Not just Compete, but Collaborate: Local Image-to-Image Translation via Cooperative Mask Prediction

Daejin Kim[1] Mohammad Azam Khan[2] Jaegul Choo[1]
[1]KAIST [2]Dhaka Power Distribution Company Ltd.
{kiddj,jchoo}@kaist.ac.kr, azam@dpdc.org.bd

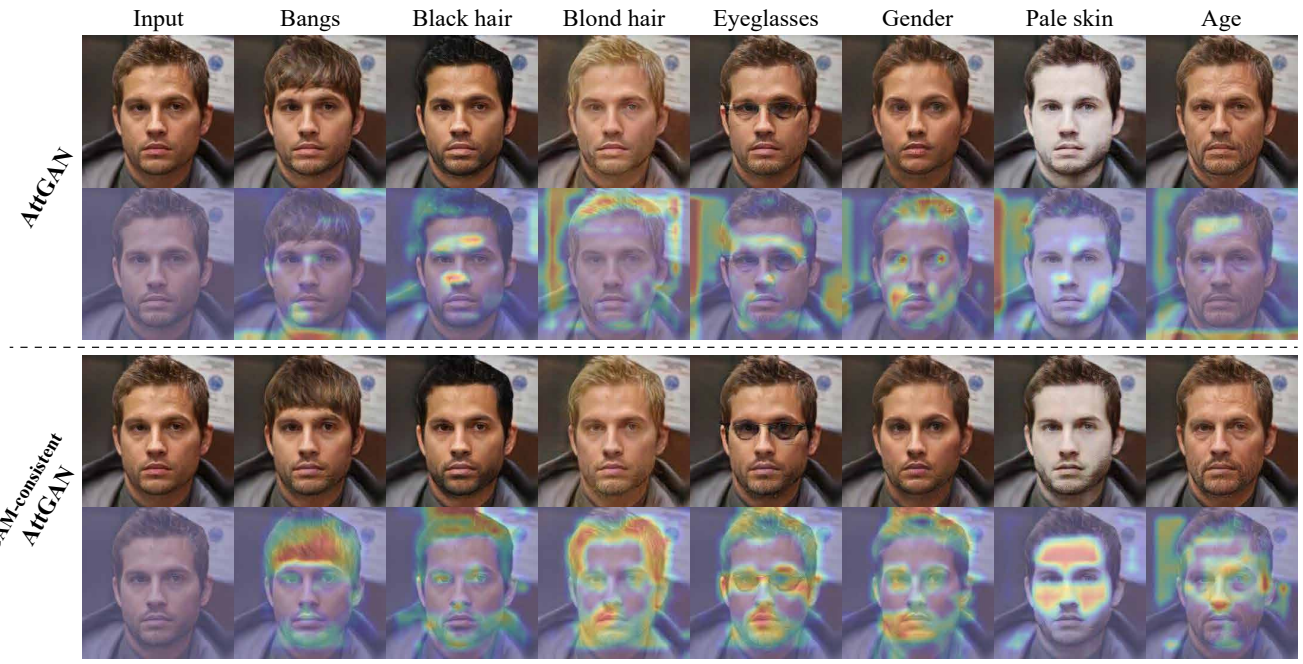


Figure 1: Results of facial attribute editing (256 × 256), and the Grad-CAM for each attribute. The first two rows are the results from the original AttGAN while the latter two are those from the AttGAN trained with our proposed CAM-consistency loss.

Abstract

Facial attribute editing aims to manipulate the image with the desired attribute while preserving the other details. Recently, generative adversarial networks along with the encoder-decoder architecture have been utilized for this task owing to their ability to create realistic images. However, the existing methods for the unpaired dataset cannot still preserve the attribute-irrelevant regions properly due to the absence of the ground truth image. This work proposes a novel, intuitive loss function called the CAM-consistency loss, which improves the consistency of an input image in image translation. While the existing cycle-consistency loss ensures that the image can be translated back, our approach makes the model further preserve the attribute-irrelevant regions even in a single translation to another domain by using the Grad-CAM output computed from the discriminator. Our CAM-consistency loss directly optimizes such a Grad-CAM output from the discriminator during training, in order to properly capture which local regions the generator should change while keeping the other regions unchanged. In this manner, our approach allows the generator and the discriminator to collaborate with each other to improve the image translation quality. In our experiments, we validate the effectiveness and versatility of our proposed CAM-consistency loss by applying it to several representative models for facial image editing, such as StarGAN, AttGAN, and STGAN.

1. Introduction

Image-to-image translation is a key task in computer vision, the aim of which is to learn the mapping of an input image in a source domain to the one in a target domain. Since generative adversarial networks (GANs) [6] have been proposed, showing their ability to create realistic images, numerous studies on image translation, such as facial attribute editing, have been conducted due to its practicality. However, it is unrealistic to find all the paired datasets for various attributes (*e.g.*, the same person with a different gender), so unpaired image translation approaches have also been studied. For example, Zhu *et al*. [31] introduced CycleGAN, which can translate the images between different domains using the unpaired dataset via the cycle-consistency loss. StarGAN [3] and AttGAN [8] were also proposed to edit the facial attributes while achieving multi-domain translations using a single generator.

Facial attribute editing, which aims to manipulate the particular attribute with the given face image, is still a challenging task. One such challenge lies in difficulty in preserve attribute-irrelevant regions while changing the desired attribute of a given image. For example, the existing models often change the overall color of an image to a golden color when imposing a blond hair attribute. Recently, additional modules were proposed to preserve attribute-irrelevant regions. Zhang *et al*. [29] proposed SaGAN that changes only the partial region of an image based on the estimated segmentation masks. However, such local manipulation may not be applicable in the case of changing the global attribute (*e.g.*, gender and age) translation. In response, CAFE-GAN [12] attempted to preserve the attribute-irrelevant regions by predicting the attribute-relevant information in feature maps, not at a pixel level, using the attention branch network (ABN) [5]. RelGAN [16] uses the relative attributes and utilizes a conditional adversarial loss by taking triplets consisting of two images and a vector of modified attributes for image translation. However, these approaches are not easily applicable to the general architectures since they require specific modules, and they do not employ an explicit loss for pixel-level preservation.

To further address this issue, we propose a novel, intuitive loss function called the CAM-consistency loss for image-to-image translation by utilizing the Grad-CAM [20] in adversarial training. Our proposed loss is widely applicable to the existing image translation approaches such as StarGAN [3], AttGAN [8], and STGAN [18] without modifying their architectures. Furthermore, our CAM-consistency loss can overcome various limitations of the existing methods since it enforces the generator to preserve the irrelevant regions of an image at a pixel level while making the discriminator attend to the attribute-relevant information at a feature level. This allows the model to generate the image at once while preserving the attribute-irrelevant regions as shown in Figure 1. In summary, our contributions include:

- We propose a novel loss function called the CAM-consistency loss, which can directly enforce the generator to preserve the attribute-irrelevant regions while the discriminator handles the attribute-relevant regions. It works even without any additional information such as segmentation maps of each attribute or any modification of the network architectures, and also it allows the generator and the discriminator collaborate with each other for the better performance.

- Our proposed CAM-consistency loss overcomes the limitations of the existing image-to-image translation approaches by directly preserving attribute-irrelevant regions computed by the discriminator.

- We demonstrate the possibility of using the Grad-CAM directly as training objectives, rather than just a visualization tool.

2. Related Work

2.1. Generative Adversarial Networks

Since originally proposed by Goodfellow *et al*. [6], GANs have shown impressive results in computer vision tasks such as image translation [3, 10, 17, 24, 31], super-resolution [13, 22], image manipulation [11, 14, 26], synthesizing realistic output images. Mirza and Osindero [19] introduced conditional GANs (cGANs) to generate the images with the given properties by taking the condition variables as input. The need for the expensive paired data for image translation, however, makes it difficult for real-world deployment. To overcome the lack of a paired training dataset, CycleGAN [31] introduced the cycle-consistency loss by restricting the generated image to be converted back to the original image. Furthermore, Choi *et al*. [3] proposed StarGAN to tackle multi-domain image-to-image translation tasks by leveraging an auxiliary classifier.

2.2. Facial attribute editing

Facial attribute editing is a prominent task for real-life applications with the increasing need to manipulate humans' facial images. StarGAN utilized a single model trained on several image datasets and performed the facial image editing with respect to the facial attributes shared among them. He *et al*. [8] introduced AttGAN, which divides the attribute information and image features by applying the attributes in the decoder part of the generator. The smooth warp fields were applied to GANs to perform semantic image editing at arbitrary resolutions including a very high resolution (4k images) [4]. Lee *et al*. [14] introduced MaskGAN that allows interactive facial image

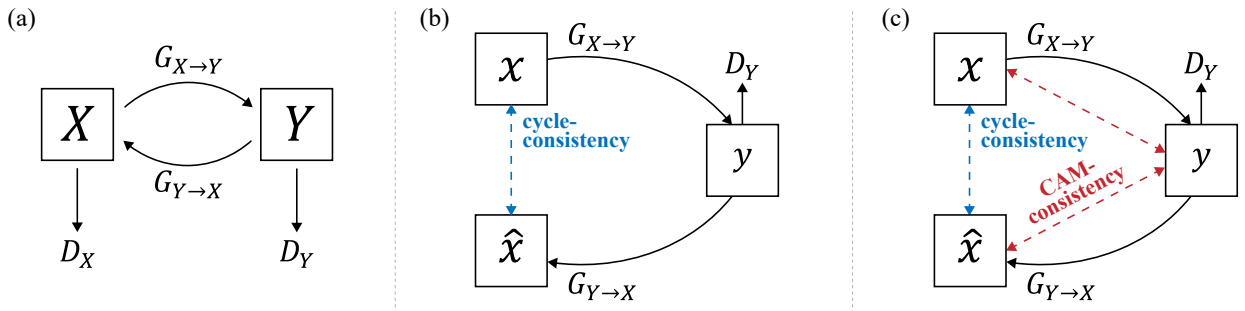


Figure 2: (a) Standard image-to-image translation approach composed of Generator $\boldsymbol{G}$ and the associated discriminator $\boldsymbol{D}$. The model is trained to translate a given image between different domains. (b) The cycle-consistency loss enforces $\boldsymbol{G}$ to recover the original image from the two consecutive translations ($X \to Y$ and $Y \to X$) so that $\boldsymbol{G}$ can preserve the original information of a given image. (c) Our CAM-consistency loss explicitly ensures consistency in each of the two translations by allowing the comparison between two different domains by attending only to the attribute-irrelevant regions. To capture the attribute-irrelevant regions, we utilize the Grad-CAM from the attribute classifier in D as the mask (*Grad-CAM mask*).

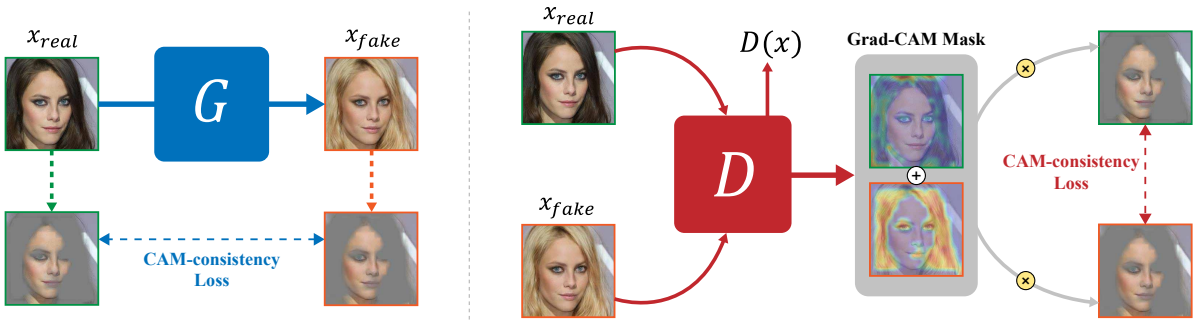
manipulation using semantic masks. In their work, semantic masks serve as an effective intermediate representation and enable users to manipulate the face images with flexibility and fidelity preservation. As the real-world applications of facial attribute editing have received increasing attention, it becomes crucial to change only the relevant part of interest to users while preventing the model from modifying the background or the other attributes the users intend to preserve. To solve such issues, further studies [7, 12, 16, 18, 25, 27, 29] were proposed. GANs with the replacing strategies [7, 27, 29] were proposed to preserve attribute-irrelevant regions. However, they replace the output with the calculated mask to manipulate face images, and hence, are not suitable for manipulating global attributes. On the other hand, Liu *et al*. [18] introduced STGAN with the style transfer unit (STU) similar to the gated recurrent unit (GRU) [2] to selectively transfer the style. ResAttr-GAN, a residual attributes learning model based on the Siamese network, were also proposed by Tao *et al*. [25] to learn the attribute differences in the high-level latent space. CAFE-GAN [12] utilizes the attention branch network (ABN) [5] and allows the discriminator to properly identify the regions for the specified attributes. Moreover, RelGAN [16] suggested a binary label transformation for using relative attributes in a multi-domain image-to-image translation task. They train the model using the conditional adversarial loss by taking triplets consisting of two images and a vector of modified attributes. However, these approaches does not generally employ the explicit loss for pixel-level preservation and require specific modules incompatible with the general architectures. In this work, we propose a novel loss function to overcome such limitations by restricting the generator to preserve other regions while improving the overall performance of the discriminator.

2.3. Interpreting Convolutional Neural Networks

The interpretation of convolutional neural networks (CNNs) [1, 5, 20, 21, 23, 28, 30] has attracted significant attention to understand the neural network behavior. Zhou *et al*. [30] proposed class activation mapping (CAM), which highlights the model's attention for the specific class. However, the requirement of the global average pooling (GAP) [15] layer in the CAM-based techniques hinders the network from training and hurt its versatility. The ABN [5] have been proposed to obtain the attention maps explicitly, which can be adapted to the classifier for visual explanation and the attention mechanism. Gradient-based visual explanation [1, 20, 21] is also widely used due to its ease of use in interpreting CNN models. Selvaraju *et al*. [20] proposed Grad-CAM, a generalized version of CAM that can be applied without structural changes of the network. Note that the existing studies mainly used the Grad-CAM for visual explanation for a given model. However, we propose a novel loss function that directly involves the Grad-CAM module as an optimizable part in the training process.

3. Proposed Method

To begin with, we consider the limitations of the existing cycle-consistency loss that does not preserve the detailed information when manipulating the facial image. To address it, this work proposes a straightforward approach to compare the real image and the generated one. Since it is generally impractical to compare the real image and the generated one directly, we propose the CAM-consistency loss that make the model attend only to those attribute-irrelevant regions such as the background when comparing the two images. Concretely, we utilize the Grad-CAM output from the auxiliary classifier to mask the relevant region to the attribute that the user wants to change. Figure 2 shows the dif-



(a) CAM-consistency for the generator

(b) CAM-consistency for the discriminator

Figure 3: (a) From the perspective of the generator, the CAM-consistency loss enforces it to preserve the attribute-irrelevant (unmasked) regions. (b) On the other hand, the CAM-consistency loss makes the discriminator properly determine the regions for the generator to change. In this manner, both the generator and the discriminator are trained to preserve the attribute-irrelevant regions.

ference between the cycle-consistency loss and the CAM-consistency loss. Note that the CAM-consistency loss is widely applicable to various architectures with an auxiliary classifier since the Grad-CAM module can be flexibly added without modifying the existing network structure.

In order to achieve this intuition, we adapt the CAM-consistency loss for both the generator and the discriminator of GAN models in the training process as depicted in Figure 3. Interestingly, by simply adding the CAM-consistency loss to the training objective, we observe that the generator and the discriminator collaborate with each other. In GANs, as the name says, the discriminator and the generator are basically trained in a manner that they compete against each other. However, to optimize our CAM-consistency loss, they have to collaborate with each other. In other words, from the perspective of the generator, it is essential that the discriminator properly attends to the attribute-relevant regions for preserving attribute-irrelevant regions. On the other hand, the discriminator has to properly update the Grad-CAM region corresponding to the given attributes to follow the manipulated regions by the generator. Thus, the generator has to make minimal necessary changes corresponding to the attribute-relevant regions so that the discriminator can correctly classify the attribute. As training proceeds, the generator can change the part where the discriminator considers to be highly correlated with the given attributes, while the discriminator can capture the attribute-relevant regions by focusing on the region where the generator made changes. Of course, our training objective is fully aligned with the original purpose of the image manipulation so that the application of our CAM-consistency loss does not harm the quality of generated images.

3.1. Grad-CAM Mask

Grad-CAM is the network-agnostic method that visualizes where the classifier attends for a particular class. Although Grad-CAM has been originally proposed for interpreting CNN models, we use this mechanism for masking attribute-relevant regions. For facial attribute editing, we define the *Grad-CAM mask* as the mask covers where the Grad-CAM attends for the given relative attribute. The relative attribute $\mathbf{att}_{s\to t}$ indicates the difference between the target attributes and the source ones, *i.e.*,

$$\mathbf{att}_{s\to t} = \mathbf{att}_t - \mathbf{att}_s. \tag{1}$$

The attribute vector is represented as a binary value. Therefore, the relative attribute has the value of 1 when a new attribute is introduced and the value of -1 if an attribute is removed. By taking the relative attribute, we can consider the changes in the attributes and also expect Grad-CAM to find the counter-factor with a negative value. To obtain the *Grad-CAM mask* for generated image $M_{s\to t} \in \mathbb{R}^{w\times h}$ of width w and height h, we compute the gradients for class scores y multiplied by the relative attribute vector $\mathbf{att}_{s\to t}$, with respect to feature map activations A^k of a convolutional layer, *i.e.*, $\partial(y \odot \mathbf{att}_{s\to t})/\partial A^k$. Following the original work, the neuron importance weights $\alpha^k_{s\to t}$ are obtained by global-average-pooling these gradients across the width and height dimensions, *i.e.*,

$$\alpha^k_{s\to t} = \frac{1}{Z}\sum_i\sum_j \frac{\partial(y \odot \mathbf{att}_{s\to t})}{\partial A^k_{ij}}. \tag{2}$$

In order to obtain the *Grad-CAM mask* $M_{s\to t}$, we apply normalization after performing a weighted combination. In addition, we only take the value between 0 and 1 to use it as a mask for image comparison. This also restricts the maximum value to 1 so that the regions with the values higher than the specific threshold (1 in our work) can be treated equally. Hence, the *Grad-CAM mask* $M_{s\to t}$ can be derived

as

$$M_{s\to t} = \text{Normalize}\left(\sum_k \alpha^k_{s\to t} A^k\right)_{[0,1]}. \quad (3)$$

In practice, after obtaining the mask for the generated image, we also get the mask for the original image in a similar manner using the corresponding relative attribute vector, $\mathbf{att}_{t\to s} = \mathbf{att}_s - \mathbf{att}_t$. The final *Grad-CAM mask* M_{CAM} is obtained by adding the two masks, one from the generated image, and another from the original image.

3.2. CAM-Consistency Loss

For both the generator and the discriminator, we implement the CAM-consistency loss using the *Grad-CAM mask* obtained in Section 3.1. Since the Grad-CAM has higher values in more relevant regions, we invert their values and multiply them to both the original image and the generated image. After masking the attribute-irrelevant regions, we compare the original image and the generated one, *i.e.*,

$$\mathcal{L}_{CAM}(x, x') = \mathbb{E}_{x,x'}[(\mathbb{1} - M_{CAM}) \odot ||x - x'||_1]. \quad (4)$$

Generator. In the generator, the CAM-consistency loss plays a role of enforcing the generator to preserve those regions not covered by the *Grad-CAM mask*. Assuming that the discriminator well predicts the regions for the classification, the generator can learn to change the region of an image that is highly related with the attributes while preserving its other regions. We add the CAM-consistency loss to the existing objective function of an image translation model for both forward and backward directions between the two domains. That is, our new objective function of the generator is written as

$$\mathcal{L}_G = \mathcal{L}_G^{origin} + \frac{1}{2}\lambda_{CAM}\left(\mathcal{L}_{CAM}(x, y) + \mathcal{L}_{CAM}(y, \hat{x})\right), \quad (5)$$

where x is the input image, y is the generated image, and $\hat{x}$ is the image reconstructed from y. λ_{CAM} is the hyper-parameter that controls the importance of the preservation of the attribute-irrelevant regions. To train the generator, we set the value of λ_{CAM} as 5 in all the experiments. Since optimizing output images for the Grad-CAM (rather than preservation) results in lower image quality in practice, we do not involve the terms related to the *Grad-CAM mask* generation while training the generator.

Discriminator. For training the discriminator, we add the CAM-consistency loss to the objective function in the same manner used as in the generator. However, it enforces the discriminator, but not the generator, to cover the changed regions of the image via the *Grad-CAM mask*. The objective function of the discriminator is written as

$$\mathcal{L}_D = \mathcal{L}_D^{origin} + \lambda_{CAM}\mathcal{L}_{CAM}(x, y), \quad (6)$$

where x is the input image and y is the generated image. λ_{CAM} is the hyper-parameter that controls how much the discriminator should focus on the region of an image where the generator make changes. Similar to the generator, we set the value of λ_{CAM} as 5 in all the experiments. Since the discriminator often uses shared layers to classify the attributes as well as to classify whether a given image is real or fake, the CAM-consistency loss can affect how the discriminator determine whether an image is real. In Section 4.6, we show that even applying the CAM-consistency only to the discriminator can indirectly enforce the generator to preserve the attribute-irrelevant regions.

To summarize, if the generator does not correctly perform image translation, the discriminator will not be able to properly attend to the given attributes since the Grad-CAM is trained to focus on the regions where the generator changes. Similarly, if the discriminator fails to attend to the attribute-relevant regions, the generator will not be properly guided to the region that should be preserved. Note that the CAM-consistency loss is applied to the existing models, indicating that both the generator and the discriminator are still trained via the adversarial loss and the classification loss. This allows the *Grad-CAM mask* to properly mask attribute-relevant regions, rather than arbitrary regions.

4. Experiments

4.1. Dataset

We evaluate our proposed loss function on the CelebA dataset containing 202,599 aligned facial images with 40 associated binary attributes. In all experiments, we consider 13 attributes, including *Bald, Bangs, Black Hair, Blond Hair, Brown Hair, Bushy Eyebrows, Eyeglasses, Male, Mouth Slightly Open, Mustache, No Beard, Pale Skin, and Young*, which have strong visual impact and widely used for the relevant work. Each image is cropped to 178×178 and then resized to 128×128. We take 2,000 images as test data for evaluation.

4.2. Baseline Models and Proposed Loss

We validate our CAM-consistency loss using widely-used models for facial attribute editing. Specifically, we apply our proposed loss in the training process of three different image-to-image translation models; StarGAN [3], AttGAN [8], and STGAN [18].

StarGAN uses a single shared generator for multi-domain image translation. The generator takes an image and the target attribute vector as input and generates the manipulated image. The cycle-consistency loss [31] is used for preserving the existing information of the image.

AttGAN aims to model the relations between the target vector and the latent representation of the image by treating the attribute vector as part of the latent representations.

StarGAN	*StarGAN	AttGAN	*AttGAN	STGAN	*STGAN
21.14 (±6.77)	**17.98 (±3.24)**	14.27 (±6.29)	**13.27 (±4.04)**	11.61 (±4.68)	**10.49 (±3.66)**

Table 1: Average FID scores for 13 translation tasks (lower is better). The models with * correspond to those trained with the CAM-consistency loss, and the values in parentheses denote the standard deviation.

Due to its autoencoder-like model architecture, AttGAN can preserve the information of the original image without the cycle-consistency loss.

STGAN utilizes the style transfer unit (STU) similar to GRU to overcome the limitations of the existing skip connections used by AttGAN. Unlike StarGAN and AttGAN, it takes the relative attribute vector indicating the difference between target and source attribute vectors to identify those attributes that need to be changed.

We use StarGAN[1], AttGAN[2], and STGAN[3] as our baseline models and then compare the performance after applying the proposed CAM-consistency loss. For each baseline model, we experiment with two types of models with and without the CAM-consistency loss. Since the loss can be applied without changing the network structure, we add the CAM-consistency loss to the objective functions for both the generator and the discriminator without modifying the structure or searching for new hyper-parameters.

4.3. Quantitative results

For quantitative evaluations of the visual quality of translated images, we use Fréchet Inception Distance (FID) [9] and peak-signal-to-noise ratio (PSNR) as our evaluation metrics. When computing PSNR, the corresponding regions to the changed attributes are masked using a pre-trained face parser model, in order to calculate the mean squared error (MSE) only for the attribute-irrelevant regions that should be preserved.

FID score. FID indicates how far the distribution of the manipulated image is from the distribution of the original one; thus, this metric is generally consistent with human evaluations. We calculate the FID scores of the translated images and the original images for each model to examine how well they keep the existing information with high quality. Table 1 reports the average FID scores for 13 translations of each attribute. As expected, the models with our proposed CAM-consistency loss can achieve lower FID scores. This indicates that our proposed loss helps the model preserve the overall image quality.

PSNR of attribute-masked images. We evaluate how well the CAM-consistency loss preserves the information of the given image when changing the local attributes by calculating PSNR with known segmentation masks. We use a pre-trained face parser trained with the CelebA-HQ dataset [14] to mask the attribute-relevant regions. We mask the hair regions for *Black hair, Brown hair, Blond hair* attributes, and mask skin regions (*e.g.*, skin, nose, and neck) for *Pale skin* attribute. Table 2 shows the PSNR results for each attribute.

[1] https://github.com/yunjey/stargan
[2] https://github.com/elvisyjlin/AttGAN-PyTorch
[3] https://github.com/bluestyle97/STGAN-pytorch

	Black hair	*Blond hair*	*Brown hair*	*Pale skin*	*Average*
StarGAN	24.29	21.78	25.06	21.86	23.25
*StarGAN	**25.46**	**25.30**	**26.94**	**24.96**	**25.66**
AttGAN	28.03	25.65	29.89	25.47	27.26
*AttGAN	**28.77**	**26.37**	**31.05**	**27.70**	**28.47**
STGAN	26.80	23.56	28.03	26.21	26.15
*STGAN	**28.84**	**25.31**	**29.70**	**29.22**	**28.27**

Table 2: PSNR results of attribute editing for validating the preservation ability and the visual quality (higher is better). We used face parser, which is pre-trained by CelebAMask-HQ dataset [14] for masking the local attributes (*e.g.*, hair and skin). The models with * correspond to those trained with the CAM-consistency loss.

	Brown Hair	*Blond Hair*	*Pale skin*	*Gender*	*Blond hair Pale skin*	*Gender Age*
AttGAN	35.0%	31.7%	16.7%	45.0%	38.2%	41.8%
*AttGAN	**65.0%**	**68.3%**	**83.3%**	**55.0%**	**61.8%**	**58.2%**

Table 3: User study results on facial attribute editing tasks (higher is better). The models with * correspond to those trained with the CAM-consistency loss.

4.4. Qualitative results

Figure 4 shows the results from the baseline models and those with our CAM-consistency loss. The existing models seem to preserve the important features from the original information with the cycle-consistency loss or with the encoder-decoder architectures. However, the lack of an explicit loss for preserving the attribute-irrelevant regions, those models often change the overall color of the images including the background regions. On the contrary, the models with the CAM-consistency loss maintain the attribute-irrelevant regions by simply adding our proposed loss function during training. In particular, it prevents the existing models from generating new faces with the given attributes or changing the color of the background on different attributes, such as *Pale skin* or *Blond hair*. Interestingly, we can observe that the changed regions appear differently for each model, regardless of the application of our CAM-consistency loss, indicating that the architecture of the base model itself is still important.

4.5. User Study

We conduct a user study for validating the effectiveness of our proposed CAM-consistency loss, as the results are

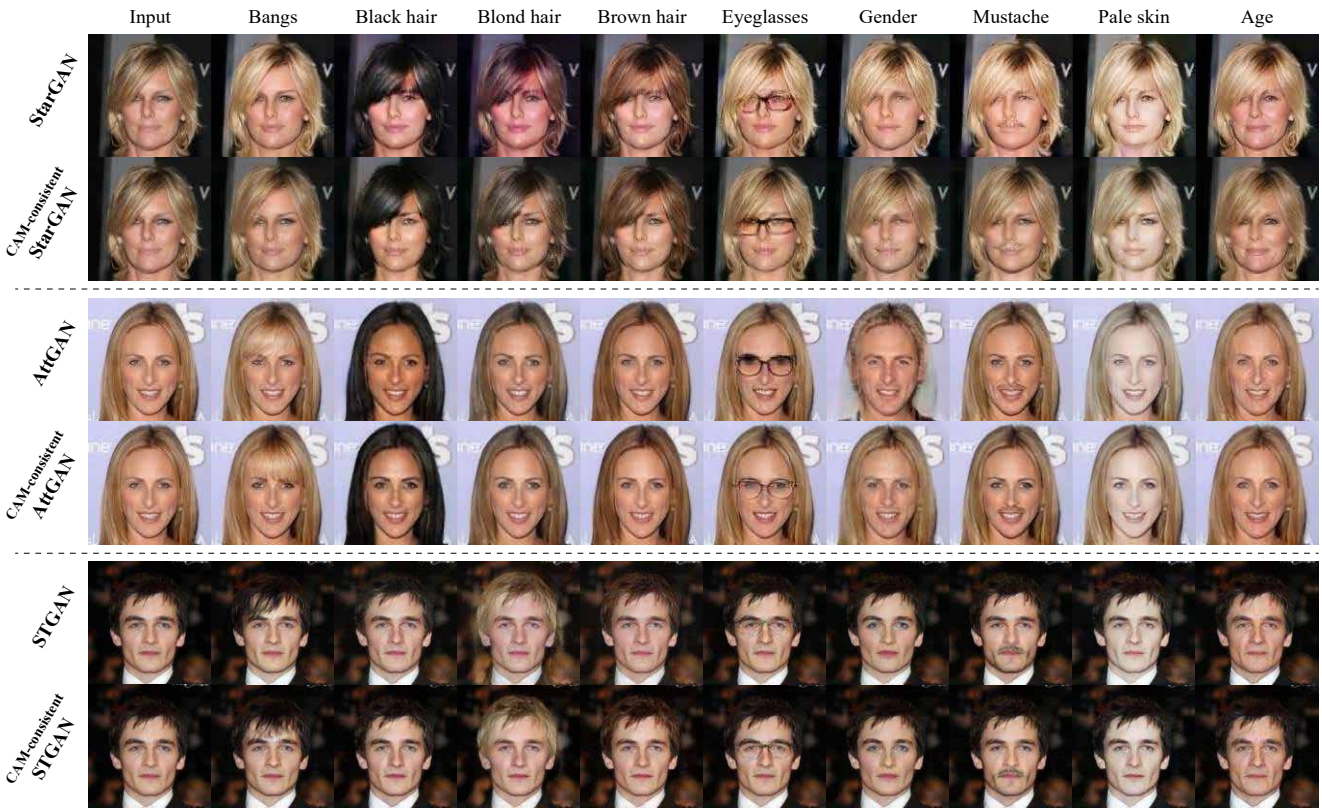


Figure 4: Results for each baseline and the ones trained with the CAM-consistency loss. The first column shows the input image while the other columns show the attribute-editing results. If the input image already has an attribute, it will be translated into an image without such an attribute. Note that we add the CAM-consistency loss without any modifications to the model structure or hyper-parameter tuning.

shown in Table 3. We consider six tasks that significantly impact the overall quality, including preserving the irrelevant parts (*e.g.*, background) of the image. AttGAN is used as our baseline due to its superior capability in preserving the original information, as shown in Table 2. Since our proposed loss offers the preservation of the original images, we provide the gif-format examples where the real image and the generated one are shown in the same place to visually highlight the differences between the two images. We provide the examples from AttGAN with the pairs of the original one and the one with the CAM-consistency loss. Then we request the participants to choose the best one, which properly translates an image with respect to the given attributes while preserving the remaining regions of the image.

4.6. Ablation Study

In this work, we apply our CAM-consistency loss to both the generator and the discriminator by default. We conduct an ablation study to validate the impact of the CAM-consistency loss in the generator and the discriminator. Figure 5 shows the results of facial attribute editing and the corresponding Grad-CAM from the discriminator in different settings. We use AttGAN as the baseline model. *Generator only* and *Discriminator only* show the results of applying the CAM-consistency loss only to the generator and applying it only to the discriminator, respectively. When the CAM-consistency loss is applied only to the generator, the discriminator cannot fully attend to the attribute-relevant regions. Due to the incomplete mask, the proper change of the attributes cannot be made; even the generated image may look identical to the original one. On the contrary, when the CAM-consistency loss is applied only to the discriminator, the discriminator seems to capture the attribute-relevant regions. However, since there is no direct influence to the generator, it still fails to manipulate the attributes while keeping all the attribute-irrelevant regions. For example, the generator cannot preserve the hair regions when translating the input image to a different gender. Nevertheless, interestingly, enforcing the discriminator to capture the region where the generator should change can slightly help to preserve the attribute-irrelevant regions even without restricting the generator by the loss functions. We conjecture that the discriminator indirectly affects the

Figure 5: Results of facial attribute editing and the corresponding Grad-CAM in different settings of applying the CAM-consistency loss. In our work, the CAM-consistency is applied to both the generator and the discriminator as shown in the last two rows (*Both*).

generator in adversarial training, and this highlights the importance of the properly trained classifier in an image translation model. The last two rows (*Both*) show the results of the CAM-consistency loss applied to both the generator and the discriminator, as proposed in our work. The generator and the discriminator perform better than any other setting by collaborating with each other. The generator can fully preserve the hair regions when it translates the image to a different gender. This indicates that the generator and the discriminator can improve the overall performance of the model by collaborating with each other.

5. Conclusions

In this work, we proposed a novel, intuitive loss function called the CAM-consistency loss, which can consistently improve the performance of the existing image-to-image translation models. Using our proposed approach, the generator and the discriminator can collaborate with each other while improving their performances even in adversarial settings, showing the potentials in using Grad-CAM directly as the training objective. We hope that our work will encourage researchers to re-think how to fully utilize the interactions between the generator and the discriminator.

Acknowledgments. This work was partially supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2018M3E3A1057305 and No. NRF-2019R1A2C4070420).

References

[1] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision, (WACV)*, 2018. 3

[2] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 3

[3] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 2, 5

[4] Garoe Dorta, Sara Vicente, Neill D. F. Campbell, and Ivor J. A. Simpson. The GAN that warped: Semantic attribute editing with unpaired data. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2020. 2

[5] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019. 2, 3

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2

[7] Zhenliang He, Meina Kan, Jichao Zhang, and Shiguang Shan. PA-GAN: progressive attention generative adversarial network for facial attribute editing. *CoRR*, abs/2007.05892, 2020. 3

[8] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Trans. Image Process.*, 2019. 2, 5

[9] 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. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 6

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017. 2

[11] Youngjoo Jo and Jongyoul Park. SC-FEGAN: face editing generative adversarial network with user's sketch and color. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2019. 2

[12] Jeong-gi Kwak, David K. Han, and Hanseok Ko. CAFE-GAN: arbitrary face attribute editing with complementary attention feature. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[13] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017. 2

[14] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2020. 2, 6

[15] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *Proc. the International Conference on Learning Representations (ICLR)*, 2014. 3

[16] Yu-Jing Lin, Po-Wei Wu, Che-Han Chang, Edward Y. Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2019. 2, 3

[17] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[18] 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 *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019. 2, 3, 5

[19] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2

[20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2017. 2, 3

[21] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. 3

[22] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised MAP inference for image super-resolution. In *Proc. the International Conference on Learning Representations (ICLR)*, 2017. 2

[23] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *Proc. the International Conference on Learning Representations (ICLR)*, 2015. 3

[24] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *Proc. the International Conference on Learning Representations (ICLR)*, 2017. 2

[25] Rentuo Tao, Ziqiang Li, Renshuai Tao, and Bin Li. Resattr-gan: Unpaired deep residual attributes learning for multi-domain face image translation. *IEEE Access*, 2019. 3

[26] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018. 2

[27] Xuan Xia, Fengqi Yu, Nan Li, Yansong Qu, Jiajia Zhang, and Chengguang Zhu. Self-attention-masking semantic decomposition and segmentation for facial attribute manipulation. *IEEE Access*, 2020. 3

[28] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2014. 3

[29] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2018. 2, 3

[30] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016. 3

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of the IEEE international conference on computer vision (ICCV)*, 2017. 2, 5