructed output. In the training phase, the pixel loss is computed from reconstructed image $G(x, S_x)$ from style maps S_x with the same domain $c_x \rightarrow \mathcal{X}$ as the input image x and ground-truth $\hat{x}$.

$$\mathcal{L}_{pix} = |G(x, S_x) - \hat{x}|_1 \tag{3}$$

Where S_x denotes style maps obtained from mapper with source domain condition c_x, $G(x, S_x)$ denotes the reconstructed output from x using S_x with the decoder's weight modulation.

Adversarial Loss. The synthesized image should be realistic enough to be comparable to the original image. To achieve realistic reconstruction output beyond multiple attributes, we employed a multi-domain discriminator [31,33] based on Wasserstein GAN [34]. We also applied the R1-regularization gradient penalty [35] to the adversarial loss for the discriminator's stable training with high convergence. The adversarial loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv} = E_{\hat{x}}[-\log D(\hat{x}, c_x)] + E_{\hat{x}, c_y}[\log(1 - D(G(x, S_y), c_y))] \\ + E_{\hat{x}}[\bigtriangledown D(\hat{x}, c_x)],\end{aligned}$$

Where S_y indicates style maps from target domain $c_y \rightarrow \mathcal{Y}$, $D(G(x, S_y), c_y)$ is the output of the discriminator for the fake image with target domain c_y; and $D(\hat{x}, c_x)$ denotes the real image description with source domain c_x, which is the same as the domain of ground-truth image $\hat{x}$.

Style Consistency Loss. We define style consistency loss [36,29] in source domain c_x to guarantee that the encoder G_e and mapper $\mathcal{M}$ extract identical style maps from reconstructed image $G(x, S_x)$ and ground-truth $\hat{x}$.

$$\mathcal{L}_{sty} = \sum_{i}^{N-1} (|s_{\hat{x}}^{i} - s_{x}^{i}|_2) \tag{4}$$

Where $S_x = \{s_x^0, s_x^1, ..., s_x^{N-1}\}$ are N style maps obtained from the encoded feature of the reconstructed image $G(x, S_x)$ and mapper $\mathcal{M}$, $S_{\hat{x}} = \{s_{\hat{x}}^0, s_{\hat{x}}^1, ..., s_{\hat{x}}^{N-1}\}$ denotes style maps from the feature of ground-truth image $\hat{x}$ and mapper $\mathcal{M}$ so that the style consistency loss can be computed from the summation of L-2 distances between S_x and $S_{\hat{x}}$.

Identity Preserving Loss. After computing the above three losses, we have reconstructed images with source attribute domain $\mathcal{X}$ and target attribute domain $\mathcal{Y}$. From those images we introduced identity preserving loss, which guarantees that reconstructed outputs from an image x preserve the same identity. Similar to [37], we adopted the face identity recognizer network ψ, which is the pre-trained ArcFace [38] with CASIA-WebFace [39]. The identity loss for our model is defined as:

$$\mathcal{L}_{id} = |\psi(G(x, S_x)) - \psi(G(x, S_y))|_2 \tag{5}$$

Full Objective. Finally, the total objective for training our proposed SAC-GAN can be described as a combination of aforementioned losses:

$$\mathcal{L}_{total} = \lambda_{pix} \cdot \mathcal{L}_{pix} + \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{sty} \cdot \mathcal{L}_{sty} + \lambda_{id} \cdot \mathcal{L}_{id} \tag{6}$$

Where $\{\lambda_{pix}, \lambda_{adv}, \lambda_{sty}$ and $\lambda_{id}\}$ denote hyper-parameters for controlling the importance of each component. Experimentally, we conducted training and test with the hyper-parameters conditions of $\lambda_{pix} = 100$, $\lambda_{adv} = 1$, $\lambda_{sty} = 1$ and $\lambda_{id} = 0.1$. Detailed training strategy is shown in Algorithm 1.

4 Experiments

4.1 Implementation and Datasets

Datasets Aligned face datasets have the possibility that they basically have spatial information because their significant facial components are fixed. In order to confirm that our spatial-aware method has an effect even in the unaligned face images, we prepared wild CelebA dataset [41] using only face detection based on MTCNN [42]. We cropped the face images with 25 margins and 128×128 size. CelebA contains more than 200,000 face images including various facial attributes, which are mainly used for quantitative evaluation for restoring facial images in image inpainting. For quantitative experiments, we mainly used 'smiling' attribute class data because the amount of data in each class is well balanced compared to other attributes.

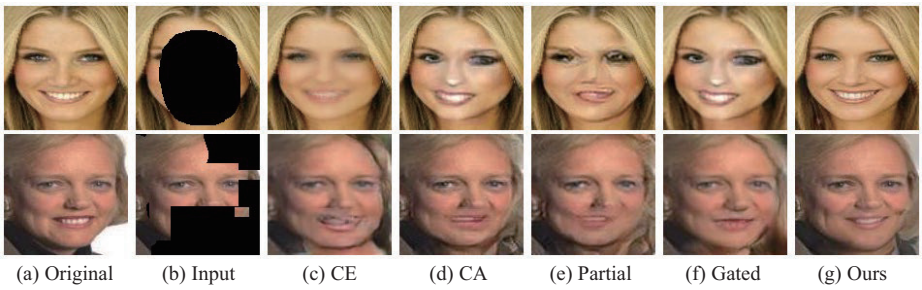

Fig. 5. Comparison of inpainting result with CelebA dataset. (a) Ground-truth. (b) Input. (c) context encoders (CE) [4]. (d) contextual attention (CA) [19]. (e) partial convolution [20]. (f) gated convolution [21]. (g) Ours.

Implementation Details. Our network is implemented in PyTorch and trained for 100,000 iterations using Adam optimizer [43] with a batch size of 20. Training and experiments are conducted on the NVIDIA TITAN RTX GPU. For occlusion masks, we adopted two strategies: fixed masks and irregular masks. The fixed masks cover all main facial components, including the eyes, nose, and mouth. The irregular masks are same used in gated convolution (free-form) [21], which creates random brushes, circles, and rectangles.

In order to compare the quantitative performance with other methods, we employed the facial attribute classifier based on ResNet-50 [44] to compare the visual certainty of specific attributes from the reconstructed image and FID [45] to measure the quality of generated images using Inception-V3 [46] network. We exploit several facial attributes for qualitative experiments, including 'smile', 'gender', 'glasses' and 'mustache' which are visually evident, and location information was expected to be important because those attributes tend to appear in certain areas of the face. For each attribute, we conducted training and test as two-class domains and used 90% of images for training and 10% for evaluation.

For comparative experiments with previous other inpainting models, we conducted two types of comparisons. 1) Comparing our model's inpainting performance with previous image inpainting models like context encoders (CE) [4], contextual attention (CA) [19], partial convolution [20], and gated convolution [21]. To compare our SAC-GAN with these models, we reconstructed the masked image by giving the same domain condition value as ground-truth. 2) Comparing with other models with conditional-based inpainting. In this experiment, we consider the condition as attribute domains such as wearing glasses, smiling and gender. Besides COMOD-GAN [13], we combined CGAN [40] and CE [4] as a baseline inpainting model for facial attribute manipulation without modulation. We denoted conditional CE as 'C-CE' which takes additional input condition value and reshapes it to tensor in a similar way to CGAN [40].

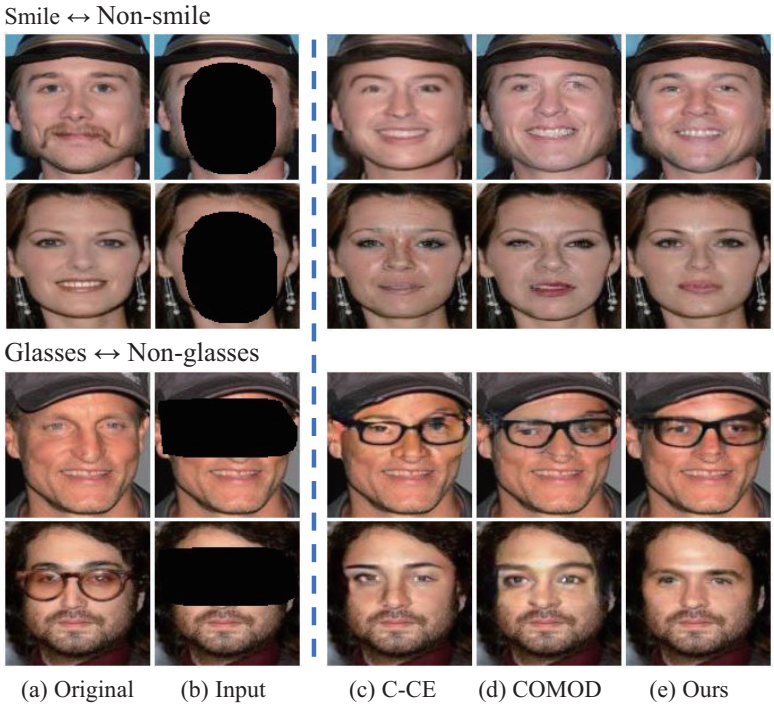

Fig. 6. Examples of attribute controllable face inpainting with given domain attribute condition. From left to right are: (a) Ground-truth. (b) Input masked image. (c) Conditional CE [40,4]. (d) COMOD-GAN [13]. (e) Ours.

4.2 Quantitative Comparisons

As mentioned above, we evaluated the inpainting task of our proposed model by providing the same condition attribute with the ground-truth of the input masked image. For example, we split test set into 'smile' and 'non-smile' groups and evaluated the average accuracy in 'smile → smile' and 'non-smile → non-smile' reconstruction. Table 1 presents the quantitative comparative results for the inpainting task with various generative models.

Next, we conducted an inpainting test with facial attribute manipulation, considering an attribute as a domain. Because facial attribute translation is a hard task to evaluate using only with visual metrics, we deployed the ResNet-50 based facial attribute classifier to verify numerically that the attributes we expected were applied well during the reconstruction process. Fig. 7 shows the two-class classification for specific attributes, including 'smile' and 'eye-glasses'. Our model produced higher performance and accuracy than other methods. Results of quantitative comparison in attribute manipulating experiments are shown in Table 2. In this table, we reported the results of test for 'non-smile → smile' inpainting task. Acc denotes accuracy of attribute classification about two classes: smile and non-smile.

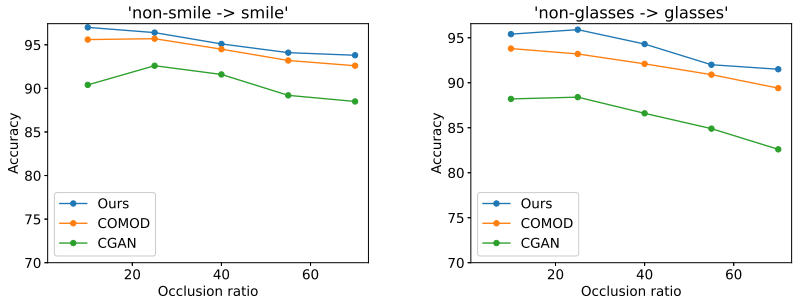

Fig. 7. Accuracy of facial attribute classification about smiling(left) and eyeglasses(right) from reconstructed images using various facial manipulation inpainting methods. Input images are occluded by irregular masks and the x-axis denotes proportion of occluded pixels from mask.

	Fixed Mask			Free Mask		
	FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑
CE [4]	13.83	19.46	0.641	11.86	20.09	0.759
CA [19]	10.35	21.87	0.694	9.13	21.87	0.803
Partial [20]	7.98	21.02	0.710	6.05	22.54	0.786
Gated [21]	6.43	22.65	0.748	4.23	23.09	0.801
Ours	**4.55**	**23.39**	**0.769**	**3.18**	**24.68**	**0.814**

Table 1. Quantitative comparison on CelebA with inpainting task for 'non-smile → non-smile' and 'smile → smile' with fixed and free form masks. The highest performances are marked in bold.

	Fixed Mask				Free Mask			
	FID↓	PSNR↑	SSIM↑	Acc↑	FID↓	PSNR↑	SSIM↑	Acc↑
C-CE [40,4]	21.49	19.28	0.621	88.54%	13.49	0.784	21.92	91.15%
COMOD [13]	10.28	20.65	0.665	92.60%	8.18	0.811	22.57	93.28%
Ours	**10.04**	**21.30**	**0.701**	**93.82%**	**7.70**	**0.823**	**22.84**	**95.28%**

Table 2. Quantitative comparison on CelebA with attribute manipulation inpainting for 'non-smile → smile' with fixed and free form masks. The highest performances are marked in bold.

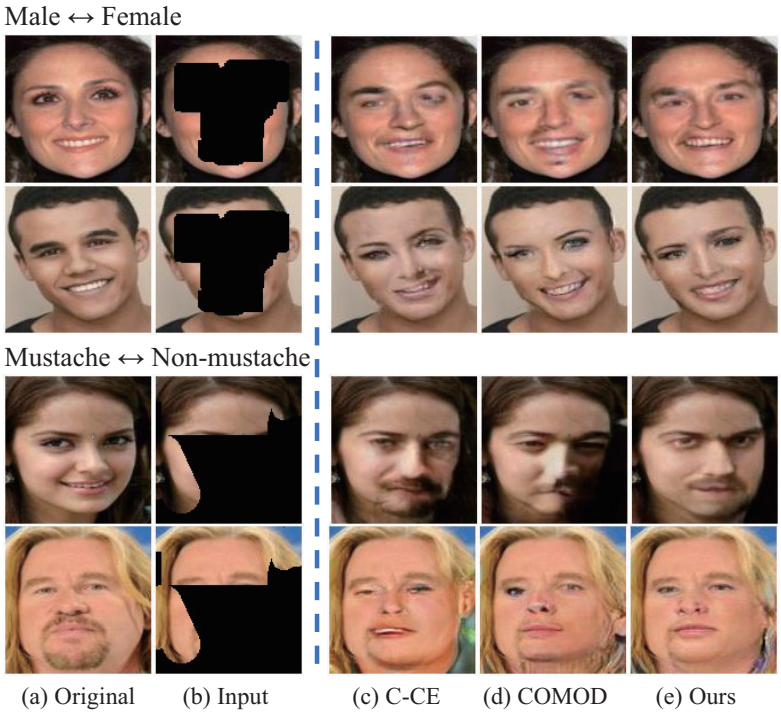

Fig. 8. Examples of attribute controllable face inpainting with given domain attribute condition. From left to right are: (a) Ground-truth. (b) Input masked image. (c) Conditional CE [40,4]. (d) COMOD-GAN [13]. (e) Ours.

4.3 Qualitative Comparisons

We also conducted qualitative comparisons with the same condition as experiments for quantitative results. Reconstructed images from inpainting task are shown in Fig. 5. Although other methods produce slightly distorted outputs, applying weight modulation synthesized output images with better quality visually.

In Fig. 6, we present results of image inpainting with face attribute manipulation. From top to bottom: 'non-smile → smile', 'smile → non-smile', 'non-glasses → glasses', and 'glasses → non-glasses'. Our model synthesized visually natural output by filling masked areas with intended condition. Furthermore, our SAC-GAN generated higher quality output with spatial style maps compared to previous linear-based style modulation like COMOD-GAN [13]. Additionally, we presented another results with 'mustache' and 'gender' attributes in Fig. 8.

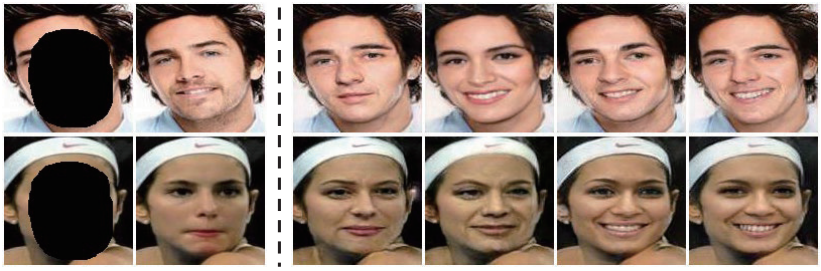

Fig. 9. Ablation study results for our model on the various losses and cross attention in inpainting task of 'non-smile → smile'. (a) Input image. (b) Ground-truth. (c) Using $\mathcal{L}_{pix} + \mathcal{L}_{adv}$. (d) Using $\mathcal{L}_{pix} + \mathcal{L}_{adv} + \mathcal{L}_{sty}$. (e) Using $\mathcal{L}_{pix} + \mathcal{L}_{adv} + \mathcal{L}_{sty} + \mathcal{L}_{id}$ and (f) Using all losses and cross attention module.

$\mathcal{L}_{pix}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{sty}$	$\mathcal{L}_{id}$	FID↓	PSNR↑	SSIM↑	Acc↑
✓	✓			14.27	20.68	0.681	90.29%
✓	✓	✓		10.81	20.48	0.663	92.10%
✓	✓	✓	✓	10.28	20.86	0.685	93.01%
All losses + Cross Attention				**10.04**	**21.30**	**0.701**	**93.82%**

Table 3. Quantitative comparison for the ablation study with various losses on CelebA with fixed mask and facial attribute 'non-smile → smile'.

4.4 Ablation Study

We trained our model on the auxiliary losses or module to check the effect of our loss terms. We conducted our ablation study with fixed mask and various loss conditions in 'non-smile → smile' reconstruction task. As shown in Table 3. it demonstrates preserving identity loss term produce more stable facial attribute-aware inpainting by maintaining overall identity information beyond domain conditions. The result in the bottom achieves better output, showing the benefits of the cross attention module. The visual examples are shown in Fig. 9. We can check that our cross attention improves the quality of synthesized images in visual metrics and attribute accuracy. Since style consistency loss is excluded in model (c), the 'smile' attribute was not well applied to the reconstructed output. In (d), although the expected attribute was well applied to output, it showed a limitation that the identity is not maintained.

5 Conclusions

We presented SAC-GAN : Spatial-aware attribute controllable GAN for image inpainting in this paper. Our network is able to restore masked image with appropriate contents and intended attribute by using style tensors preserving

spatial dimension instead of conventional linear style vectors. Through extensive experiments, we demonstrated that our proposed SAC-GAN generating visually remarkable outputs using spatial style maps. Additionally, our proposed cross attention module achieved the advantage of long-range dependency between feature from image and style map, which enhanced the performance of style-aware image inpainting. Moving forward, we expect our proposed model to reconstruct images with high quality using more complex and extensive facial attributes.

Acknowledgements This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2017-0-00897, Development of Object Detection and Recognition for Intelligent Vehicles) and (No.B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis)

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28** (2009) 24
2. Wilczkowiak, M., Brostow, G.J., Tordoff, B., Cipolla, R.: Hole filling through photomontage. In: BMVC 2005-Proceedings of the British Machine Vision Conference 2005. (2005)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
4. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2536–2544
5. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (ToG) **36** (2017) 1–14
6. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 3911–3919
7. Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D.: Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 5741–5750
8. Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 1745–1753
9. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. (2019) 0–0
10. Yang, Y., Guo, X., Ma, J., Ma, L., Ling, H.: Lafin: Generative landmark guided face inpainting. arXiv preprint arXiv:1911.11394 (2019)
11. 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. (2018) 8789–8797

12. 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. (2019) 4401–4410
13. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021)
14. Lu, W., Zhao, H., Jiang, X., Jin, X., Wang, M., Lyu, J., Shi, K.: Diverse facial inpainting guided by exemplars. arXiv preprint arXiv:2202.06358 (2022)
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. (2017) 1501–1510
16. 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. (2020) 8110–8119
17. Kim, H., Choi, Y., Kim, J., Yoo, S., Uh, Y.: Exploiting spatial dimensions of latent in gan for real-time image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 852–861
18. Wang, Q., Fan, H., Sun, G., Cong, Y., Tang, Y.: Laplacian pyramid adversarial network for face completion. Pattern Recognition **88** (2019) 493–505
19. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 5505–5514
20. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European conference on computer vision (ECCV). (2018) 85–100
21. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 4471–4480
22. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 1438–1447
23. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-gan: Probabilistic diverse gan for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 9371–9381
24. Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., Miao, C.: Diverse image inpainting with bidirectional and autoregressive transformers. In: Proceedings of the 29th ACM International Conference on Multimedia. (2021) 69–78
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
26. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning, PMLR (2019) 7354–7363
27. 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. (2020) 8188–8197
28. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1125–1134

29. 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. (2017) 2223–2232
30. He, Z., Kan, M., Zhang, J., Shan, S.: Pa-gan: Progressive attention generative adversarial network for facial attribute editing. arXiv preprint arXiv:2007.05892 (2020)
31. Yang, G., Fei, N., Ding, M., Liu, G., Lu, Z., Xiang, T.: L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 2951–2960
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
33. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 10551–10560
34. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning, PMLR (2017) 214–223
35. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning, PMLR (2018) 3481–3490
36. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3d-guided cycle consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 117–126
37. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 2287–2296
38. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019) 4690–4699
39. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
40. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
41. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. (2015) 3730–3738
42. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters **23** (2016) 1499–1503
43. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
44. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
45. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

46. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2818–2826