IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 January 2023, accepted 1 March 2023, date of publication 8 March 2023, date of current version 14 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3253935

RESEARCH ARTICLE

Correlation-Concealing Adversarial Noise Injection for Improved Disentanglement in Label-Based Image Translation

SEONGUK PARK, JOOKYUNG SONG, DONGHOON HAN, AND NOJUN KWAK, (Senior Member, IEEE)
Department of Intelligence and Information, Seoul National University, Seoul 08826, South Korea
Corresponding author: Nojun Kwak (nojunk@snu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) through the Korean Government under Grant 2021R1A2C3006659, and in part by the Institute for Information & communication Technology Planning & evaluation (IITP) through the Korean Government under Grant 2021-0-01343 and Grant 2021-0-00537.

ABSTRACT Deep learning models in image synthesis have proven their applicability in various image translation areas. However, although the synthesized image may reflect the user's intention, some of its properties may be different from those of real images. In this study, we introduce an undesirable property that we discovered in the multi-domain label-based image translation techniques: Once the image is translated to one domain, the translated image cannot be adequately translated again to another domain. We refer to this problem as the failure of recursive translation, and analyze this phenomenon from the viewpoint of attribute disentanglement and establish a hypothesis: Unlabeled or unknown attributes that are correlated with the direction of translation hinder the network from learning the correct direction of translation. Based on our hypothesis, we also devise a solution that endows the generator with the power of recursive translation, which is achieved by injecting additive perturbations during model training. Our method is simple and easy to implement on various translation models without requiring much hyperparameter adjustment. Beyond enabling recursive translation, it is worth noting that solving the recursive translation problem improves the disentanglement of single translations, which eventually strengthens its practicability.

INDEX TERMS Adversarial attack, image translation, GAN, disentanglement.

I. INTRODUCTION

Among the various tasks of GAN [1], image-to-image translation task is to translate an image from a source domain to a designated target domain. Image-to-image translation models can be categorized by various criteria, and we divide them into two based on how domains are defined: explicit-domain translation which uses domain labels and implicit-domain translation which explores latent vectors to define domains.

In this paper, we introduce an interesting phenomenon that we discovered: explicit-domain translation commonly suffers from one problem; once the model translates an image into one domain, the translated image cannot be adequately translated into another domain. We coin this problem as the '**failure of recursive translation**' or the '**lack of transparency**'. The problem is due to the model's deficiency of attribute disentanglement. A target domain has unknown correlations with other unintended attributes that result in unexpected biases, and the model translates the input images following the direction of the unknown correlation. As a result, although the translated image may appear as if it belongs to the target domain in human perception, it actually may not lie in the distribution of real images belonging to the target domain. Thus, when a model deals with an already-translated image, that image is out of the input domain distribution, prohibiting the model from translating the image to the right direction. It is worthwhile to address the recursive translation problem from three perspectives. 1) Mitigating this from an academic point of view would naturally lead to a more reliable model. 2) From a practical point of view,

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

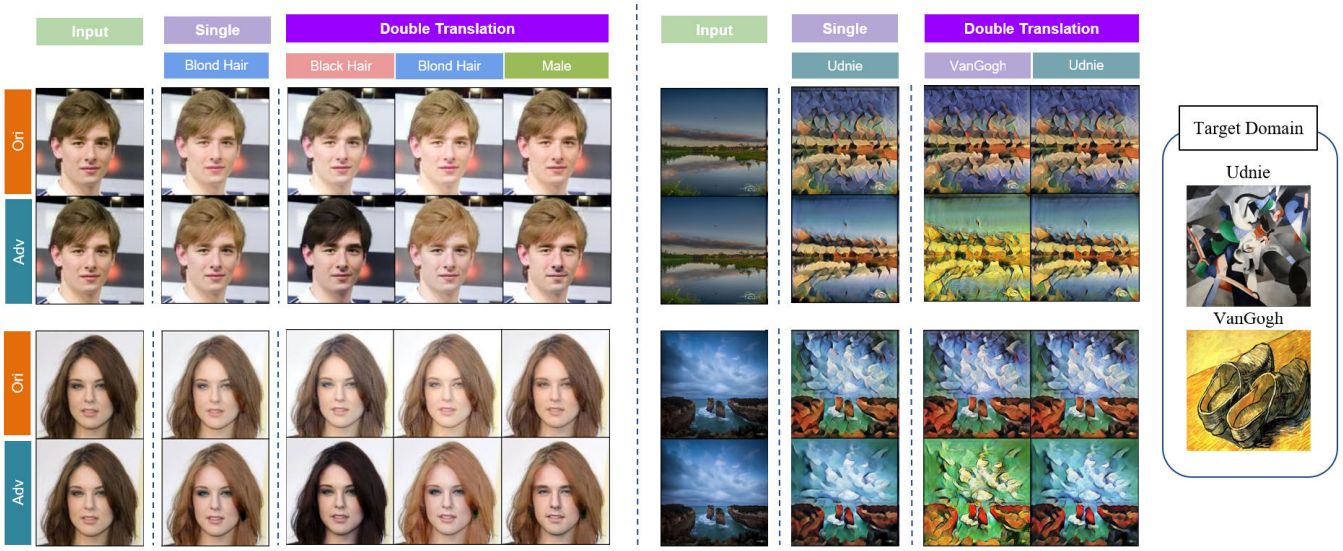

FIGURE 1. Results of single image translation and double image translation using RelGAN. The 'Ori' rows indicate the results of the original method, and 'Adv' rows show the results when our method is applied. We used single translated image $x^1 = G(x, v)$ as the input of the double translation, i.e, $x^2 = G(x^1, v)$.

removing the bias will lead to an improved model with better disentanglement. 3) Failure of recursive translation can be a threat to those trying to deploy image translation methods for data augmentation.

In order to solve this problem, we propose a method that injects adversarial perturbations into the input during the training. Our *adversarial attack method* on GAN is based on the **projected gradient descent** (PGD) [2], which iteratively injects an adversarial perturbation to the input image by maximizing the mean squared error (MSE) between the initially generated image and the generated image from the perturbed image.

Some examples of the recursive translation problem and our solution can be found in Fig. 1. In the original model (denoted as 'Ori'), when the desired translation is '*changing to blonde hair*', the image's overall brightness changes as well. In this case, the unintended attribute would be '*increasing brightness*', which is correlated with an intended attribute '*changing to blonde hair*'. Since the original model ('Ori') cannot disentangle these two attributes, it makes the whole image look paler, whereas our 'Adv' model exhibits alleviated behavior.

Our main contributions can be summarized as:

- To the best of our knowledge, we are the first to unveil the '**failure of recursive translation**' problem, that current label-based explicit image translation models cannot properly translate images recursively.
- Examining the cause of this problem, we establish the '**undesired correlation hypothesis**', that the model is vulnerable to learning the undesired correlations with unintended attributes that are inherent in the data.
- Harnessing the power of adversarial attack, we propose a modified PGD attack, which injects randomness to the direction of the image translation during training phase. Our method is simple to implement, but powerful enough to translate images into the right target distribution.
- Our method not only enables **robust recursive translations**, but also shows **better disentanglement** results on single translation, which improves its practicability.

Our work has strength in that our method fundamentally rectifies the model's incorrect behavior, and also gains controllability over the model by disentangling attributes. In the discussion section and supplementary materials, we provide more in-depth analysis on additional experiments with accompanying explanations that can suggest new directions of future researches.

II. RELATED WORK

Image-to-image translation refers to the task of translating an image from one domain to another. In this context, a *domain* refers to a group of images with similar characteristics, such as gender, hair color, and facial expressions pre-defined by a human.

Explicit-domain translation is a subset of image translation, where the target domains are explicitly defined (*e.g.* by labeling the painting styles, labeling facial attributes, semantic maps, a single style image, etc). Some early works include Pix2Pix [3] which uses supervised labels to learn the mapping from the input image to the output image. Reference [4] proposed perceptual loss that translates an image toward a single style image, and used a residual [5] generator. CycleGAN [6] and DiscoGAN [7] enabled learning the mapping between two domains in an unpaired manner, introducing the cycle-consistency loss. For **label-based translations**, IcGAN [8], StarGAN [9], and SingleGAN [10] use hard

labels representing the target domain as input conditions to the generator network. These networks use a single generator network to translate to multiple domains. Furthermore, RelGAN [11] pushed ahead by training a generator with relative attributes based on the change between source and target domains, enhancing the attribute interpolation performance. Sym-parameterized GAN [12] proposes a technique of translating images into any mixed-domain, using Sym-parameter and various mixed losses.

Adversarial attack is known as a dangerous security threat by exploiting the vulnerability of deep models for malicious purposes. The objective of an adversarial attack is to create unrecognizably small distortions η to the input x which maximize the difference between the outputs of input x and perturbed input $x + \eta$. The optimal perturbation used in the attack is expressed as

$$\eta^* = \underset{\eta \in S}{\operatorname{argmax}}\, L(x, x + \eta), \tag{1}$$

where $S = \{\eta \mid \|\eta\|_\infty \leq \epsilon\}$ and L is the adversarial loss, which denotes the distance between the outcomes of x and $x + \eta$. To find a stronger adversarial example, numerous gradient-based methods have been proposed. Reference [13] proposed Fast Gradient Signed Method (FGSM), which adds noise in the same direction as the gradient of the cost function. Projected Gradient Descent (PGD) attack [2] is one of the popular methods among first-order gradient-based attacks, which applies the gradient descent and the projection for multiple steps so that the perturbed image lies in the constrained region S. PGD can be expressed as

$$x'_0 = x,\ x'_{t+1} = \Pi_S(x'_t + \alpha\ \text{sign}(\nabla_x J(x'_t, y))), \tag{2}$$

where x'_t denotes the adversarial example at the t^{th} iteration, J is the target loss function, y is the ground truth label for x, Π_S is the projection operation to the constrained region S, and α is the step size. PGD also suggested that adversarial training via solving a minimax optimization problem can boost the model's robustness. Some studies tried cooperating GAN with adversarial attack [14], [15], [16], [17], and tried attacking the image synthesis process of GANs [1].

III. PROBLEM IN RECURSIVE TRANSLATION

In this section, we first introduce the definition of the *recursive translation problem* that we discovered in the label-based image-to-image translation GAN models. Next, we analyze the cause by proposing a hypothesis of undesired correlation between attributes. Throughout this paper, the word 'attribute' of the translation refers to a vector that represents the change of domains. For example, translating from domain A to domain B is expressed by $\vec{Att}_{AB}$.

A. DEFINITION OF RECURSIVE TRANSLATION PROBLEM

We observed the phenomenon that the existing label-based multiple domain image translation models are not able to translate already translated images to another domain. We define this as the *failure of recursive translation problem.* Suppose a generative network $G(x_d, v_d)$ that is trained to conduct multiple domain image translation, 3 domains for example, $d \in \{A, B, C\}$. Given an input image x_A which belongs to domain A, we can translate it into domain B using the generator G conditioned on the target label v_B. The output can be expressed as $x_B^1 = G(x_A, v_B)$, where its intended attribute is $\vec{Att}_{AB}$.[1] If one wants to translate the output image recursively to domain C, then $x_C^2 = G(x_B^1, v_C)$, where the superscript denotes the number of recursive translations. With a transparent generator, the crafted image x_B^1 and x_C^2 would lie on the distribution of B and C respectively. However, according to our observation, with existing methods, the resulting $x_C^2 \notin C$ for human eyes although $x_B^1 \in B$. To the best of our knowledge, we are the **first** to point out this problem. This phenomenon can be observed in Fig. 1. Following the figure, the double translation refers to translating a single translated image in the second column (i.e, x_B^1) into other domains (i.e, C). The outputs of the existing models (denoted as 'ori') look plausible in a single translation, but the outputs of the double translation do not.

1) UNDESIRED CORRELATION HYPOTHESIS

We hypothesize about the cause of the failure of recursive translation stated above, and name it *Undesired Correlation Hypothesis*: The direction of the image translation has a non-zero correlation with unintended attributes that lie in the training dataset so that the model learns these undesired correlations with unintended attributes during training, resulting in an unintended bias as a result of the translation.

Suppose there are three predefined domains $d \in \{A, B, C\}$ (e.g. Blond, Male in Fig. 1) and an undefined domain D (e.g. Pale face) which does not belong to the domain label space of the training dataset. Also, suppose a situation in which $\vec{Att}_{AB}$ has a positive correlation with the attribute vector $\vec{Att}_{AD}$. Then, the sample $x_B^1 = G(x_A, v_B)$ can have been translated toward domain D. As a result, the distribution of resulting x_B^1 does not match with the real distribution of domain B. This situation is depicted in Fig. 2(A), where Fake B distribution is biased toward the unintended domain D. Now, let us think about the situation of recursive translation, $x_C^2 = G(x_B^1, v_C)$. During the training, $G(\cdot)$ has only received real images, x_A and x_B, as its inputs, but it is not aware of x_B^1 as its inputs. Therefore, $G(\cdot)$ cannot properly cope with x_B^1, which lies in unseen distributions.

IV. PROPOSED METHOD

In Fig. 2(A), we want to fit the distribution of Fake B ($x_B^1 = G(x_A, v_B)$) to the distribution of domain B: $G^* = \text{argmin}_G ||x_B^1 - x_B^*||$. For that, the model has to disentangle the intended attribute vector $\vec{Att}_{AB}$ from the unintended attribute vector $\vec{Att}_{AD}$ that are correlated. Ideally, under the undesired

[1] Informally speaking, it can be considered as $\vec{Att}_{AB} = x_B^* - x_A$, where the superscript $*$ denotes the desired sample.

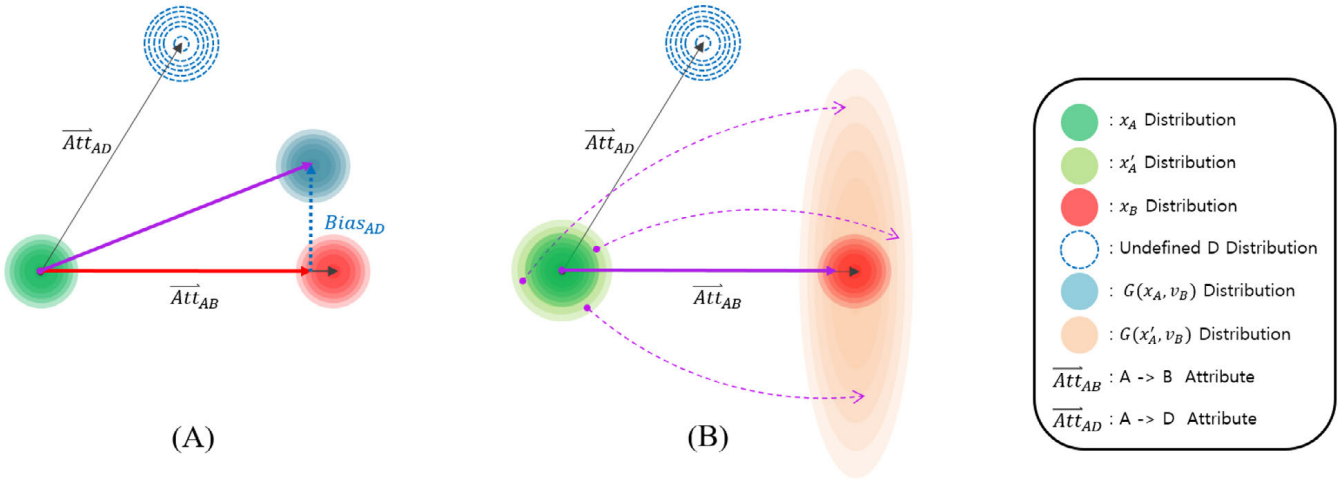

FIGURE 2. **(A): Illustration of our hypothesis about the cause of *failure of recursive translation*. The Domain *B* refers to an intended target domain, and Domain *D* refers to an undefined domain. The purple arrow refers to expected translation direction. (B): Illustration of our proposed solution. The expected translations of images during the training phase is depicted by dashed purple arrows. Note that the distribution of x'_A is actually not very different from x_A, since the magnitude of an adversarial perturbation is constrained to be very small.**

correlation hypothesis, the successful disentanglement would result in the removal of $Bias_{AD}$, and we can solve the problem.

A. MOTIVATION

Numerous solutions may exist for disentangling the unintended attributes, and let us suppose that the solutions can be largely sorted into two methodologies: Either training a model to learn to de-correlate, or making the model **correlation-incognizable** during the training. The main reason the former disentangling is difficult is that it is complicated to directly apply some explicit formulation for disentanglement, because the unintended attributes are implicit, rather than clearly defined as a form of target label. Therefore, in the task of image-to-image translation, we decided to focus on the latter approach that makes the data-inherent bias imperceptible for the generator. As a way to make the model unable to recognize the data-inherent correlation, we propose to grant randomness to the direction of image translations $G(x', v)$ other than the intended attribute. Similar approach exists in the *information perturbation* [18], which shares the motivation of randomizing information to prohibit the trained model from learning unintended tendency that lies in the data.

B. RANDOM NOISE FOR DIVERSE TRANSLATION

A straightforward way to grant randomness during the GAN training is adding noise to the training samples, but we found that there also exist some accompanying limitations that cannot be resolved using random noises. Fig. 3 is the translation result of StarGAN when Gaussian random noises are added to the input images during the network training, and clean images are used for the testing. Whereas the result of original multi-domain image translation models fails to translate recursively in Fig. 1, the model trained noisy input

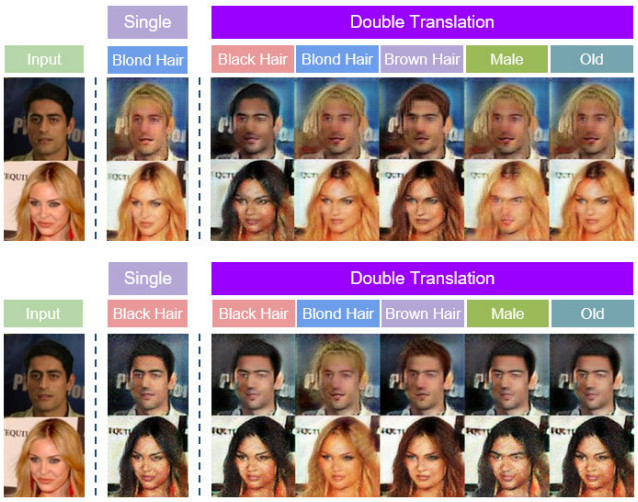

FIGURE 3. **The test result of StarGAN model with Gaussian noise added during the training. Though the results are noisy, the recursive translation problem is mitigated.**

succeed in changing attributes such as hair color and facial expressions.

Injecting Gaussian random noise seems to enable recursive translation, which supports our conjecture that the randomization will make the direction of image translation diverge, which in turn will alleviate the correlation between the intended and unintended attributes so that it will ultimately enable the recursive translation. However, the performance (in terms of visual quality) of the single translation is severely degraded. The reason behind this is the mismatch of the input distribution between the noisy images for training and the clean images for testing. Since the input images are added with Gaussian noises, there are no guarantees that the clean

and noisy data have similar distributions, so the model trained on only noisy data cannot handle clean data properly.

Now, we are to make the direction of image translation diverge by randomization, while not changing the input distribution during the training of the generator. Thus, the distribution of the output $x_B'^1$ should be widened while maintaining the distribution of the input. Fortunately, there exists a way that kills two birds with one stone: 'Adversarial attack', which makes the output of the target network diverge by adding subtle noise to the input.

C. MODIFIED PGD ATTACK FOR DIVERSE TRANSLATION

Inspired by PGD attack [2] which is prominent in adversarial perturbation, we devised a method of diversifying translated images. Our method maximizes the distance between $x^1 = G(x, v)$ and $x'^1 = G(x', v)$ where x is the input and x' is a perturbed image. In Eq. (1), if we use the Mean Squared Error (MSE) as the adversary loss L and then replace the original loss J in Eq. (2) with L, it becomes

$$x'_0 = x, x'_{t+1} = \Pi_S(x'_t + \alpha \ \text{sign}(\nabla_x L(G(x'^1_t), G(x^1)))) \quad (3)$$

and finally an adversarial input x' is obtained by $x' = x'_T$ for some predefined iterations T. Precisely, our method is quite different from the PGD adversarial training, as the PGD training is to solve the minimax optimization problem whose outer minimization directly minimize the adversary loss given by the outer attack problem ($L = J$), whereas our adversary loss in Eq. (1) does not aim to maximize the outer minimization loss J ($L \neq J$). This implies that our method does not conflict with the training of GAN.

After generating x', **all** the real input samples x and fake output samples x^1 are replaced with x' and x'^1 during the model training. Specifically, in the field of image-to-image tasks, it is common to use the adversarial loss of GAN and the domain classification loss jointly. For example, in StarGAN, these can be modified using our method as:

$$\mathcal{L}_{adv} = \mathbb{E}_{x'}[\log D(x')] + \mathbb{E}_{x',v}[\log(1 - D(G(x', v)))] \quad (4)$$

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x',v}[-\log D_{cls}(v \mid x')] \quad (5)$$

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x',v}[-\log D_{cls}(v \mid G(x', v))] \quad (6)$$

where $D_{cls}(\cdot)$ refers to the domain classifier head of the discriminator [19].

The expected aspect of our proposed training is illustrated in Fig. 2(B). Since we perturb the input with a norm-constrained noises, the variance of the input domain distribution will not increase as much as the Gaussian random noises. On the other hand, since the input perturbation maximizes the MSE loss in Eq. (3), the variance of the output distribution (Fake B) will increase largely. The reason behind the choice of MSE as the adversary loss is that we do not want to interrupt the originally intended attribute, which corresponds to our motivation. Here, the distribution of Fake B will not vary much toward the direction of the intended attribute, because the generator is trained to make the classifier head of the discriminator correctly classify $x'^1_B = G(x'_A, v_B)$. On the other hand, the direction of attributes other than A to B (intended attribute) will be severely randomized, so that it is difficult for the generator to perceive any correlations with other attributes that inhere in the training dataset. It is important to note that in Fig. 2(B), the mean value of Fake B distribution will not be biased much toward the direction of $\overrightarrow{Att}_{AD}$, because the generator could not learn the correlations. We will discuss this later in detail.

As a result, only the intended attribute vector is learned by the generator, and the distribution of Fake B is more likely to be close to that of the target domain, and ideally, the recursive translation $x^2_C = G(x^1_B, v_C)$ will be equivalent to $x^1_C = G(x_B, v_C)$. With the generator trained using adversarial input x', the results of single translation and double translation are both improved impressively.

Algorithm 1 Pseudocode of the Modified PGD Attack

Require: Training samples x, perturbation bound ϵ, generator G, number of steps T, perturbed sample x'
$S = \{\eta \mid \|\eta\|_\infty \leq \epsilon\}$
η *is randomly initialized*
for $t = 0, \ldots, T$ **do**
  $\eta = \eta + \alpha \ \text{sign}(\nabla_x MSE(G(x + \eta), G(x)))$
  $\eta = \Pi_S(\eta)$ {projection to set S}
end for
$x' = x + \eta$
return

V. EXPERIMENTS

In this section, we apply our method to various image-to-image translation models using diverse datasets. We first qualitatively compare our method's results with the original model on the tasks of single translation and recursive translation. Next, we provide quantitative analysis using various methods such as principle component analysis, PSNR, and FID score. Detailed explanations of the codes used, datasets and experiment settings are addressed in the supplementary.

A. BASELINE MODELS

We implemented our method on StarGAN [9] and RelGAN [11], which are the representative models for label-based image-to-image translation of facial attributes. In addition, we compared our model on SGN [12], which conducts painting style transfer to mixed domains. In our experiment, the loss domains of SGN are defined as the VanGogh dataset for GAN loss, and the Udnie image for perceptual loss.

B. FACIAL ATTRIBUTE TRANSFER

Fig. 4 illustrates the comparison on the StarGAN, and Fig. 5 illustrates the comparison on the RelGAN. 'Adv' and 'Ori' refer to the models trained with our method and the original methods. 'Single' indicates the result of the single translation

FIGURE 4. Visual comparison of the results of the original StarGAN (Right) and our proposed training method applied to StarGAN (Left).

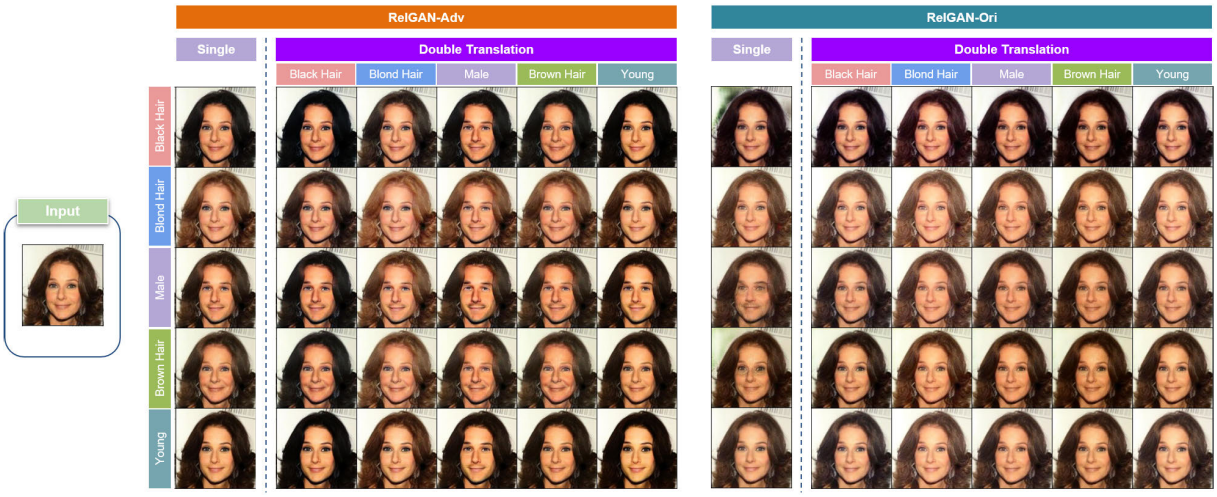

FIGURE 5. Visual comparison of the results of the original RelGAN (Right) and our proposed training method applied to RelGAN (Left).

to the target domain denoted on the left, while 'Double' indicates the result of recursive translation of the single translated image with the target domain of the double translation denoted on the top of the images. Note that 'Gender' in StarGAN refers to translating images into reversed gender, while 'Male' in RelGAN only masculini zes. Both of the models show similar tendencies. First, there are unintended changes in the original model's single translation results. For example, in the original RelGAN, when translating into 'Blond Hair', the overall brightness of the image goes up, and the makeup is erased, whereas our model maintains these properties properly. Furthermore, the single translation result of our model on RelGAN shows better quality in all domains.

Second, both original models show recursive translation problem, and the double translation even ruins the properties of the single translation image. Especially in StarGAN, all of the double translation results seem as if the model tries to map back to the original input image. We assume that the cause is the cycle-consistency loss. During training, the generator reconstructs the generated image back to the original image with the original label. The only recursive translation that the model experience during training is the reconstruction loss which solidifies the mappings between the original image and the reconstructed image, thus the model is overfitted. RelGAN also shows a similar tendency in double translation. The attribute of a single translation such as gender or age seems to disappear in the double-translated results. In contrast, our double-translated image preserves the attributes of a single translation. For example, the translation of gender from the 'blond hair' image still preserves the attribute of 'blond hair', only changing the facial properties regarding gender. The result of more samples with additional explanations can be found in the supplementary.

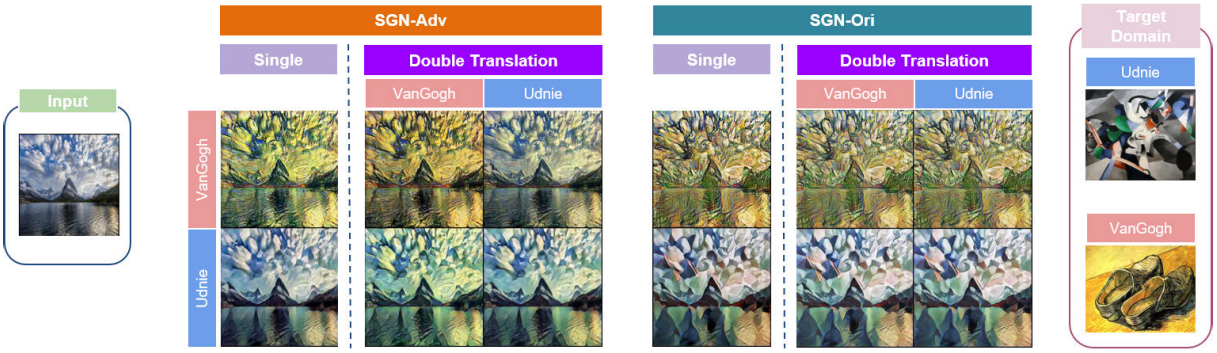

FIGURE 6. Comparing our method to the original SGN.

C. PAINTING STYLE TRANSFER

Fig. 1 and Fig. 6 shows the comparison between our method and original SGN [12]. Similar to the previous facial attribute results, the original SGN failed at recursive translation. The results of double translation is not very different from the single translation. In contrast, our method is robust to multiple translations, conserving the property of the single translated image as well. Notice that the results of a single translation of the original model and our model are different. Our model seemingly preserves the *content* of the input image better, while the original SGN model focuses on translating *style* more rather than preserving the content of the original image. To be more specific, in Fig. 1-right, the 'Ori' model care less about the contents of the input image, even strongly stylizing the empty skies. In contrast, the 'Adv' model mainly stylizes the main contents such as clouds and grounds. In Fig. 6, the single translation toward the Udnie domain using the SGN-Adv model successfully preserve the shape of the clouds in the sky, while the clouds in the SGN-Ori model results are undistinguishable.

For this phenomenon, we thought in terms of the properties of the perceptual loss. Unlike existing GANs which roughly predicts binary labels at the very abstract level with binary classification loss, perceptual loss requires regression of feature maps to align the udnie picture in the feature space of the vgg network. The former can be seen as requiring more specific adjustment than the latter one, which makes it especially difficult to learn the adversarial images that are drawn using the modified PGD attack. Differently with the conventional PGD, our modified PGD attack magnifies the MSE loss with the original output image, so that the outer minimization problem and inner maximization problem can cause more conflict compared to the GAN losses. Therefore, for the generator, it is more difficult to match the perceptual loss using the samples that are generated with the modified attack, requiring more specific adjustments to its outputs. As a result, the generator does not change the entire image, but strongly changes the styles only in the specific area where the important contents exist and preserves the rest part of the image.

In addition, the perceptual loss is quite sensitive to hyper-parameter adjustment. Since our method assumes that the rest of the losses in the existing network are not touched, in practice, putting more weight on the style loss can make the change in style stronger. We presume that this can be balanced through re-scaling the hyper-parameter of the style and content losses, but we left it untouched because the current phenomenon is satisfying, and also we want to avoid applying heuristic adjustments to our method.

D. QUANTITATIVE RESULTS

PCA analysis In order to concretely verify our hypothesis, we conducted principal component analysis (PCA) using StarGAN models, and the results are shown from (A) to (D) in Fig. 7, which are the experimental demonstration of Fig. 2. The 'D-ori' represents the real CelebA data with 'Black Hair', and the 'D-tar' represents the real CelebA data with 'Blond Hair'. The 'adv' represents the translated images from 'Black Hair' to 'Blond Hair' using our model ($G_{adv}(x, v_{blond})$), while 'ori' represent the translated image from 'Black Hair' to 'Blond Hair' using the original model ($G_{ori}(x, v_{blond})$). For each group, the average values are displayed using square markers. Each data group has 100 different base images randomly selected from CelebA dataset. We train a separate classifier that distinguishes 2 classes, 'Blond Hair' and 'Black Hair' domains, and use it as a feature extractor for PCA. We visualized the results of five top different component pairs, with the first component fixed at the x-axis. Obviously, the first component represents translation from 'Black Hair' to 'Blond Hair'. The 'ori' distribution is far apart from the 'D-tar' distribution, while the 'adv' distribution almost overlaps with the 'D-tar' distribution. The average distance between the mean values of 'D-tar' and the 'ori' is 4.8, whereas the average distance between the mean values of 'D-tar' and 'adv' is only 0.9. This clearly visualizes the existence of the inherent bias in the original model that keeps the translated images away from the real data distribution and it is much alleviated by our model.

Orthogonal regularization RelGAN [11] adopts the orthogonal regularization loss (L_{orth}) of the BigGAN [20],

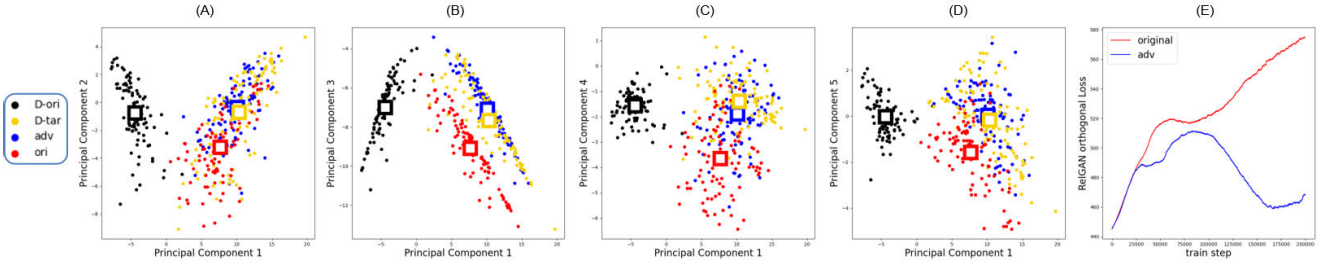

FIGURE 7. (A) to (D) are visualization of 100 images for each of the original (black) and target (blond) domain using PCA. The red and blue dots are the translated images of the original StarGAN and our Adv-StarGAN respectively. □ denotes the mean of 100 samples, which clearly shows that our method better translates to the target domain than the original StarGAN (compare the distance between □ and □ and that between □ and □). (E) is the comparison of orthogonal regularization loss from RelGAN.

TABLE 1. PSNR and FID score for quantitative analysis. A higher PSNR scores and lower FID scores indicate better image quality.

Double Translation	PSNR (↑)				
	Black	Blond	Male	Old	Young
Ori	24.2	22.5	20.2	22.0	23.6
Ours	27.4	28.5	26.9	28.2	26.1
Single Translation	**FID (↓)**				
	Black	Blond	Male	Old	Young
Ori	42.3	55.0	62.6	39.0	46.6
Ours	39.5	50.1	56.6	39.5	44.5

which induces the weight matrix close to be orthogonal in channel-wise. During the training of the RelGAN-Ori, we observed the increase of L_{orth}, which implies that it learns correlations. On the other hand, in RelGAN-Adv, L_{orth} increases at first but gradually decreases, showing the ability of disentangling, which reflects the disentangling behavior in Fig. 2. The comparison is shown in Fig. 7(E).

Numerical Evaluation Table 1 represents the PSNR scores of the double-translated image on RelGAN (top) and the FID scores of the single-translation images (bottom). The PSNR scores are averaged for each 5 domains. Our method resulted in higher PSNR scores and lower FID scores, which indicates that our method is superior in both double and single translation.

For evaluating the PSNR scores, we set the ground truth for the double translation as the equivalent single translation result of attribute changes in the double translation. For example, suppose there are 3 domains $d \in \{A, B, C\}$, if the target label for first and double translation is $v_A = [1, 0, 0]$ and $v_B = [0, 1, 0]$, then the ground truth target domain label of double translation is $v_{AB} = [1, 1, 0]$. Thus, the ground-truth target for $G(G(x, v_A), v_B)$ is $G(x, v_{AB})$.

The FID scores are used for evaluating the visual quality of single translation results to empirically show that our adversarial-attack-based training doesn't harm its original performance, but rather slightly improve its performance by improved disentanglement. We compared the FID scores of the single translation results, which are commonly used for evaluating image qualities.

VI. DISCUSSIONS

A. VISUAL QUALITY OF SINGLE TRANSLATION

Through experiments, we confirmed that our modified PGD definitely succeeds in recursive translation. For single translation, we found some pros and cons of our method. For both StarGAN and RelGAN, the single translations of our training strategy disentangle better than the original models, but for StarGAN, our results occasionally look a bit noisier than the original one (a closer look at Fig. 4). This might be from the injected noise to the input. Fortunately, in RelGAN, our training improves both disentanglement and visual quality.

B. RECURSIVE TRANSLATION IN REFERENCE-BASED IMAGE TRANSLATION

As we mentioned in our paper, the domain of unpaired image translation can be largely divided into an implicit-domain and an explicit-domain translation. Through some experiments, we found that implicit-domain models such as StarGANv2 [21] and StyleGAN [23] do not suffer from the recursive translation problem. However, we want to add that they have quite different properties producing different behaviors concerning the recursive translation. Also, its behavioral differences result in pros and cons in terms of practicality and applicability.

1) DIFFERENT BEHAVIOR CONCERNING THE RECURSIVE TRANSLATION

Implicit-domain translation models have quite different properties compared to explicit-domain translation models *(e.g. StarGANs, RelGANs)*. First, while the behaviors of implicit- and explicit-domain translation models are the same for the training phase at testing phase, implicit domain models do not conduct image translation at the training phase. They only learn to generate random samples at the training phase, and conduct latent space manipulation at the test phase. Especially, the image translation process has a large difference.

Explicit models usually use residual encoder-decoder-based models that forward images with their target domain labels paired as inputs, whose procedure is very straightforward and intuitive. Implicit models have to conduct latent space exploration for the trained model to find a specific

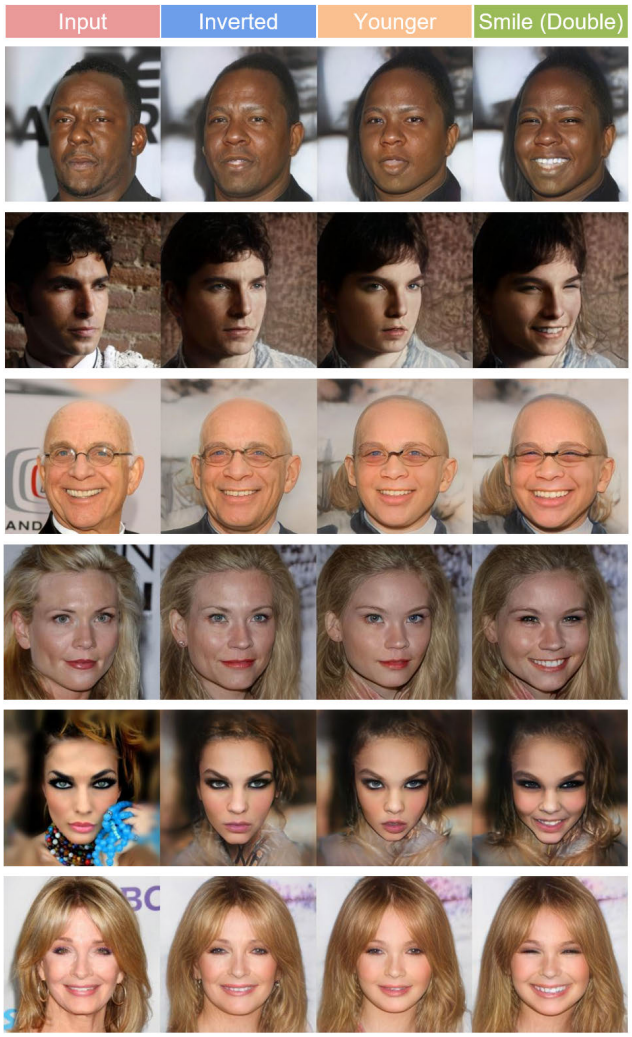

FIGURE 8. Result images of StyleGAN V2, an implicit-domain model. The images at the 'Younger' column are single translated images, and the 'Smile (Double)' column are double-translated images.

latent vector for carrying out each domain, and add that latent vector to translate domains for their own generated (fake) samples. To translate real samples, they conduct GAN-Inversion to get corresponding latent vectors, and then manipulate the latent vectors to conduct image translation.

Figure 9 shows the behavioral differences between the original explicit-domain models, our improved models, and implicit-domain models. The $Bias_{BD}$ can still occur at the second translation. Implicit domain models may fail to accurately disentangle the output with the unintended attribute D, but the GAN-Inversion procedure force the unaligned output to align through latent vector Z, which becomes the input space for the second translation. Thus, an implicit model is able to translate images recursively regardless of the existence of the bias. In the first three rows of Fig. 8, we can clearly find that 'Younger' and 'Smile' attribute correlates with 'hairy' attribute, which shows the correlations with the undesired attributes.

2) PROS AND CONS OF IMPLICIT-DOMAIN IMAGE TRANSLATION

Cons Shortcomings for these procedures of implicit-domain model are: 1) The need of latent space exploration to define domains which is a heuristic procedure. 2) The choice of latent space is ambiguous: The choice of latent spaces results in different tradeoffs between editing power and reconstruction power. 3) For real samples, the generator has to invert a real image and find its latent, which is also a heuristic procedure. 4) The inverted samples show decent performance in terms of an intended attribute, but perform poorly at reconstruction [22], [23], [24] which limits the applicability of implicit-domain models. Ignoring the bias and forcing the output to align with latent Z in Fig. 9 may also ascribe the poor reconstruction of GAN-Inversion. Meanwhile, the residual encoder-decoder structure of explicit models generally does not suffer from misconducting reconstruction. For all the images in Fig. 8, this phenomenon is very obvious: observe the fourth shortcoming that the translated images fail to reconstruct the details other than the areas that are interested (e.g. backgrounds, hair).

Pros The implicit models also exhibit some superiority: 1) Though conducted in a heuristic manner, with a trained generator, they can transfer numerous styles that are explicitly defined during the training phase. 2) Image quality in terms of sharpness and resolution is superior compared to conventional explicit models. However, poor reconstruction power of inverted real samples limit their practicality in conditional image translation.

In this paper, we mainly deal with the phenomena and solutions in the explicit-domain models, and only compare their characteristics with the implicit-domain model. However, since our learning method not only enables recursive translation, but also fundamentally increases disentangling performance, we believe that if it can be properly applied to the implicit-domain model, the reconstruction power of the GAN-inversion in implicit-domain models can be increased through stronger disentangling, and we leave it as our future work.

C. N-TIMES TRANSLATION

For the translations of more than two times, we confirmed our model still performs appropriately (we tested up to 5 times of recursive translation). The translation results of more than two times are shown in Fig. 10.

D. ON THE POSSIBILITY OF UTILIZING OUR METHOD FOR DETECTING DEEP-FAKES

Deepfake is an image synthesis technique that alters a person's appearance in existing photos or videos, and it can be exploited for malicious purposes. However, if the existing image synthesis inherits the problem of recursive translation, checking the recursive transfer-ability can be a possible means for future researchers. Experimentally, we found that the problem of recursive translation can occur between

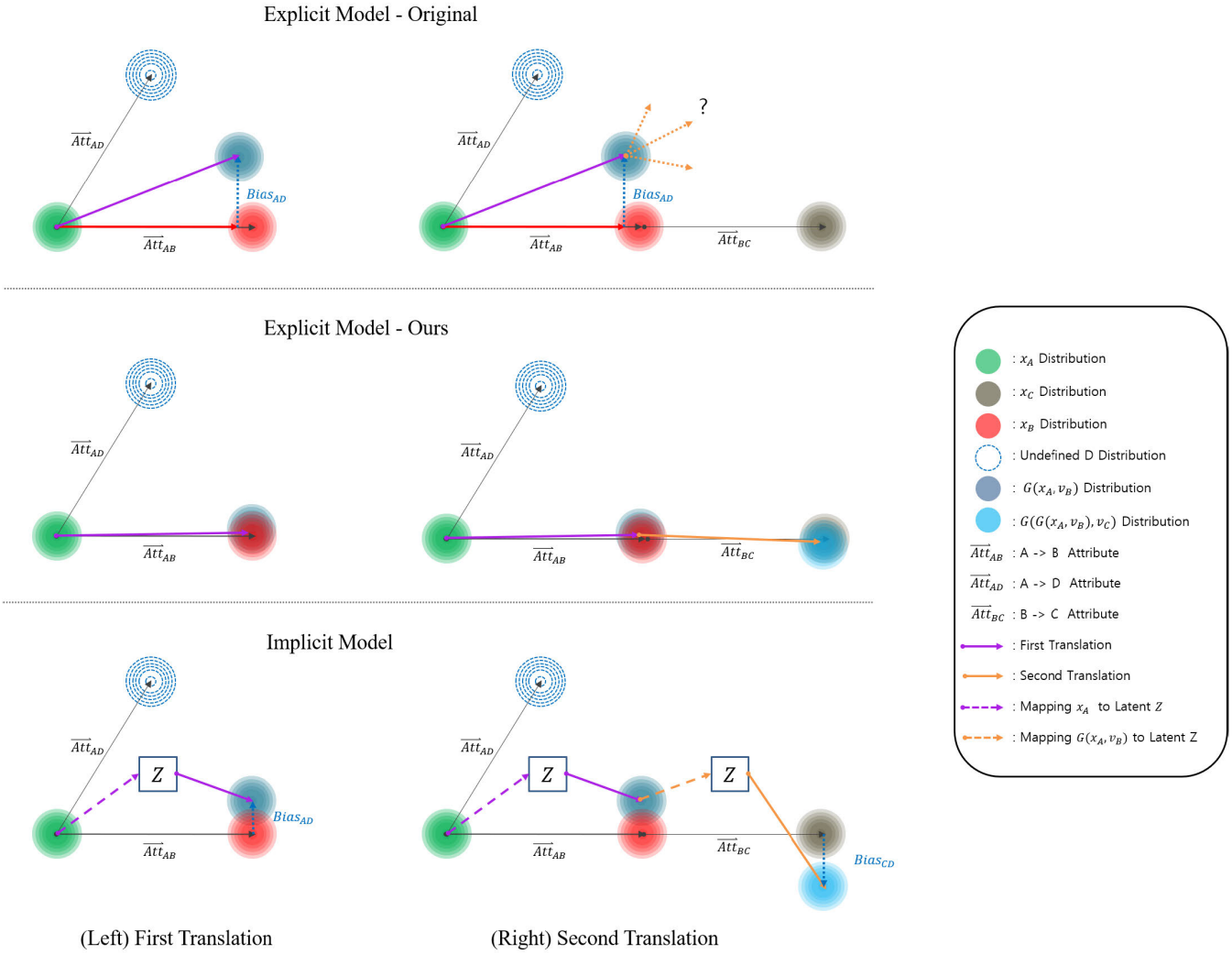

FIGURE 9. **(Left) is the illustration of the first translation of image synthesis models, and (Right) is the illustration of the second translation of image synthesis models.**

different models as well. We separately trained two original StarGAN generators G_1 and G_2, and as can be seen in Figure 11, we could find that both $G_2(G_1(x, v_1), v_2)$ and $G_1(G_2(x, v_1), v_2)$ fail to translate appropriately. If a model fails to translate an image, then one can infer that the image had already been synthesized once.

Furthermore, this behavior supports the existence of the *Bias* in Fig. 9. Although G_1 and G_2 are trained separately, the real x_B distribution remains unchanged, so that the *Bias* arises for both G_1 and G_2. As a result, the output of the first translation (distribution $G_1(x_A, v_B)$ and $G_2(x_A, v_B)$) serves as an unseen input distribution of its opponents (G_1 and G_2).

E. BIAS OF OUR MODEL

In Figure 1, one may doubt whether our modified PGD attack can really remove $Bias_{AD}$, because the bias is the inherent property that lies in the dataset. Also, one can ask whether our adversarial noise can cause another bias to our model, which can make unexpected side effect that harm the original goal of training. It is worthwhile saying that Figure 1 is the concept illustration of the network which already learned the undesirable correlations. Using a modified PGD attack, our model does not learn the undesired correlation from the beginning of the GAN training, so that the resultant bias will be alleviated much. This behavior is also observed through the PCA analysis in Figure 7. Also, in this figure, we can find that beyond removing the unintended bias, the 'adv' model also conduct the intended translation ('D-ori' to 'D-tar') better than the 'ori' model.

F. COMPARISON ON MODIFIED PGD ATTACK

Figure 12 shows the effect of modified PGD attack. The perturbed image '$x+\eta$' in the second column is less distinguishable from the input 'x', while the translated image '$G_{ori}(x)$' and '$G_{ori}(x+\eta)$' are far different. However, the translated image with our method '$G_{adv}(x+\eta)$' seems reasonably

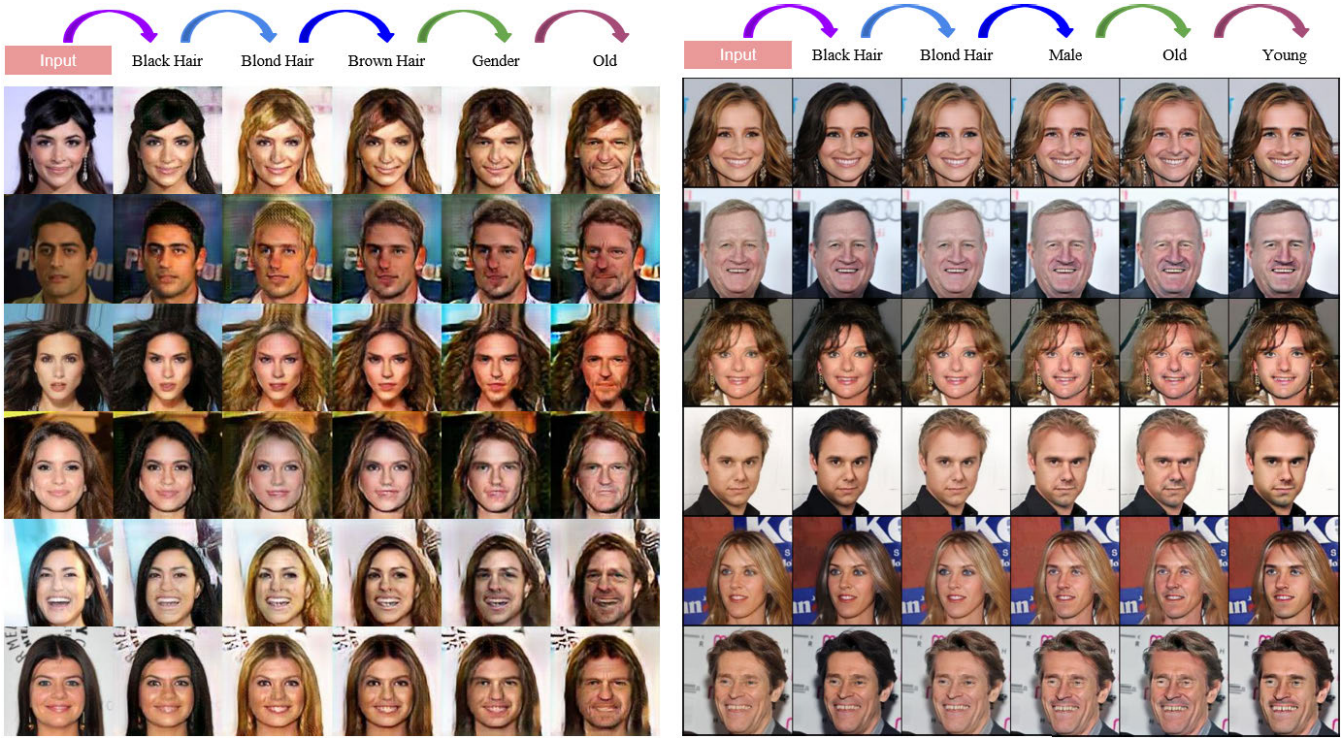

FIGURE 10. Left is the N-times recursive translation result of our method on StarGAN and right is the N-times recursive translation result of our method on RelGAN.

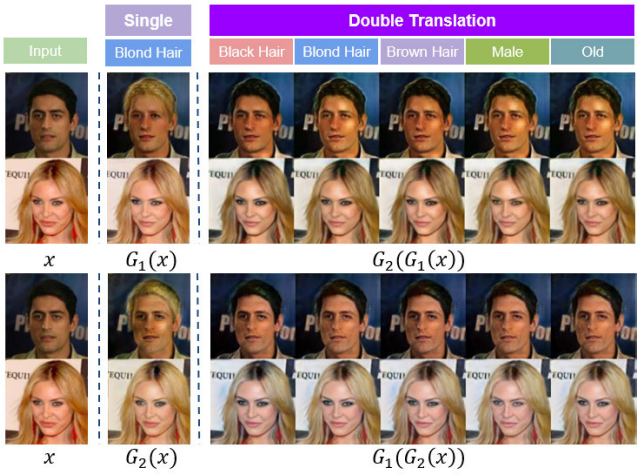

FIGURE 11. The translated image of 2 separate StarGAN original models. The recursive translation problem can occur between two different models.

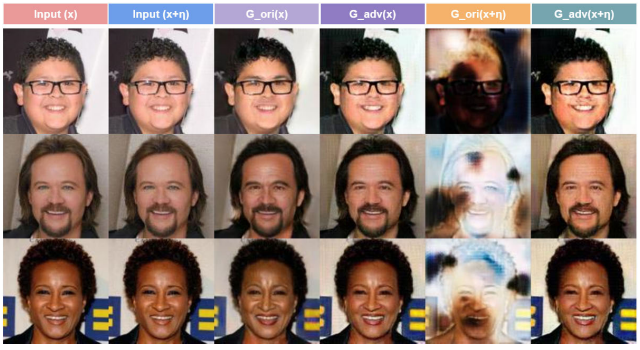

FIGURE 12. The comparison of perturbed image in StarGAN model. '$x+\eta$' denotes the perturbed image by modified PGD attack, 'G_{ori}' denotes the original model, and 'G_{adv}' is the model with our method.

translated into the target domain, but there are minor changes detected in the facial expression. It implies that our method finds the unique direction to the target domain, while diversifying other attributes to the undesired domains.

Implementing our method enables multiple recursive translations. Figure 10 shows the result of up to 5 times recursive translation. The arrows on above indicate the object image to be translated. For example, the second column is the translated image of the first column, and the third column is the translated image of the second column. Our method on StarGAN shows a decline in quality with more translation we do, however, our method on RelGAN has no noticeable quality difference in the multiple recursive translations.

VII. DETAILED EXPERIMENT SETTINGS

A. DATASET

1) CelebA

The CelebFaces Attributes Dataset (CelebA) is a large-scale dataset with 202,599 face images of celebrities, each labeled

with 40 facial attributes. We center-cropped these images to 178 × 178. When implementing our methods on RelGAN, we resize them into 256 × 256 and on StarGAN, we resize them into 128 × 128.

2) Photo2Art

We used various landscape photos and paintings from *Flickr* to train our model on Symparameterized-GAN. *Udnie of Francis Picabia* and *VanGogh* painting datasets are used for transferring styles of images.

B. ENVIRONMENTS

Our method first generates the adversarial examples and replaces them with the original training dataset. The adversarial examples were crafted with the l_∞-bounded attacks, with the constraint of ϵ equals to 8/255, the step size α equals to 0.01 and the step iteration equals to 10. We followed the settings from PGD-attack. We followed the exact same settings from the original paper, modifying the code from https://github.com/yunjey/stargan, https://github.com/elvisyjlin/RelGAN-PyTorch, and https://github.com/TimeLighter/pytorch-sym-parameter.

VIII. CONCLUSION

In this paper, for the first time we reveal that existing label-based image translation models commonly suffer from the problem of *failure of recursive translation*, that once the image is translated to another domain, that image cannot be translated recursively. Regarding this phenomenon, we propose an undesired correlation hypothesis, and based on our explanation, we propose a neat solution using an adversarial attack that is easy to implement. Our solution results in a valid model that not only enables recursive translation, but also enables the model to disentangle attributes better on single translations. In this paper, we confine our contributions to label-based translations, and expect to expand our work in implicit-domain translations for future work. We hope our work and interesting results can encourage researchers to think back at conventional image translation methods for better tradeoffs between manipulation and reconstruction, and discuss more about our findings.

ACKNOWLEDGMENT

(Seonguk Park and Jookyung Song contributed equally to this work.)

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[4] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 694–711.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[6] 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. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[7] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1857–1865.

[8] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional GANs for image editing," 2016, *arXiv:1611.06355*.

[9] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[10] X. Yu, X. Cai, Z. Ying, T. Li, and G. Li, "SingleGAN: Image-to-image translation by a single-generator network using multiple generative adversarial learning," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Perth, WA, Australia: Springer, 2018 pp. 341–356.

[11] Y.-J. Lin, P.-W. Wu, C.-H. Chang, E. Chang, and S.-W. Liao, "RelGAN: Multi-domain image-to-image translation via relative attributes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5914–5922.

[12] S. Chang, S. Park, J. Yang, and N. Kwak, "Sym-parameterized dynamic inference for mixed-domain image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4803–4811.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[14] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4422–4431.

[15] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," 2018, *arXiv:1805.07894*.

[16] Y. Yoo, S. Park, J. Choi, S. Yun, and N. Kwak, "Butterfly effect: Bidirectional control of classification performance by small additive perturbation," 2017, *arXiv:1711.09681*.

[17] C.-Y. Yeh, H.-W. Chen, H.-H. Shuai, D.-N. Yang, and M.-S. Chen, "Attack as the best defense: Nullifying image-to-image translation GANs via limit-aware adversarial attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16188–16197.

[18] H.-S. Choi, J. Lee, W. Kim, J. H. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," 2021, *arXiv:2110.14513*.

[19] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.

[20] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.

[21] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8188–8197.

[22] Z. Wu, D. Lischinski, and E. Shechtman, "StyleSpace analysis: Disentangled controls for StyleGAN image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12863–12872.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[24] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.

SEONGUK PARK received the B.S. degree in electrical engineering from Sungkyunkwan University, Seoul, South Korea, in 2017. He is currently pursuing the Ph.D. degree with the Department of Intelligence and information, Seoul National University, Seoul. His research interests include generative models, model compression, and low-level vision.

JOOKYUNG SONG received the B.S. degree in political science and international relations from Korea University, Seoul, South Korea, in 2020. She is currently pursuing the Ph.D. degree with the Department of Intelligence and Information, Seoul National University, Seoul. Her research interests include generative models, image translation, and adversarial attack.

DONGHOON HAN was born in Bucheon, South Korea, in 1996. He received the B.S. degree in computer science and international studies from Kyung Hee University, Suwon, in 2021. He is currently pursuing the M.S. degree with Seoul National University.

Since 2021, he has been with the Machine Intelligence and Pattern Recognition Laboratory, Seoul National University. His current research interests include computer vision, generative modeling, and multimodal learning with deep neural networks.

NOJUN KWAK (Senior Member, IEEE) was born in Seoul, South Korea, in 1974. He received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, in 1997, 1999, and 2003, respectively. From 2003 to 2006, he was with Samsung Electronics, Seoul. In 2006, he joined Seoul National University as a BK21 Assistant Professor. From 2007 to 2013, he was a Faculty Member with the Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea. Since 2013, he has been with the Graduate School of Convergence Science and Technology, Seoul National University, where he is currently a Professor. His current research interests include feature learning by deep neural networks and their applications in various areas of pattern recognition, computer vision, and image processing.