Consistency Preservation and Feature Entropy Regularization for GAN based Face Editing

Weicheng Xie, Wenya Lu, Zhibin Peng, Linlin Shen*

***Abstract*—Generative Adversarial Network (GAN) has been widely used for image-to-image translation-based facial attribute editing. Existing GAN networks are likely to generate samples with anomalies, which may be caused by the lack of consistency preservation and feature entanglement. For preserving image consistency, many studies resorted to the design of the network framework and loss functions, e.g. cycle-consistency loss. However, the generator with the cycle-consistency loss could not well preserve the attribute-irrelevant features, and its feature-level noises may possibly cause synthesis abnormalities. For feature disentanglement, previous works were devoted to mining the implicit semantics of feature spaces, while these semantics are not stable and intuitive enough. For consistency preservation, we propose a target consistency loss to complement the cycle-consistency loss, and enable the network to learn to preserve features of the image more directly. Meanwhile, we filter out outlier feature maps to reduce the synthesis abnormalities and propose a dynamic dropout to better preserve the attribute-irrelevant features. For feature disentanglement, we encode the image semantics more stably and intuitively and propose an entropy regularization to decouple these semantics to allow independent editing of different attributes. The proposed modules are general and can be easily integrated with available image-to-image-based GAN models like StarGAN, AttGAN, and STGAN. Extensive experiments on CelebA dataset show that the our strategy can largely reduce the artifacts and better preserve the subtle facial features, and thus significantly improve the facial editing performance of these mainstream GAN models, in terms of FID, PSNR and SSIM. Additional experiments on realistic expression editing show that our method outperforms StarGAN on RaFD, and achieves much better generalization performances than the three baselines on datasets of FFHQ, RaFD and LFW.**

***Index Terms*—GAN; Consistency preservation; Entropy regularization; Self-adaptive dropout.**

I. Introduction

GENERATIVE adversarial network (GAN) has been frequently applied to image-to-image translation and has achieved appealing results [1]. StarGAN [2], AttGAN [3] and STGAN [4] achieved multi-domain translations using a single generator. In the facial attribute editing, artifacts frequently occur in the samples generated by existing models. For example, the existing models often change the features of a face and produce artifacts when translating a hair color attribute. We collectively refer to the changes of irrelevant attributes and artifacts in the sample as abnormalities. What is more, preliminary experiments show that these abnormalities will be more serious as the number of training iterations increases. There are two possible reasons for these abnormalities in the generated samples. (i) It has been a challenge to preserve the consistency of irrelevant attributes in facial attribute editing. (ii) Feature entanglement causes attribute entanglement, which in turn causes other attributes to change when one attribute is edited.

To preserve consistency, many studies proposed different network architectures and losses to constrain the generator. Zhu et al. [5] proposed a cycle consistency loss for unpaired image-to-image translation. Cycle consistency loss preserves consistency by encouraging the generator to re-synthesize the images toward the original images during training. This reconstruction in face attribute editing has widely used in many GAN networks [2], [6], [7], [8], [9]. Some works made improvements on cycle consistency loss to adapt to different tasks, e. g. handling asymmetric unpaired image-to-image translation [10] and preserving consistency in latent space to maintain the identity of the person [11]. In recent years, the preservation of attribute-irrelevant regions has been extensively studied [6], [8], [9]. Specifically, based on cycle-consistency loss that guides the network to preserve the consistency of attribute-irrelevant regions, these works learned a mask of the editing area. However, in facial attribute editing where a single generator realizes multi-attribute translation, cycle consistency loss may not well preserve the features of the image. For example, when the network takes the already edited image as input and uses its attribute as the target label, it can be observed that the generator does not tend to preserve the features of this image. This is due to the lack of consistency constraints of the generator in the target attribute space. In this work, to give the generator a more comprehensive and stable consistency constraint, we propose a target consistency loss to complement the cycle consistency loss.

On the other hand, some works resort to additional regularizations to preserve the consistency of the GAN network. Zhang et al. [12] and Zhao et al. [13] focused on penalizing the sensitivity of generators or discriminators on the augmented data. This allows the network to learn more robust features from the original and augmented data, which can in turn stabilize training and improve the quality of synthesis. SPA-GAN [14] used the attention in the discriminator to narrow the distribution gap between the source and target domains. SReGAN [15] regularized the distinction among classes in the feature space. Wei et al. [16] proposed dropout-based regularization in the discriminator to maintain the consistency between the images before and after dropout. However, noises

W. Xie, W. Lu, Z. Peng and L. Shen are with Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, 518060, China. W. Xie was also an academic visitor in School of Computer Science, University of Nottingham, Nottingham, UK. Corresponding author: Prof. Linlin Shen, Tel: 86-0755-86935089, Fax: 86-0755-26534078, llshen@szu.edu.cn

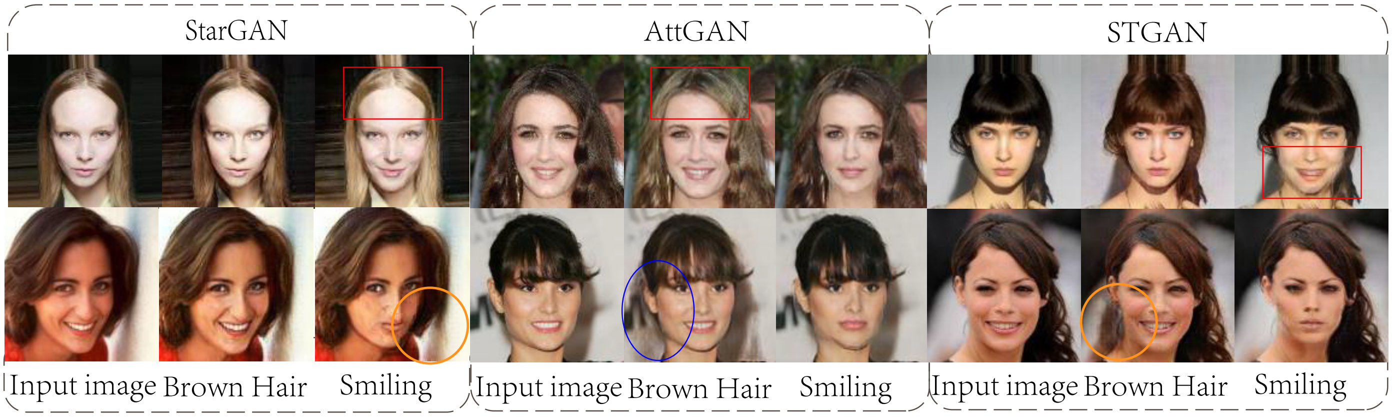

Fig. 1. Visualization of the face attribute editing results of three popular GANs, i.e. StarGAN [2], AttGAN [3] and STGAN [4]. The first row presents the edited images where the attribute-irrelevant regions are changed. The second row shows the problem of inconsistency between images before and after editing.

may possibly appear in the feature space of the generator, which will affect the editing of attributes. Thus, the suppression of such noises is useful to reduce the abnormalities in the generated samples. Dropout [17] is a common regularization method, but the original dropout based on random dropping is not desirable to the GAN generator. In recent years, many works have proposed new dynamic dropout strategies for different tasks [18], [19], [20], [21]. Thus, we propose to filter out the feature maps with such noises, and introduce dynamic dropout to suppress these features to improve the generator.

In addition, feature entanglement will also cause abnormalities in the generated samples. More intuitively, the entanglement of features makes the editing of multiple attributes entangled as well, i.e. manipulating each feature map to edit the attribute-related regions will undesirably affect other semantic regions. To reduce feature entanglement, many works have proposed improved network structure [24], [25] and feature map-based orthogonality [26], [27]. However, the devised network structures are often task-specific, which are difficult to generalize to other tasks. Node-based and feature map-based disentanglement are not efficient enough since a large number of nodes and feature maps are included in each network. Some studies [28], [29], [30] excavated and pruned the features in the context of hidden semantics, i.e. these features are mapped into the semantic space for disentanglement. PA-GAN [22] disentangled attribute semantics via an overlapping loss of the attention maps specific to the attributes with disjoint regions, while hidden semantics are not stable and intuitive enough to achieve accurate attribute editing. EigenGAN [23] proposed to control the semantic attributes via an unsupervised mining of interpretable and controllable eigen-dimension, i.e. a latent feature from each layer of the generator, while the specific network architecture employed by EigenGAN cannot be directly applied to image-to-image translation. In this work, we propose a regularization module that can offer a more stable and intuitive guidance to achieve semantic disentanglement for the feature maps.

A. Contributions

In this work, to give insight into the abnormalities, i.e. changes of irrelevant attributes and artifacts (see labeled regions in Fig. 1 for examples), caused by GAN generator, we propose two modules for consistency preservation and one module for feature entropy regularization. As far as we know, this is the first work to specifically study the abnormalities in generators.

These proposed modules are general and can be easily transferred to most image-to-image GANs for face attribute editing. Our contributions are summarized as follows

- A target consistency loss is proposed to preserve the generation consistency within the target attribute space, which help generators to preserve the global image consistency, especially for the face-irrelevant attributes.
- A dropout algorithm that adapts to the response location is proposed to suppress outlier feature maps and retain more subtle features, where a filtering mechanism with an efficient hash encoding and a dynamic dropout is introduced to detect and suppress outlier features.
- We propose a semantic entropy regularization on feature maps to decouple the features of different attributes, to enable each feature map in the generator to edit semantic attributes independently.
- Three popular GANs equipped with our approach achieve better visual, quantitative and generalization performances than the baselines on four face datasets, where the artifacts of the synthesized images are largely reduced and the attribute-irrelevant identity cues are better preserved.

II. Related Work

A. Facial Attribute Editing by Image-to-Image Translation

GANs [1] has shown impressive results in image-to-image translation. Isola et al. [31] used a conditional adversarial network to learn a mapping from paired images. Since paired images are not easy to obtain, CycleGAN [5] constrained the conditional generation of unpaired images. Odena et al. [32] proposed auxiliary classifier for multi-domain translation. Furthermore, several unified network architectures based on GAN are proposed to train a single generator for the task of multiple attributes editing [2], [3], [4]. For photo-realistic attribute editing, Dorta et al. [33] proposed WrapGAN to learn the smooth wrapping fields. Xiao et al. [34] proposed ELEGANT to exchange attributes in a pair of images by exchanging attribute-related latent codes. Yang et al. [25] proposed L2M-GAN, a model based on StarGAN-V2 [7], to separate and orthogonalize style vectors into attribute-relevant and irrelevant codes for preserving irrelevant attributes. Though well preserving the attribute-irrelevant facial regions, L2M-GAN could possibly cause background distortion. Thus, existing models still change irrelevant attributes and generate artifacts in the editing of local facial attributes. In this work, we propose to suppress these artifacts from two perspectives, i.e. consistency preservation and attribute disentanglement.

B. Consistency Preservation

Zhu et al. [5] proposed the cycle-consistency loss to encourage model to preserve consistency, which is also used in [2], [6], [7]. To further preserve consistency in image-to-image translation, Arantes et al. [9] proposed a new cycle-consistency loss based on semantic segmentation. Some works resort to the preservation of semantic structure [35] and spatial structure [36] during translation. The studies [6] [18], [37] enabled the network to learn a mask of the edited region to preserve the consistency of irrelevant attributes. Most of these modules are based on cycle consistency loss, while this loss may ignore image feature preservation in the target attribute space of the generation. In this work, we propose a target consistency loss to preserve this image consistency.

C. Regularization of GAN

The regularization in GAN can play important role in stabilizing training, preventing overfitting, feature disentanglement, and improving the quality of generation. To stabilize training and improve the model generation ability, DQE-GAN [38] dynamically evaluated the image quality with discriminator, TWGAN [39] selected the optimal discriminator to optimize the generator, and many works [40], [41] proposed normalization and regularization of feature maps. As noise is common in the feature space of the generator, we introduce a regularization-based method to suppress such abnormal features with noises.

For feature disentanglement, some works have proposed methods based on feature orthogonality [26], [27], [42] and network architecture improvement [24], [42]. The studies [22], [28], [29], [43], excavated the semantic information in the feature space to achieve independent editing of attributes. Previous works were devoted to mining the hidden semantics of feature space, and decoupling them to achieve independent editing of attributes. In this work, we resort to the segmentation to explicitly encode the semantics in a stable way. Based on this, an entropy regularization is introduced to disentangle the semantics for independent editing.

III. THE PROPOSED ALGORITHM

The entire framework of the proposed generator is shown in Fig. 2, where the motivation and method of each proposed module are introduced below.

A. Target Consistency Loss

*1) **Motivation**:* Although most GANs introduced the reconstruction loss, during the training process, it is not easy to preserve irrelevant attributes and features. As shown in red boxes in Fig. 1, when we change the smiling attribute, the hair color is undesirably changed as well. And when AttGAN is requested to translate the color of brown hair to brown, the hair color is not well preserved.

Specifically, in current GAN-based face attribute editing, in order to preserve consistency, i.e. retaining attribute-irrelevant regions, the following reconstruction loss has been proposed.

$$\mathcal{L}_{CCL} = E_{x,c,c'}[\|x - G(G(x,c'),c)\|_1] \tag{1}$$

where x is the input image, c and c' are original and target attribute labels, respectively, $G(x,c')$ is the generated image with target attribute c', $E[\cdot]$ is the expectation operator.

During the training of facial attribute editing, the generator G learns the mapping from the original attribute space $\mathcal{X}$ to the target attribute space $\mathcal{X}'$. Cycle consistency loss (CCL) constrains the mapping function learned by G to be cycle-consistent, as follows

$$\begin{cases} \mathcal{L}_{CCL} = E_{x,x',\hat{x}}[\|(x-x') - (\hat{x}-x')\|_1] \\ x' = G(x,c') \in \mathcal{X}', \hat{x} = G(G(x,c'),c) \in \mathcal{X} \end{cases} \tag{2}$$

As shown in Fig. 3 and Eq. (2), for $x \in \mathcal{X}$, CCL encourages G to generate an image $\hat{x}$ similar to x in a translation cycle, i.e. $x \to x' \to \hat{x} \approx x$. In the translation cycle, G performs mapping $x \to x'$ and mapping $x' \to \hat{x}$, respectively, and guides these two mappings to be the inverse mapping of each other. More intuitively, CCL guides the changes during the two rounds of attribute editing to offset each other at the pixel level, i.e. $x - x' \approx \hat{x} - x'$. CCL emphasizes the consistency of changes during the mapping between different attribute spaces in a translation cycle, but does not emphasize the consistency of features within generation space of the target attribute.

Our intuition is that G should preserve consistency in the target attribute space while preserving cycle consistency. For example, taking x' and target attribute labels c' as input, G generates a sample $\hat{x}' = G(x',c')$. In this case, we expect that $\hat{x}'$ is similar to x', i.e. G can well preserve the features of x' during the translation in target attribute space. However, G under the mere CCL constraint may produce unexpected results, e.g. G will apply the learned information from x to x' and generate $\hat{x}' \approx x' + (x' - x)$. With a large capacity of possible mapping space, there may be a large difference between the input and desired images when x' and c' are used for the input, especially when G lacks the consistency constraint specific to the target attribute space.

*2) **Method**:* To give G more comprehensive and stable consistency constraints, a target consistency loss (TCL) in addition to CCL is proposed to improve image consistency. As shown in Fig. 3, by inputting the generated image x' and target

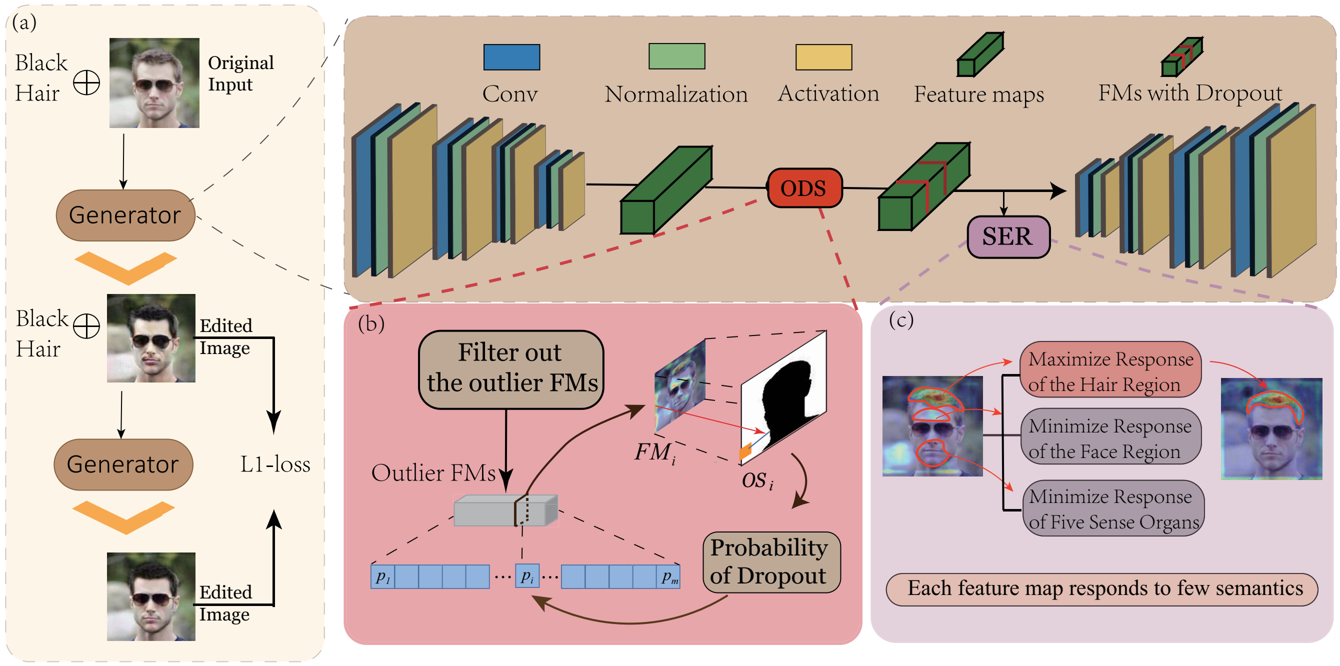

Fig. 2. The framework of the proposed generator: (a) Target consistency loss (TCL). (b) The suppression of outlier feature map based on dynamic dropout (ODS), where os_i is the distance from the maximum response area to the critical area of feature map FM_i, and p_i is the dropout probability specific to FM_i. (c) The semantic information entropy regularization (SER) based on region segmentation for making each feature map respond to as few regions as possible.

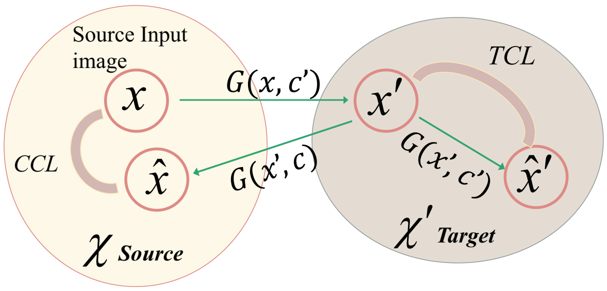

Fig. 3. The framework of cycle consistency loss (CCL) and target consistency loss (TCL). $\mathcal{X}$ and $\mathcal{X}'$ denote original and target attribute spaces, respectively, and x is the original input image. c and c' are source and target attribute labels, respectively.

attribute c' to G, TCL guides G to preserve image consistency in target attribute space, i.e. $x' \rightarrow \hat{x}' \approx x'$, as follows

$$\mathcal{L}_{TCL} = E_{x,c'}[\|x' - G(x',c')\|_1] \tag{3}$$

where $x' = G(x,c')$. The loss TCL guides the two generated samples x' and $\hat{x}' = G(x',c')$ to approximate each other as close as possible.

Through the joint guidance of CCL and TCL, the mapping learned by G satisfies the cycle consistency and the consistency within the target attribute space.

B. Dynamic Suppression of Outlier Feature Maps

*1) **Motivation**:* Many reconstruction losses are based on the constraints specific to the entire image, which does not pay enough attention to the critical regions and important regions with less pixels. As shown in blue ellipses in Fig. 1, in the translation of brown hair, a noisy block is undesirably produced to extend the hair region. Meanwhile, the shapes of the eyes and nose are also slightly changed.

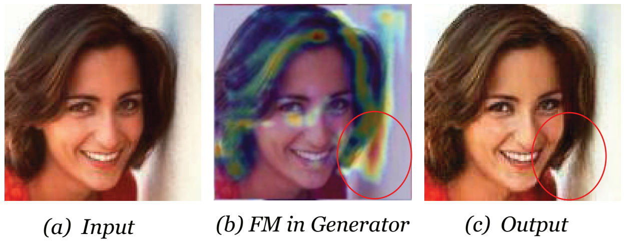
(a) Input (b) FM in Generator (c) Output

Fig. 4. (a) An input image. (b) Feature map (FM) of the middle layer of StarGAN. (c) The image synthesized by StarGAN.

To shed light on the cause of these artifacts, we analyze their generation in the feature map representation. As shown in Fig. 4, the feature space of the generator incorrectly responds to the shaded region in the red circle, which causes artifacts in the output samples during hair color translation. These artifacts appear as abnormal responses in the feature space, mostly around the critical regions to be edited, and affect the preservation of attribute-independent regions. We define these maps with abnormal responses as abnormal feature maps.

In this work, we identify outlier feature maps far from the cluster centers as candidates of abnormal feature maps. Furthermore, as abnormal responses appear around the key regions, the feature maps with the maximum response being close to the edited region are intensively suppressed using the dropout algorithm. In this way, image information is preserved as much as possible.

The framework of the proposed outlier feature map suppres-

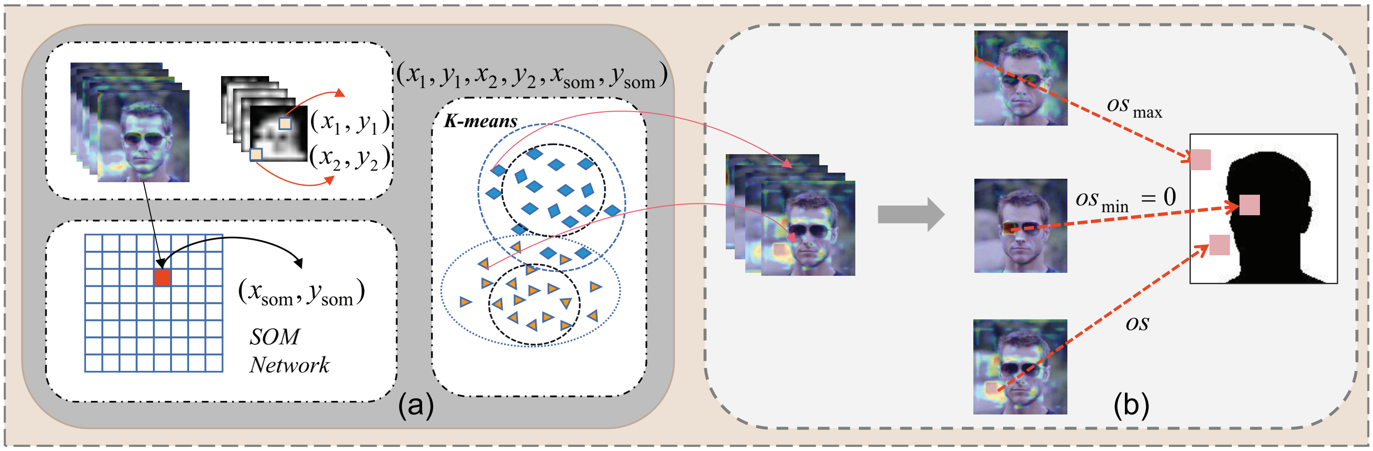

Fig. 5. The detection and suppression of outlier feature maps. (a) The clustering of feature maps based on hash encoding. (x_1,y_1) and (x_2,y_2) are coordinates of top-2 largest responses. (x_{som},y_{som}) is the coordinate of the winning neuron of SOM network output. (b) The probability of dropout is obtained based on the distance os from the largest response to the face, where os_{min} and os_{max} are the minimum and maximum distances.

sion is presented in Fig. 5.

2) Method: **Hash encoding for outlier FMs screening:** To filter out the outlier feature maps, we cluster the feature maps and select those far from centers as the outliers. Motivated from the study [21], we use hash coding to encode each feature map and speed up the clustering.

In [21], the coordinates of the top-nc largest response areas are formulated as follows

$$hr_{lr} = (x_{i_1}, y_{i_1}, \cdots, x_{i_{nc}}, y_{i_{nc}}) \tag{4}$$

where (x_{i_j}, y_{i_j}) is the coordinate of the top j-th largest average response. In this work, the setting of $nc = 2$ is employed.

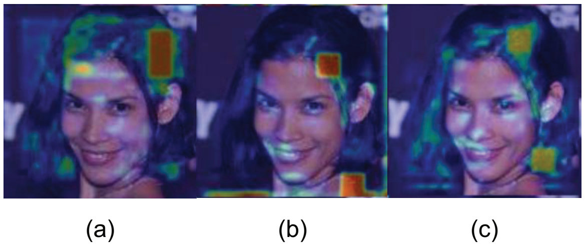
(a) (b) (c)

Fig. 6. (a)-(c) are three example feature maps of the 10-th layer of the StarGAN generator, with the size of 32×32. We divide the feature map into 8×8 regions, and the red labels the largest activation region of the feature map. Both (a) and (b) respond to the hair region and $L_2(a,b) < L_2(b,c)$ (L_2 denotes L_2-norm distance), while $L_2(hash(a), hash(b)) > L_2(hash(b), hash(c))$ after hash encoding in Eq. (4).

However, as shown in Fig. 6, the two feature maps with similar maximum-response coordinates may still respond to regions that are largely different, due to the introduced noises in the feature space. Given this consideration, the self-organizing mapping network (SOM) [44] that can map similar inputs to the closer neurons is further used.

More precisely, we use the coordinate $c_{lr} = (x_{som}, y_{som})$ of the winner neuron in the output layer to encode the feature map after inputting it into a SOM network, while the distance with winner neuron can well reflect the feature map similarity. By concatenating hr_{lr} in Eq. (4), the final hash encoding is formulated as follows

$$hc = (hr_{lr}, c_{lr}) \tag{5}$$

Based on the K-means clustering with hash coding of the feature maps, the outlier feature maps [21] are defined as those far from the cluster centers as follows

$$d_i > \kappa \frac{\gamma}{\sqrt{2nN^{(io)}}}, \tag{6}$$

where $\kappa = 1.5$ is a predetermined value; γ is the L_2-norm of the feature map; n is the dimension of feature map after vectorization; d_i is L_2-norm distance between feature map f_i and the center of the class it belongs to, $N^{(io)}$ is the number of feature maps in this class.

We argue that the outlier feature maps mainly reflect the attribute-irrelevant cues [21], which are also useful for preserving reconstruction consistency. Thus, instead of getting rid of these outlier feature maps directly, we suppress them with dynamic dropout.

Feature re-screening based on semantic regions: Motivated from the study [45], we highlight the attribute-irrelevant features that are not well preserved, and propose a dropout with dynamic probabilities in Fig. 5. In these outlier feature maps, we argue that those with maximum response close to the face region are associated with attribute-relevant features, which should be suppressed with a larger dropout probability. In this way, the feature maps specific to the attribute-irrelevant cues are more prone to be retained, which enables the generator to preserve more subtle feature information.

To this end, a segmentation algorithm, i.e. BiseNet [46] is used to locate the face region, then the distance from the largest response to the face region is used for the dynamic update of the specific dropout probability, which is formulated as follows

$$os_i = \inf_{cd \in CD} \|cd_{f^i} - cd\|_2 \tag{7}$$

where cd_{fi} is the coordinate of maximum response area of the feature map f^i, CD is the collection of the coordinates specific to the attribute-relevant regions via BiseNet [46].

The dropout probability specific to the i-th feature map f^i is updated as follows

$$p_i = \begin{cases} 0 & f^i \notin FM_{ol} \\ p_{max} - \frac{p_{max}-p_{min}}{os_{max}-os_{min}} \cdot (os_i - os_{min}) & f^i \in FM_{ol} \end{cases} \quad (8)$$

where FM_{ol} is the set of outlier features, os_{min} and os_{max} are the minimum and maximum distance values. $p_{max} = 0.7$ and $p_{min} = 0.3$ are predetermined hyperparameters.

In order to make up for the loss of information caused by feature map dropout, we minimize the difference of generated images with and without dropout as follows

$$\mathcal{L}_{ODS} = E_{x,c'}[\|G(x,c') - G_{drop}(x,c')\|_1] \quad (9)$$

where G_{drop} denotes the generator with the proposed dropout, $\|\cdot\|_1$ denotes the L_1-norm distance.

C. Semantic Information Regularization

*1) **Motivation**:* In addition to the outlier feature maps, the generator of GAN may also suffer from feature entanglement. As shown in yellow circles in Fig. 1, the generator mistakenly treats the shadow area as the hair to be edited, and yields abnormal generation.

We speculate one possible reason of generating these artifacts is that the semantic information represented by each feature map is entangled. More precisely, each feature map may represent the cues associated with multiple sematic regions, which makes it hard for the generator to edit the attributes independently. As shown in Fig. 4, a feature map that mainly responds to the hair region also represents a shadow region in the red circle. In this way, when the hair region is edited by this feature map, feature entanglement will inevitably produce this shadow region.

Another possible reason maybe that the semantic information learned by the generator is inaccurate, e.g. generators often treat hair-surrounding shadow as a hair region for editing, causing abnormalities.

In this work, in order to make each feature map represent as few semantic regions as possible, an entropy regularization of the semantic information is proposed, where the semantics are explicitly represented based on face segmentation model.

*2) **Method**:* Specific to the task of facial attribute editing, we divide the face into three semantic regions, i.e. hair, face, and the five sense organs.

The framework of the proposed entropy regularization is presented in Fig. 7, where we obtain the information encoding of each feature map (FM) for the following entropy regularization of semantics.

For the sample x, we obtain the masks of the three semantic regions based on BiseNet [46], i.e. $M = \{m_1, m_2, m_3\}$, where m_j is the $0-1$ mask of j-th semantic region, i.e. a pixel is assigned with 1 if it belongs to the j-th semantic region, and 0 otherwise. As shown in Fig. 7, for each feature map $f^i \in F = \{f^1, \cdots, f^n\}$, we obtain its information encoding $\{S(f^i,m_1), S(f^i,m_2), S(f^i,m_3)\}$ for three semantic regions by weighted averaging as follows

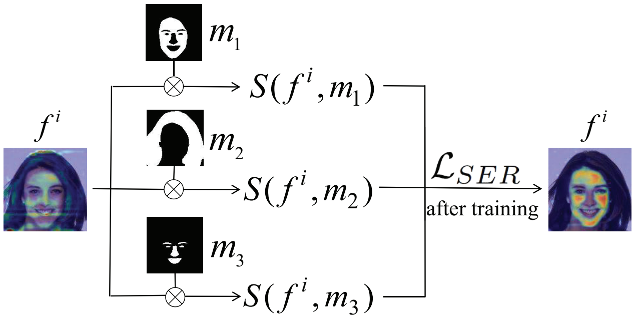

Fig. 7. The framework of the proposed semantic entropy regularization (SER). m_j is the mask of j-th semantic, and $S(f^i,m_j)$ is the information value specific to the i-th feature map of the j-th semantic. $\otimes$ denotes element-wise multiplication.

$$\begin{cases} S(f^i,m_j) = mean(abs(f^i) \otimes m_j) \\ S(f^i,m_j) \leftarrow S(f^i,m_j)/\sum_j \|S(f^i,m_j)\|_2 \end{cases} \quad (10)$$

where $abs(\cdot)$ is to take the absolute value of each element, $\otimes$ and $mean(\cdot)$ are element-wise multiplication and the averaging operation.

Based on the information encoding of each semantic region of the feature maps, we introduce the information entropy regularization by minimizing the following loss

$$\mathcal{L}_{SER} = -\frac{1}{n}\sum_i^n \sum_j^3 S(f^i,m_j) log(S(f^i,m_j)) \quad (11)$$

By minimizing $\mathcal{L}_{SER}$, the gradient backpropagation causes the largest value $S(f^i,m_{j_0}) = max(S(f^i,m_j)) \rightarrow 1$, and makes $S(f^i,m_j) \rightarrow 0$ $(j \neq j_0)$.

By minimizing the loss in Eq. (11), we not only encourage each feature map to respond to as few semantic regions as possible, but also offer the model more prior knowledge to represent the semantic region. This can facilitate the decoupling of the semantic attributes during editing and allow the model to edit these attributes independently.

D. Network Training

In a nutshell, we propose three modules in the generator and design the corresponding loss functions.

Target Consistency loss (TCL). Besides of the cycle consistency loss enabling the generator to preserve consistency between source and target attribute spaces in Eq. (1), a target consistency loss in Eq. (3) is proposed to preserve consistency in the target attribute space.

Dynamic Suppression of Outlier Feature Maps (ODS). In order to suppress abnormal FMs and preserve more subtle attribute-irrelevant information, ODS with the specific loss in Eq. (9) is proposed.

Semantic Information Entropy Loss (SER). To enable each feature map to reflect as few semantic regions as possible to allow independent editing, the entropy regularization loss is proposed in Eq. (11).

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3289757

Generator. All the three modules are generator-specific operators, yielding the final loss as follows

$$\mathcal{L}_G = \mathcal{L}_{G_{ori}} + \lambda_{TCL} \cdot \mathcal{L}_{TCL} + \lambda_{ODS}\mathcal{L}_{ODS} + \lambda_{SER}\mathcal{L}_{SER} \tag{12}$$

where $\mathcal{L}_{G_{ori}}$ denotes the original loss of GANs. The setting of the regularization parameters, i.e. $\lambda_{ODS} = 10$ and $\lambda_{TCL} = 0.5 \cdot \lambda_{CCL}$ is employed, while λ_{CCL} is set as 15 for StarGAN and STGAN, and 70 for AttGAN. The number of clusters in ODS is set the same as that of semantic regions in SER.

The training process of the proposed generator G is summarized in Algorithm 1.

Algorithm 1 Training process of the proposed algorithm

```
Require: The input image $x$; original label $c$; target label $c'$;
    the set of semantic segmentation masks of $x$: $M$.
    Initialization of $G(\cdot)$ and $D(\cdot)$.
Ensure: Trained $G(\cdot)$ and $D(\cdot)$.
for each all training steps do
    Train discriminator $D(\cdot)$ in the original way.
    # Employ TCL module
    Obtain $x' \leftarrow G(x,c')$, $G(x',c)$, and $G(x',c')$ in Eqs. (1) and (3);
    Calculate $\mathcal{L}_{CCL}$ and $\mathcal{L}_{TCL}$ in Eqs. (1) and (3);
    # Employ ODS module
    Cluster feature maps (FMs) based on hash encoding in Eq. (5) and select the outlier FMs in Eq. (6);
    Perform dropout on the outlier FMs with the loss in Eq. (9);
    # Employ SER module
    Calculate the regularization loss $\mathcal{L}_{SER}$ in Eq. (11);
    #Perform forward calculation and backpropagation
    Calculate the loss $\mathcal{L}_G$ in Eq. (12);
    Update $G(\cdot)$ with gradient backpropagation;
end for
```

IV. Experimental Results

We use StarGAN [2], AttGAN [3] and STGAN[4] as baselines. StarGAN [2] learns mappings among multiple domains. AttGAN [3] encodes conditional information into the decoder of generator. To study the performance of the proposed modules for face attribute editing, CelebFaces Attributes Dataset (CelebA) [47] is used for training and testing, which includes more than 200K images. We pick 4 out of 40 attributes for the experiments. While 198K images of CelebA are used for training, the remaining images are used for testing. To explore the performance of our proposed modules on facial expression editing, the Radboud Faces Database (RaFD) [48] is used for the evaluation with the baseline of StarGAN, where each participant displays eight expressions in three poses and three gaze directions. To study the generalization performance of our model, RaFD, Flickr Faces-HQ (FFHQ) [49] and Labled Faces in the Wild (LFW) [50] are used. FFHQ consists of 70k high-quality images with the size of 1024 × 1024. LFW consists of 13,233 images with a resolution of 250 × 250, where the images are all derived from natural scenes in life. For FFHQ and LFW, the preceding 2k images are resized to be the resolution of 128 × 128 for testing. We get the mask of face semantic segmentation based on BiseNet [46]. The baseline models are trained for 300K steps using images with the size of 128 × 128 as input. We integrate both ODS and SER on the 10th or the 2nd layer of the baseline encoder for StarGAN or STGAN, respectively. For AttGAN, we integrate SER and ODS on the 2nd and 3rd layers, respectively. Meanwhile, StyleGAN-based generator [53] and GuidedStyle [54] are further used to test the performance of the proposed modules on the latent code-based generator.

A. Feature Maps Visualization

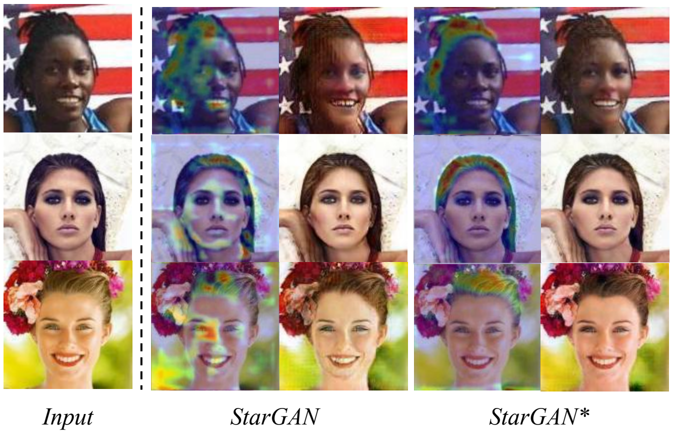

Fig. 8. Feature maps visualization and synthesized results of StarGAN and our proposed model StarGAN* on CelebA.

To shed light onto the hidden layer with our proposed method, we present the synthesized results together with the feature maps of StarGAN and StarGAN*, i.e. StarGAN equipped with the proposed three modules, in Fig. 8. Both the synthesized images and the feature map visualization of StarGAN in Fig. 8 show that the generator could not well edit the attribute-relevant regions. For example, accessories, shadows, and brown skin are made more likely to be confused as hair region. Besides, each feature map of StarGAN may respond to both hair and accessory regions. By contrast, the feature maps of our proposed algorithm enable StarGAN to accurately edit the hair region.

To investigate how our proposed modules represent semantic regions in feature space, we show the resulted feature maps

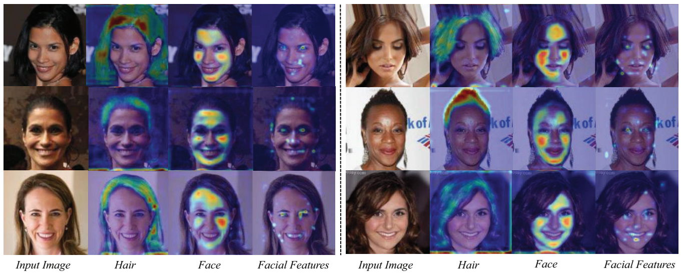

Fig. 9. Visualization results of the feature maps in the 10-th layer of StarGAN* on CelebA, where StarGAN* denotes StarGAN equipped with the proposed three modules.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3289757

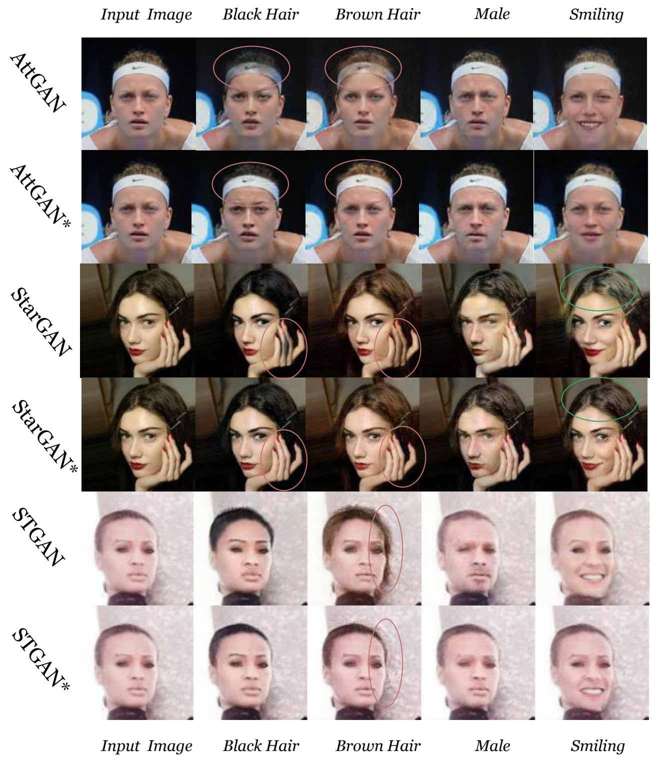

Fig. 10. Results of each baseline and the variant equipped with our proposed three modules on CelebA, where StarGAN* stands for StarGAN+TCL+ODS+SER. The red circles highlight the performances of the proposed algorithm for the suppression of artifacts, and the green circles represent the performances of consistency preservation. For more visual results, please refer to supplemental materials.

in Fig. 9. Fig. 9 shows that the our model well disentangles semantic information in the feature space. The activation regions of each feature map using the proposed modules are more clustered to reflect as few semantic regions as possible, which facilitates independent editing of semantic regions for each feature map in the generator.

B. Visual Results

The visual results of baselines and the algorithms equipped with our modules are shown in Fig. 10. The performances of StarGAN and StarGAN* for black hair translation and smiling editing, AttGAN and AttGAN* for smiling attribute translation are presented in Figs. 11 and 12, respectively. We summarize the following observations.

(1) Reduction of artifacts: In the translation of brown hair attribute, StarGAN and STGAN generate artifacts that appear on background, shadows, and occlusions adjacent to the hair region, which indicates that these generators are not good at decoupling attribute-relevant and attribute-irrelevant features. Although AttGAN reduces obvious artifacts in the background, it still could not eliminate the artifacts of the shadow region next to the hair area in the red circles of Fig. 10, possibly due to the confusion of the hair area with shadows and occlusions. By contrast, AttGAN* can yield the edited images that are more realistic. STGAN generates abnormalities on the background and earrings, while STGAN* can reduce these artifacts.

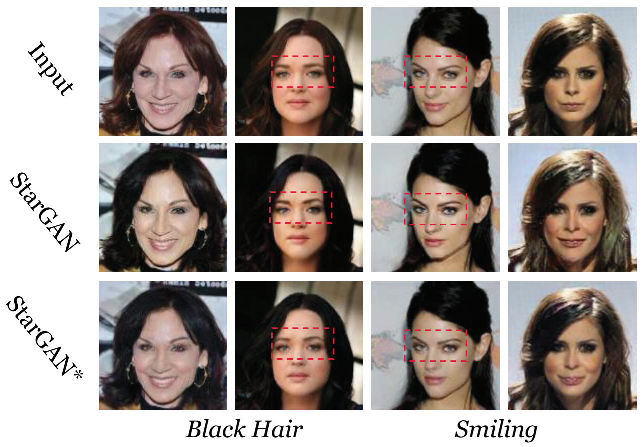

Fig. 11. The performances of StarGAN and StarGAN* on CelebA.

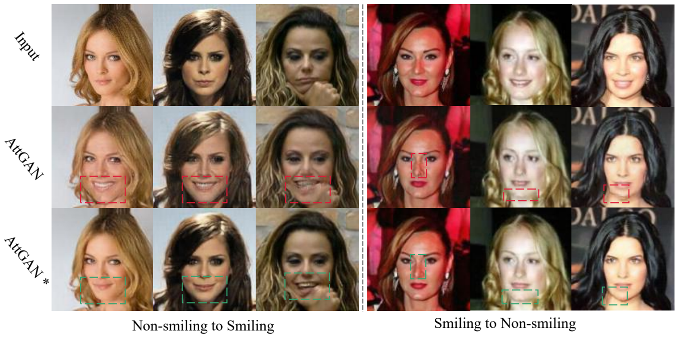

Fig. 12. Visual results of smiling attributes translation on CelebA, AttGAN* stands for AttGAN equipped with the proposed three modules. The preceding four columns show the results of translation from non-smiling to smiling, and the following four columns show the translation from smiling to non-smiling.

(2) Consistency preservation: As shown in Fig. 11, all the three baselines suffer from the undesirable changes of attribute-irrelevant regions. When editing black hair with baselines, the face, eyebrows and eye contours become darker, even their shapes are also changed. Meanwhile, for the smiling attribute translation, the baselines fail to preserve the original hair color. By contrast, StarGAN* can retain not only obvious features, but also those subtle features that are sensitive to noise. As shown in red boxes of Fig. 11, our proposed method enables the generator to better retain subtle attribute-irrelevant information, where earrings, contours of the five sense organs are well preserved. Similar observations can be concluded in Fig. 12, where the proposed AttGAN* better preserves the identity information and synthesizes more realistic five sense organs when translating from non-smiling to smiling.

C. Quantitative Results

To quantitatively evaluate the performance of our methods, we use Fréchet Inception Distance (FID) [51] scores to evaluate the realism of the synthesized image, peak-signal-to-noise ratio (PSNR) [52] that reflects the contrast and lightness of the global image to evaluate the consistency preservation of attribute-irrelevant regions and structural similarity index (SSIM) [52] to measure the preservation of image structure.

FID score. The similarity between the ground truth and synthesized images is measured by FID using visual features obtained with InceptionV3. We present FID scores for four attributes in Table I. Compared with the baselines, our proposed method achieves the lower FID scores, which indicates that it improves the quality of face synthesis.

TABLE I
THE FID SCORES OF THE THREE BASELINE GANS AND THE VARIANTS EQUIPPED WITH THE PROPOSED THREE MODULES ON CELEBA.

Model	Black Hair	Brown Hair	Gender	Smiling
StarGAN	19.18	15.79	16.59	12.67
StarGAN*	**12.62**	**9.79**	**11.26**	**8.44**
AttGAN	14.23	13.61	20.47	11.77
AttGAN*	**11.35**	**10.46**	**11.04**	**9.73**
STGAN	12.55	10.86	9.4	7.66
STGAN*	**7.52**	**6.76**	**6.73**	**6.84**

PSNR. We use PSNR with the mean squared error (MSE) to evaluate the consistency preservation for only attribute-irrelevant regions in Table II. Compared with the baseline GAN, Table II shows that our proposed method can better preserve attribute-irrelevant regions.

TABLE II
THE PSNR RESULTS FOR FOUR ATTRIBUTES ON CELEBA.

Model	Black Hair	Brown Hair	Gender	Smiling
StarGAN	19.71	19.64	26.36	19.96
StarGAN*	**23.12**	**23.6**	**31.39**	**24.28**
AttGAN	21.56	22.67	27.45	23.28
AttGAN*	**24.32**	**26.29**	**32.62**	**27.30**
STGAN	21.74	22.42	31.86	24.65
STGAN*	**25.9**	**26.14**	**32.38**	**28.06**

SSIM. SSIM measures the similarity of two images in terms of luminance contrast and structure. For the evaluation, we use the entire image for the hair color attribute, and the regions that should not be edited for the other attributes. The results of SSIM are shown in Table III, they show that our proposed method can better preserve the contrast and lightness of the global image than the baselines. Especially, STGAN* largely outperforms STGAN for the attributes of black and brown hair by margins of 9.9% and 7.2%, respectively.

D. Ablation Study

1) Feature Maps Visualization: To study the performance of each proposed module for StarGAN, we perform an ablation study in Fig. 13. Because CCL could not directly supervise the preservation of image features in attribute-irrelevant regions, StarGAN generates the artifacts on the face and background regions, while these artifacts are largely reduced by StarGAN* with the aid of our TCL. As shown in the red circles of Fig. 13, ODS pays more attention to the critical regions and suppresses noises in the feature space, thus can prevent uneven facial skin, which is not well addressed by TCL.

TABLE III
THE SSIM RESULTS OF THREE BASELINE GANS AND OUR MODELS FOR FOUR ATTRIBUTES ON CELEBA.

Model	Black Hair	Brown Hair	Gender	Smiling
StarGAN	0.757	0.794	0.944	0.873
StarGAN*	**0.840**	**0.883**	**0.970**	**0.935**
AttGAN	0.774	0.809	0.943	0.881
AttGAN*	**0.885**	**0.841**	**0.961**	**0.919**
STGAN	0.827	0.856	0.961	0.940
STGAN*	**0.909**	**0.918**	**0.978**	**0.956**

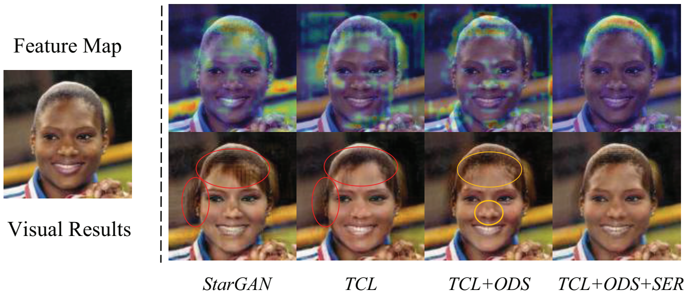

Fig. 13. Feature maps and results of brown hair editing with StarGAN and our proposed variants on CelebA.

When the noises in the background in yellow circles of Fig. 13 are suppressed by StarGAN+TCL+ODS, the noises in the feature space specific to the critical regions such as the forehead, eyes are also suppressed, which is helpful for the proposed variant to better preserve the subtle features of the generated samples. Meanwhile, SER facilitates StarGAN* to generate samples that have more even skin tones and clearer eyes, and feature maps that respond only to the the hair region, while avoiding changing the features of other regions.

2) Visual Result: (i) Consistency preservation and artifacts suppression: Since brown hair is likely to be confused with facial shadows, its editing is a challenge and used for the ablation study of the proposed three modules. Fig. 14 shows the results edited with our proposed modules based on StarGAN.

The results in the 2nd and 3rd columns of Fig. 14 show that StarGAN+TCL can reduce artifacts in the non-facial region and better preserve the features of the input image, e.g. skin tone, eyebrows, eyes, etc. The results in the 3rd and 4th columns show that the proposed ODS can facilitate StarGAN to better preserve the consistency of key facial regions, e.g. face skin color, eyes, and mouth. The results in the 5th column reveal that the proposed SER can accurately locate the hair region for editing, i.e. largely reducing the influence on the attribute-irrelevant regions.

(ii) ODS: The proposed ODS consists of two stages, i.e. the screening of outlier feature maps in the feature space and the dynamic update of the dropout probabilities in Eq. (8). To investigate this dropout update strategy, we performed ablation

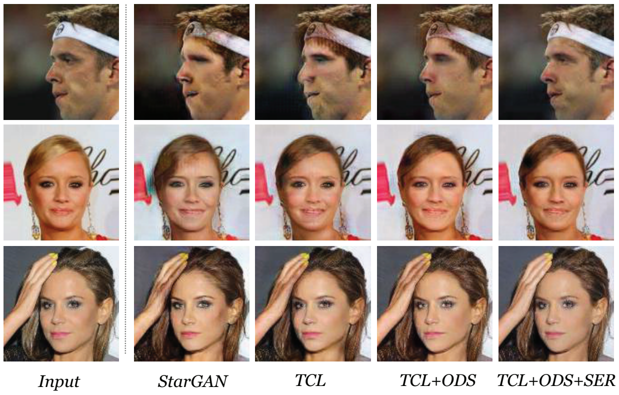

Fig. 14. Results of the brown hair translation on CelebA, where 'StarGAN+TCL' is abbreviated as 'TCL'.

experiments.

The variants of StarGAN+TCL+SER equipped with one of the following ODS strategies are used for the evaluation: a) half-ODS: random dropout with a fixed probability for outlier feature maps, b) all-ODS: drop all outlier feature maps, c) inverse-ODS: the dropout probability is set as $p_{min} + p_{max} - p_i$, where p_i is defined in Eq. (8), d) ODS. The comparison of these variants in terms of visual performances and PSNR are presented in Fig. 15 and Table IV, where brown hair translation is used for the evaluation.

TABLE IV
THE PSNR VALUE OF THE SYNTHESIZED FACE REGIONS FOR THE TRANSLATION OF BROWN HAIR ATTRIBUTE ON CELEBA.

StarGAN	half-ODS	all-ODS	inverse-ODS	ODS
20.14	24.03	25.29	24.29	**26.11**

For the strategy of all-ODS, Fig. 15 shows that the artifacts are not well suppressed, and the attribute-irrelevant regions, e.g. face regions are over-edited possibly due to the loss of much information related to the face. By contrast, the proposed ODS not only suppresses the artifacts to the maximum extent, but also retains more facial information.

Table IV also reveals the effectiveness of the proposed ODS for face regions except hair region in terms of PSNR, where a significant improvement, i.e. 5.97 over the variant without this module is achieved.

TABLE V
THE FID SCORES OF THE PROPOSED THREE MODULES ON STARGAN. FOR MORE RESULTS, PLEASE REFER TO THE SUPPLEMENTAL MATERIALS.

TCL	ODS	SER	Black Hair	Brown Hair	Male	Smiling
✗	✗	✗	19.18	15.79	16.59	12.67
✓	✗	✗	15.58	13.23	14.69	10.75
✓	✓	✗	**11.43**	10.72	11.7	**8.11**
✓	✓	✓	12.62	**9.79**	**11.26**	8.44

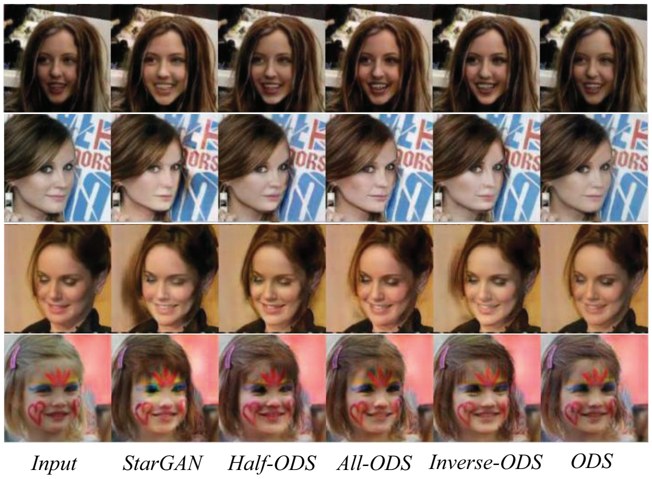

Fig. 15. Results of the brown hair translation on CelebA.

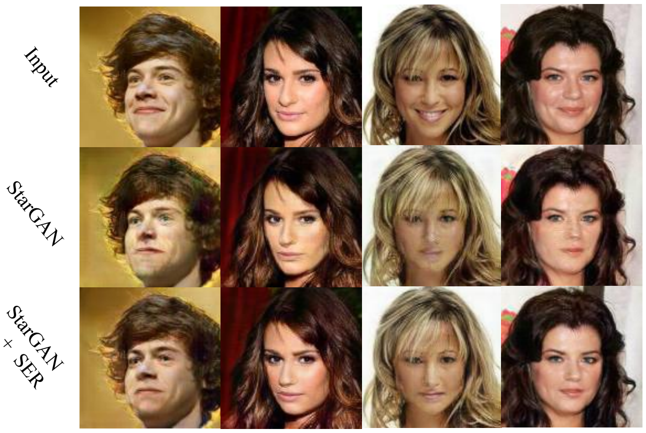

Fig. 16. Visual results of StarGAN and StarGAN+SER for translation from smiling to non-smiling on CelebA. For more visual results, please refer to supplemental materials.

(iii) SER: Since the translation from smiling to non-smiling involves much texture variation in multiple semantic regions, it is challenging for the generator to independently edit five sense organs and facial skin. We use this task to evaluate the performance of SER via ablation study, while the results based on StarGAN are shown in Fig. 16. For StarGAN, artifacts are likely to appear on the edges, eyes, and noses, possibly due to the feature entanglement in these smiling-related regions. By contrast, StarGAN+SER largely suppresses these artifacts and better preserves the fine face details.

TABLE VI
THE PSNR VALUES OF THE PROPOSED THREE MODULES ON STARGAN FOR CELEBA.

TCL	ODS	Black Hair	Brown Hair	Male	Smiling
✗	✗	19.71	19.64	26.36	19.36
✓	✗	21.71	22.19	29.7	23.35
✓	✓	**23.37**	**23.67**	**31.85**	**24.75**

3) Quantitative Result: To study the performance of each proposed module, we conduct ablation studies with the quantitative results for the translations of four attributes in terms of FID in Table V and PSNR in Table VI.

Table V shows that the proposed modules can improve FID scores for most cases, and Table VI shows that both TCL and ODS improve the ability of the generator for preserving

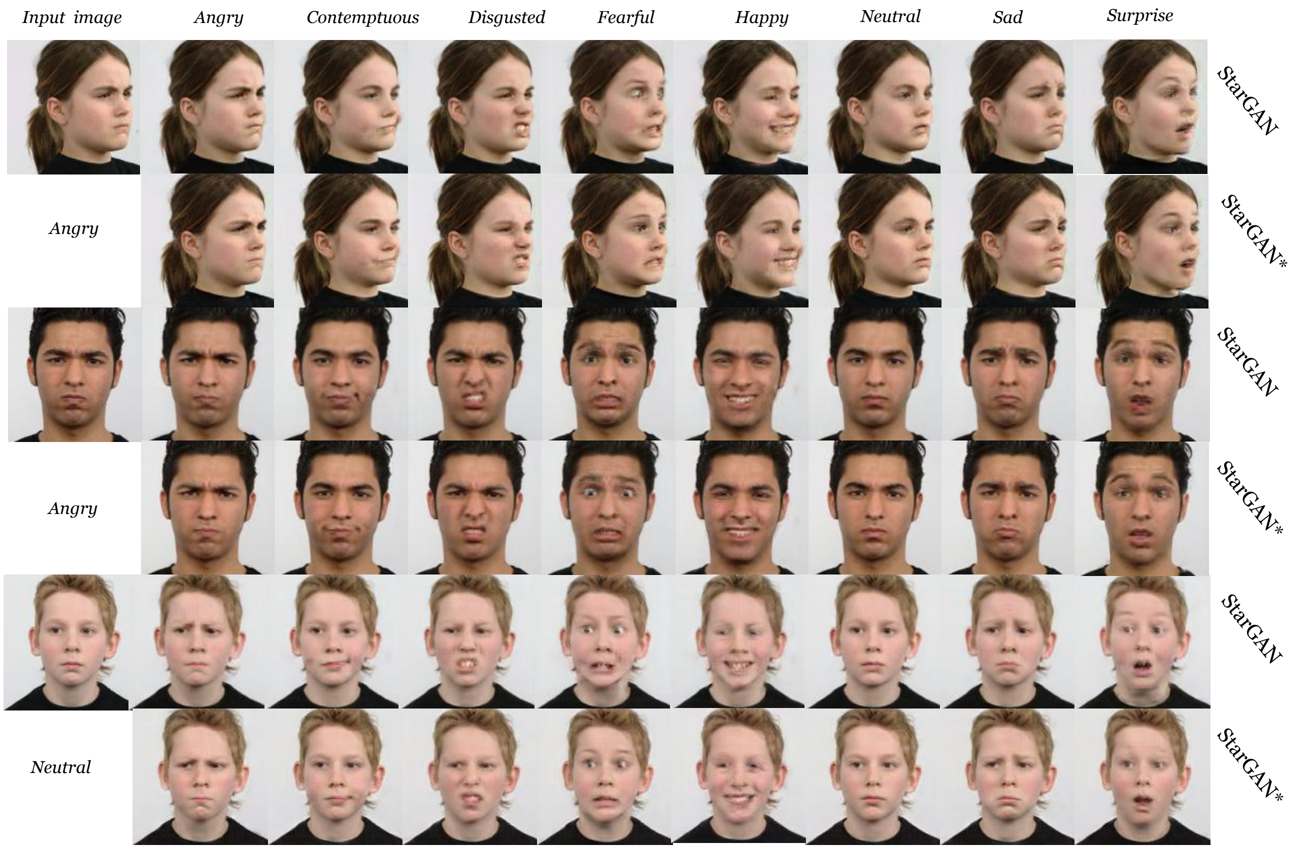

Fig. 17. Visual results of StarGAN and StarGAN* for translation of eight expressions on RaFD.

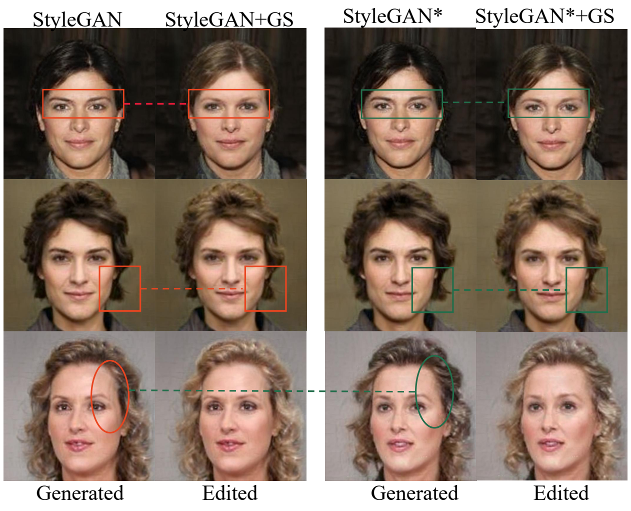

Fig. 18. Visual results of StyleGAN and StyleGAN+ODS+SER (StyleGAN*), where 'GuidedStyle' is abbreviated as 'GS'. 'Edited' denotes the result edited with GuidedStyle for image shown in 'Generated'.

attribute-irrelevant regions, and largely improve baselines for each attribute.

For consistency preservation, StarGAN+TCL outperforms StarGAN by an average margin of 13.96% in terms of PSNR, and StarGAN+TCL+ODS further achieves an improvement of 6.86% over StarGAN+TCL. For AttGAN and STGAN, similar improvements are also observed, which are presented in supplemental materials.

TABLE VII
THE ATTRIBUTE CLASSIFICATION ACCURACY (ACA) AND PSNR VALUES OF THE PROPOSED THREE MODULES ON STARGAN FOR CELEBA.

Model	StarGAN	TCL	TCL&ODS	SER	TCL&ODS&SER
ACA	91.75	89.15	90.55	**95.00**	92.40
PSNR	19.71	21.75	**23.37**	20.32	23.12

4) Attribute Classification Accuracy (ACA): In addition to PSNR, the attribute classification accuracy (ACA) is used to evaluate the editing accuracy of StarGAN and the integration of StarGAN with our proposed modules for editing of black hair in Table VII. For more results of ACA, please refer to the supplemental materials.

Table VII shows that mere TCL decreases ACA for StarGAN, since it is devised to mainly maintain image consistency in the generator, while the consistency preservation of the irrelevant attributes may affect the editing of the target attributes due to feature coupling. By imposing ODS on StarGAN+TCL, one can observe that the performances of ACA and PSNR are improved, and the best PSNR is achieved, since ODS is devised to mainly suppress feature maps with anomalous responses in local key regions. StarGAN+SER achieves the best performance in terms of ACA, since SER enables the generator to edit attribute-related semantic regions while preserving attribute-independent regions after the semantic decoupling. By contrast, our method, i.e. Star-

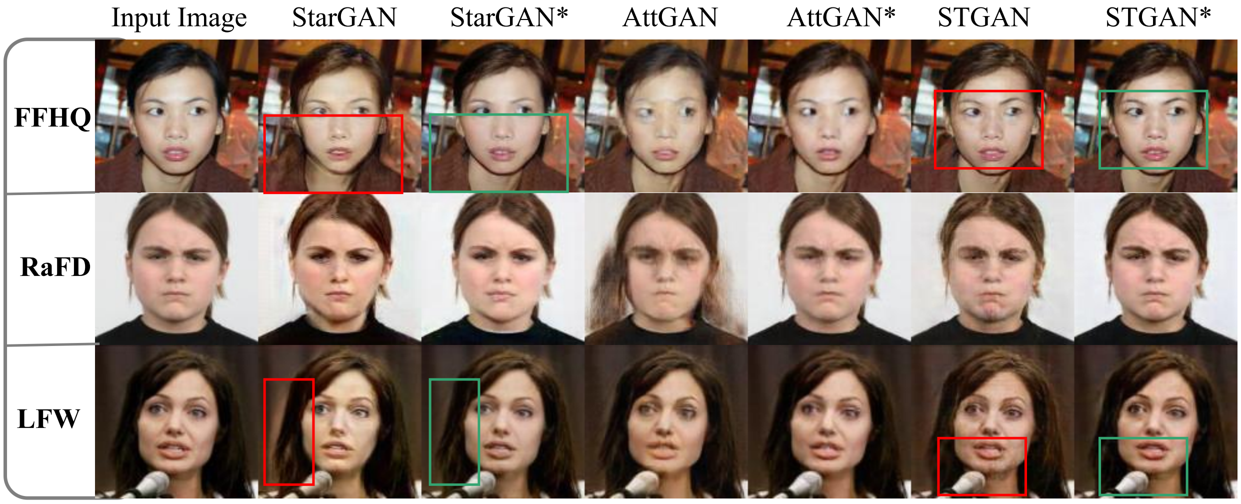

Fig. 19. Visual results of three baselines and our models, where all models are trained on CelebA and tested on FFHQ, RaFD or LFW. For more visual results, please refer to supplemental materials.

GAN+TCL+ODS+SER shows obvious advantage in balancing the performances of ACA and PSNR via both consistency preservation and semantic decoupling.

E. Facial Expression Editing

In addition to facial attribute editing on CelebA, our proposed modules are also evaluated on the task of expression editing, and compared with the baseline of StarGAN. The visual editing results for eight expression categories on RaFD are shown in Fig. 17, and the quantitative results of the FID values as well as expression classification accuracy (ECA) are presented in Tab. VIII.

Fig. 17 shows that our proposed modules can help the generator of StarGAN to produce more natural and smooth expressions while reducing artifacts in facial features, e.g. the fearful expression. Table VIII shows that StarGAN* achieves consistently better FID and ECA than StarGAN, i.e. our modules can facilitate StarGAN to edit expressions more accurately and generate samples with higher quality.

TABLE VIII
THE FID VALUES AND EXPRESSION CLASSIFICATION ACCURACY (ECA).

Angry	Contemptuous	Disgusted	Fearful	Happy	Neutral	Sad	Surprise
FID:StarGAN							
19.84	13.41	14.71	13.82	19.58	10.81	13.19	15.49
FID:StarGAN*							
16.05	**11.05**	**12.84**	**10.66**	**15.83**	**8.32**	**11.21**	**12.09**
ECA:StarGAN							
97.42	96.63	97.62	97.42	97.62	96.231	96.83	96.23
ECA:StarGAN*							
98.21	**97.82**	**98.61**	**98.02**	**98.61**	**97.42**	**97.62**	**98.02**

F. Exploration of StyleGAN-based generator

To further study the performance of the proposed modules on latent code-based generator, the StyleGAN-based generator [53] is used for testing. The generator in StyleGAN is unconditional and manipulates the latent code to synthesize images, and can synthesize images of high-quality with few artifacts. In this toy experiment, we explore the role of our proposed method by fine-tuning a StyleGAN-based generator [53] equipped with our ODS and SER for CelebA. Note that the proposed TCL is not applicable here, since its complementary cycle consistency loss (CCL) can hardly be used in Style-based generators. Based on the synthesized images, GuidedStyle [54] is used to achieve face attribute editing. The editing results of brown hair are shown in Fig. 18, while the quantitative results of FID and PNSR, together with the detailed experimental setup, are presented in the supplementary material.

TABLE IX
THE FID VALUES OF THREE BASELINE GANS AND OUR MODELS FOR BROWN HAIR TRANSLATION, WHERE ALL MODELS ARE TRAINED ON CELEBA AND TESTED ON FFHQ, RAFD OR LFW.

Dataset	StarGAN	StarGAN*	AttGAN	AttGAN*	STGAN	STGAN*
FFHQ	32.27	**14.62**	21.16	**10.89**	14.89	**9.58**
RaFD	42.84	**18.75**	60.96	**12.14**	32.15	**14.75**
LFW	43.61	**16.54**	23.28	**9.36**	12.82	**10.18**

Fig. 18 shows that our modules facilitate StyleGAN to better preserve the irrelevant attributes. As shown in the red boxes of the 1st row and the 2nd column, the eyebrow is changed and the color of the eye-surrounding region is undesirably changed to brown, while the generator equipped with our modules better preserves them. As shown in green box in the 2nd row, based on the images edited by our generator, GuidedStyle can successfully change the marginal region of the hair to brown, i.e. editing the hair area more accurately. We speculate that the proposed SER enables each feature map in the generator to represent as few semantic regions as possible, which allows GuidedStyle to edit the hair region independently and accurately. As shown in the red ellipse in the 1st column, artifacts were produced in the images generated by StyleGAN, while they are reduced in the images generated using our modules. This improvement may due to the suppression of the abnormal feature maps achieved by the proposed ODS during the fine-tuning.

G. Generalization Performances

To investigate the generalization performance of our method, we perform the training on CelebA and evaluate the performance on FFHQ, RaFD or LFW, where four attributes are used for the evaluation. While the translation for the brown hair is the most challenging for the three baselines, the generalization performance for this task appears to be worse and more unstable than those for the other attributes. Thus, we only present the results related with the brown hair attribute, i.e. the FID values are shown in Table IX and the visual results are presented in Fig. 19, while the results specific to the other attributes are appended in the supplemental materials.

Table IX shows that our proposed modules help the generator to achieve consistently better performance for brown hair editing. Especially, the FID value of our model is only around 1/5 of AttGAN on RaFD. These improvements of generalization performances maybe caused by the suppression of outlier feature maps that reduces the possibility of overfitting, or the proposed independent editing that alleviates the attribute entanglement. Fig. 19 shows that our modules enable the baselines to largely reduce the artifacts shown in red rectangles, and more detail features can also be well preserved for this task.

V. Conclusion and Discussions

In this work, we give insight into the working mechanism of the generator in GAN and propose three modules to reduce its artifacts. The proposed generator has three main novelties. First, we proposed a target consistency loss, which complements cycle consistency loss with end-to-end restriction module, while the blur of the synthesized images can be largely reduced. Second, we introduced dynamic suppression of outlier feature maps to preserve subtle feature information. Third, we proposed semantic entropy regularization (SER) to disentangle the representations of the feature semantics, and enable each feature map in the generator to independently edit different attributes. In terms of generation visualization, three quantitative metrics and the generalization performance on four face datasets, our strategies can generate more realistic samples and better preserve global image cues compared with three baseline GANs. While the proposed modules are revealed to be effective for the image-to-image based paradigm, the quantitative results of the toy experiment for latent code-based generator in the supplemental material show that our modules need to be slightly adjusted for better adaption to StyleGAN-based generators [53], [54], and other tasks, e.g. expression editing.

Acknowledgment

The work was supported by Natural Science Foundation of China under grants no. 62276170, 82261138629, the Science and Technology Project of Guangdong Province under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grants no. JCYJ20190808165203670, JCYJ20220531101412030.

References

[1] I. Goodfellow, J.-P. Abadie, M. Mirza, B. Xu, D.-W. Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672-2680.

[2] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8789–8797.

[3] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, 2019.

[4] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 3673–3682.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2223–2232.

[6] J.-G. Kwak, D. K. Han, and H. Ko, "CAFE-GAN: arbitrary face attribute editing with complementary attention feature," in *Proc. Eur. Conf. Comput. Vision*, 2020, pp. 524–540.

[7] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 8188–8197.

[8] D. Kim, M. A. Khan, and J. Choo, "Not just compete, but collaborate: Local image-to-image translation via cooperative mask prediction," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 6509–6518.

[9] R. B. Arantes, G. Vogiatzis, and D. R. Faria, "Csc-gan: Cycle and semantic consistency for dataset augmentation," in *Int. Symposium on Visual Computing*, 2020, pp. 170–181.

[10] Z. Zheng, Y. Bin, X. Lu, Y. Wu, Y. Yang and H. T. Shen, "Asynchronous generative adversarial network for asymmetric unpaired image-to-image translation," *IEEE Trans. Multimedia*, 2022.

[11] X. Hou, X. Zhang, H. Liang, L. Shen and Z. Ming, "Lifelong Age Transformation with a Deep Generative Prior" *IEEE Trans. Multimedia*, 2022.

[12] H. Zhang, Z. Zhang, A. Odena, and H. Lee, "Consistency regularization for generative adversarial networks," in *Proc. Int. Conf. Learn. Representations*, 2020.

[13] Z. Zhao, S. Singh, H. Lee, Z. Zhang, A. Odena, and H. Zhang, "Improved consistency regularization for gans," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11033–11041.

[14] H. Emami, M. M. Aliabadi, M. Dong and R. B. Chinnam, "SPA-GAN: Spatial Attention GAN for Image-to-Image Translation" *IEEE Trans. Multimedia*, vol. 23, pp. 391-401, 2021.

[15] T. Chen, S. Wu, X. Yang, Y. Xu and H. -S. Wong, "Semantic Regularized Class-Conditional GANs for Semi-Supervised Fine-Grained Image Synthesis," *IEEE Trans. Multimedia*, 2021.

[16] X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang, "Improving the improved training of wasserstein gans: A consistency term and its dual effect," in *Proc. Int. Conf. Learn. Representations*, 2018.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Mach. Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] J. Kim, S. Cha, D. Wee, S. Bae, and J. Kim, "Regularization on spatio-temporally smoothed feature for action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 12103–12112.

[19] G. Kang, J. Li, and D. Tao, "Shakeout: A new approach to regularized deep neural network training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1245–1258, 2017.

[20] Y. Tang, Y. Wang, Y. Xu, B. Shi, C. Xu, C. Xu, and C. Xu, "Beyond dropout: Feature map distortion to regularize deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5964–5971.

[21] Z. Wen, H. Wu, W. Xie, L. Shen, and J. Duan, "Group-wise feature orthogonalization and suppression for GAN based facial attribute translation," in *Int. Conf. Pattern Recongnit.*, 2020, pp.3767–3774.

[22] Z. He, M. Kan, J. Zhang and S. Shan, "PA-GAN: Progressive attention generative adversarial network for facial attribute editing," *arXiv preprint arXiv:2007.05892*, 2020.

[23] Z. He, M. Kan and S. Shan,"EigenGAN: Layer-wise eigen-learning for GANs," in *Proc. IEEE Int. Conf. Comput. Vision*, 2021, pp.14408–14417.

[24] A. H. Liu, Y. Liu, Y. Yeh, and Y. F. Wang, “A unified feature disentangler for multi-domain image translation and manipulation,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2595–2604.

[25] G. Yang, N. Fei, M. Ding, G. Liu, Z. Lu, and T. Xiang, “L2m-gan: Learning to manipulate latent space semantics for facial attribute editing,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 2951–2960.

[26] W. Peebles, J. Peebles, J.-Y. Zhu, A. Efros, and A. Torralba, “The hessian penalty: A weak prior for unsupervised disentanglement,” in *Proc. Eur. Conf. Comput. Vision*, 2020, pp. 581–597.

[27] Y. Wei, Y. Shi, X. Liu, Z. Ji, Y. Gao, Z. Wu, and W. Zuo, “Orthogonal jacobian regularization for unsupervised disentanglement in image generation,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2021, pp. 6721–6730.

[28] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *Proc. Int. Conf. Neural Inf. Process. Syst.* , 2016, pp. 2180–2188.

[29] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2337–2346.

[30] Y. Wang, Y.-C. Chen, X. Zhang, J. Sun, and J. Jia, “Attentive normalization for conditional image generation,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 5094–5103.

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1125–1134.

[32] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier gans,” in *Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.

[33] G. Dorta, S. Vicente, N. D. F. Campbell, and I. J. A. Simpson, “The GAN that warped: Semantic attribute editing with unpaired data,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 5355–5364.

[34] T. Xiao, J. Hong, and J. Ma, “Elegant: Exchanging latent encodings with gan for transferring multiple face attributes,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 168–184.

[35] F. Ma, G. Xia, and Q. Liu, “Spatial consistency constrained gan for human motion transfer,” *IEEE Trans. Circuits Syst. Video Technol.*, 2021.

[36] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao, “Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2427–2436.

[37] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[38] X. Wang, K. Jin, Y. Kong, C. L. P. Chen and Y. Cheng,“Discriminator-quality evaluation GAN,” *IEEE Trans. Multimedia*, 2022.

[39] Z. Zhang, M. Li, H. Xie, J. Yu, T. Liu and C. W. Chen, “TWGAN: Twin Discriminator Generative Adversarial Networks,” *IEEE Trans. Multimedia*, vol. 24, pp. 677-688, 2022.

[40] K. Liu, W. Tang, F. Zhou, and G. Qiu, “Spectral regularization for combating mode collapse in GANs,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 6382–6390.

[41] Z. Li, R. Tao, P. Xia, H. Chen, and B. Li, “A Systematic Survey of Regularization and Normalization in GANs,” *ACM Comput. Surv.*, 2022.

[42] B. Liu, Y. Zhu, Z. Fu, G. De Melo, and A. Elgammal, “Oogan: Disentangling gan with one-hot sampling and orthogonal regularization,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4836–4843.

[43] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 5549–5558.

[44] V. Hulle, M. M, “Self-organizing Maps,” in *Handbook of Natural Computing*, 2012, pp. 585–622.

[45] H. Liu, H. Wu, W. Xie, F. Liu, and L. Shen, “Group-wise inhibition based feature regularization for robust classification,” in *Proc. IEEE Int. Conf. Comput. Vision*, October 2021, pp. 478–486.

[46] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Bisenet: Bilateral segmentation network for real-time semantic segmentation,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 325–341.

[47] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 3730–3738.

[48] L. Oliver, D. Ron, B. Gijsbert, W.-Daniel. HJ, H.-S. T and V.-K. AD, “Presentation and validation of the Radboud Faces Database,” in *Cognition and emotion*, 2010, pp. 1377–1388.

[49] K. Tero, L. Samuli and A. Timo “A style-based generator architecture for generative adversarial networks,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 4401–4410.

[50] H.-G. B, M. Marwan, B. Tamara and L.-M. Eric“Labeled faces in the wild: A database forstudying face recognition in unconstrained environments,” in *Eur. Conf. Comput. Vision Workshop on Faces in Real-life Images.*, 2008.

[51] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[52] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in *Int. Conf. Pattern Recongnit.*, 2010, pp. 2366–2369.

[53] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 8110–8119.

[54] X. Hou, X. Zhang, H. Liang, L. Shen, Z. Lai, and J. Wan, “Guided-Style: Attribute knowledge guided style manipulation for semantic face editing,” *Neural Netw.*, vol. 145, pp. 209-220, 2022.