Neural Networks 162 (2023) 330–339

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

SuperstarGAN: Generative adversarial networks for image-to-image translation in large-scale domains

Kanghyeok Ko [a], Taesun Yeom [b], Minhyeok Lee [a,*]

[a] School of Electrical and Electronics Engineering, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, South Korea
[b] School of Mechanical Engineering, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, South Korea

ARTICLE INFO

Article history:
Received 31 October 2022
Received in revised form 19 January 2023
Accepted 28 February 2023
Available online 7 March 2023

Keywords:
Generative adversarial networks
Image-to-image translation
Domain translation
Face image translation
Image generation

ABSTRACT

Image-to-image translation with generative adversarial networks (GANs) has been extensively studied in recent years. Among the models, StarGAN has achieved image-to-image translation for multiple domains with a single generator, whereas conventional models require multiple generators. However, StarGAN has several limitations, including the lack of capacity to learn mappings among large-scale domains; furthermore, StarGAN can barely express small feature changes. To address the limitations, we propose an improved StarGAN, namely SuperstarGAN. We adopted the idea, first proposed in controllable GAN (ControlGAN), of training an independent classifier with the data augmentation techniques to handle the overfitting problem in the classification of StarGAN structures. Since the generator with a well-trained classifier can express small features belonging to the target domain, SuperstarGAN achieves image-to-image translation in large-scale domains. Evaluated with a face image dataset, SuperstarGAN demonstrated improved performance in terms of Fréchet Inception distance (FID) and learned perceptual image patch similarity (LPIPS). Specifically, compared to StarGAN, SuperstarGAN exhibited decreased FID and LPIPS by 18.1% and 42.5%, respectively. Furthermore, we conducted an additional experiment with interpolated and extrapolated label values, indicating the ability of SuperstarGAN to control the degree of expression of the target domain features in generated images. Additionally, SuperstarGAN was successfully adapted to an animal face dataset and a painting dataset, where it can translate styles of animal faces (i.e., a cat to a tiger) and styles of painters (i.e., Hassam to Picasso), respectively, which explains the generality of SuperstarGAN regardless of datasets.

1. Introduction

Generative adversarial networks (GANs) (Goodfellow et al., 2014) have shown great promise in the field of computer vision, including image generation (Brock et al., 2018; Karras et al., 2017; Kim, Kim et al., 2022; Park et al., 2022; Radford et al., 2015; Srivastava et al., 2022; Toshpulatov et al., 2021) and style transfer tasks (Isola et al., 2017; Kim et al., 2017; Liu, 2021; Taigman et al., 2016; Yuan & Zhang, 2022). The image-to-image translation is a specific sub-problem that has received much attention. This task involves taking a given image and changing it to display target features (Chen et al., 2016; Liu et al., 2021; Ojha et al., 2021; Pang et al., 2021; Xie et al., 0000; Zhuang et al., 2020). For example, given a face image, an image-to-image translation model can be used to change facial features, such as expression, hair color, and age (Zhang et al., 2017). One of the most critical applications of image-to-image translation is in the field of medical imaging. Specifically, GANs have been used to synthesize images that can be used to train machine learning models for many medical tasks that lack training data. For example, a recent study (Armanious et al., 2020) showed that a GAN-based system was able to generate realistic computed tomography (CT) images from positron emission tomography (PET) data. Such a study allows for creating training datasets that would be otherwise difficult or impossible to obtain.

In image-to-image translation, it is crucial to modify images to display the target features because it is the fundamental objective of the model. In addition, features other than the target features must be maintained after the translation. These two objectives must be satisfied when training an image-to-image translation model. Recently, deep learning-based models have been employed to address this problem.

Consequently, numerous studies using deep learning models have been conducted to control the target features of translated images. Specifically, GANs have been used to achieve this objective in recent years. In GANs, the model addresses this problem using competitive learning between two deep learning modules, called the generator and discriminator. A modification of the conventional GAN, called cycle-consistent GAN (CycleGAN) (Zhu

* Corresponding author.
E-mail address: mlee@cau.ac.kr (M. Lee).

https://doi.org/10.1016/j.neunet.2023.02.042

et al., 2017), uses two generators and two discriminators for image-to-image translation. Each generator in CycleGAN translates an image to display a specific feature. The discriminators in CycleGAN enforce the constraint that the translated images contain only the desired features. In this manner, CycleGAN can be employed for image-to-image translation with only two domains, where each generator manages a domain.

However, ${}_nC_2$ CycleGAN models must be trained for image-to-image translation with multiple domains. For instance, when the CelebFaces Attributes (CelebA) face dataset (Liu et al., 2015) is used for training, 780 CycleGAN models must be trained because the dataset has 40 domains (i.e., attributes). Although image-to-image translation with CycleGAN has demonstrated excellent performance in many studies, this limitation remains a crucial problem for datasets with multiple domains.

StarGAN (Choi et al., 2018) has been proposed for image-to-image translation with multiple domains to address this limitation. In StarGAN, features are encoded with a feature vector, which is used as an additional input of the generator. After training, synthetic images with desired features can be generated by corresponding feature vectors; in this manner, it becomes possible to use only a single generator for multiple image-to-image translations. This advantage of StarGAN significantly reduces the training time because multiple models are not necessary for multiple domains.

However, StarGAN has a few limitations. First, StarGAN generally fails to learn minor features. For example, when StarGAN is trained with a face image dataset, the model can barely learn minor facial features, such as a *big nose* or *mustache*, whereas it successfully learns significant features, such as *hair color* and *skin color*. Furthermore, if large-scale domains are trained with a single model, the performance of StarGAN significantly decreases. Specifically, the quality of the translated images is reduced, while translated images barely display the target features.

For image-to-image translation between large-scale domains, we propose a modification of StarGAN, called SuperstarGAN, which uses the framework in a controllable GAN (ControlGAN) (Lee & Seok, 2019). In ControlGAN, adopting an independent classifier with data augmentation (DA) techniques outperforms the other form of feature learning in which the discriminator manages feature learning (Perez & Wang, 2017). The independent classifier can adequately learn the features because it is independent of the GAN training, and DA is employed to enhance the performance. Thus, the discriminator focuses on its own target, enhancing the quality of the translated images. Moreover, target features can be well-trained using the independent classifier compared to conventional methods.

SuperstarGAN introduces an independent classifier to enhance feature learning, addressing the limitations of StarGAN. The generator can train with a fine classifier; thus, minor features can be translated between large-scale domains. In addition, the generator can produce realistic translated images, as the discriminator can focus on enhancing the image quality.

2. Background

2.1. A brief review of the generative adversarial network

The concept of the GAN (Goodfellow et al., 2014) is to generate a realistic fake sample through adversarial training. There are two components in the adversarial training: a generator and a discriminator. The generator produces counterfeit images, and the discriminator establishes if they are genuine or fake. The objective of the training is to make the fake images created by the generator appear as realistic as possible so that it can deceive the discriminator. During each training iteration, a mini-batch of real images (X) and noises (Z) are chosen at random. The generator network (G) then creates fake images ($G(Z)$). The discriminator network, D, outputs a probability, $P(D = 1|X)$, indicating whether X is real or not. The input for the discriminator is either X or $G(Z)$ and then it produces the probability.

Both of these neural networks are trained simultaneously. As this adversarial training process repeats, the generator becomes better at making realistic samples that cannot be distinguished from real images. The GAN can be optimized using the following objective function:

$$\hat{G} = \arg\min_{G} \{L_{adv}(r, D(G(\boldsymbol{Z})))\},$$

$$\widehat{D} = \arg\min_{D} \{L_{adv}(r, D(\boldsymbol{X})) + L_{adv}(f, D(G(\boldsymbol{Z})))\},$$

where r and f are set at one and zero, respectively.

2.2. ControlGAN

After the introduction of the GAN, many studies have been conducted to extend the GAN to a conditional model. For example, the conditional GAN (cGAN) (Mirza & Osindero, 2014) uses extra information containing the desired domain specifications as input to the generator and discriminator. The cGAN is similar to the vanilla GAN, except that both the generator and discriminator are conditioned on additional information. The vanilla GAN can be extended to the cGAN by simply adding the labels of corresponding images (Y) as an extra input to both G and D. In addition, the auxiliary classifier GAN (ACGAN) (Odena et al., 2017) adopts an auxiliary classifier to determine whether the generated samples belong to the target domain. The ACGAN can be considered an extension of the cGAN where the discriminator outputs not only the probability that the input image is real or fake, but also the probability for each class. This allows the ACGAN to generate targeted images in a more precise manner.

The ControlGAN (Lee & Seok, 2019) is a modification of ACGAN, where the model uses DA to overcome the classifier overfitting problem in the ACGAN, which significantly hinders conditional learning. While DA can hardly be used with the ACGAN structure since the ACGAN discriminator is disturbed by DA, the ControlGAN handles this problem by separating the classifier from the discriminator. Thus, a fine classifier that is trained with DA can offer sound guidance for the generator that trains with classification loss from the independent classifier. The objective functions of ControlGAN are as follows:

$$\hat{G} = \arg\min_{G} \{L_{adv}(r, D(G(\boldsymbol{Z}))) + \lambda_{cls} \cdot L_{cls}(\boldsymbol{T}, C(G(\boldsymbol{Z}, \boldsymbol{T})))\},$$

$$\widehat{D} = \arg\min_{D} \{L_{adv}(r, D(\boldsymbol{X})) + L_{adv}(f, D(G(\boldsymbol{Z})))\},$$

$$\hat{C} = \arg\min_{C} \left\{L_{cls}\left(\boldsymbol{T}, C\left(\boldsymbol{X}_{\boldsymbol{aug}}\right)\right)\right\},$$

where λ_{cls} is a hyperparameter for conditional learning, $\boldsymbol{T}$ denotes the target domain label, and $\boldsymbol{X}_{\boldsymbol{aug}}$ denotes the real image modified by DA.

2.3. Image-to-image translation and CycleGAN

The image-to-image translation is a task to change specific aspects of a given image to another. Contrary to general GAN structures, which learn the mapping from the latent space, the image-to-image translation model learns the mapping between the input and output domains of images. The image-to-image translation task is useful for various applications, such as transferring the style of a given image to another or generating an image from a semantic layout. Most recently, Kim, Kim et al.

Table 1
Comparison with recent methods for image-to-image translation. G: generator, D: discriminator, C: classifier, E: encoder, DE: decoder.

Domain	Method	Year	Conditioning	Components description
Bidirectional	UNIT	2017	Cyclic process	Two E, two G, and two D
	CycleGAN	2017	Cyclic process	Two G, two D
	SaGAN	2018	Class label	One G with two networks, one D with auxiliary C
	ERGAN	2020	Cyclic process	Four E, two G, and one D
Multiple	StarGAN	2018	Class label	Unified single model with one G, one D with auxiliary C
	AttGAN	2019	Class label	Unified single model with one encoder G, one decoder G, and one D with auxiliary C
	Kim, Park et al. (2022)	2022	Style encoder	Unified single model with one encoder G, one decoder G, and one D as style E
	GP-UNIT	2022	Style encoder	Unified single model with two E, one DE, one G, one D, and one C; pre-trained BigGAN
	SuperstarGAN	2023	Class label	Unified single model with one G, one D, and one independent C

(2022) adopted a style-aware discriminator used for style encoding as well as adversarial loss for controllable image translation. The model obtains continuous style space as pseudo-labels to substitute for class labels. They extended image translation functionality to various applications (i.e., style interpolation, content transplantation, and local image translation). Similarly, GP-UNIT (Yang et al., 2022) proposed a versatile framework that is trained with generative priors from a pre-trained class-conditional GAN (e.g., BigGAN). They achieved image translation between two domains with drastic differences (e.g., *Bird* to *Car*). Table 1 shows the comparison of several methods (He et al., 2019; Hu et al., 2020; Kim, Park et al., 2022; Liu et al., 2017; Yang et al., 2022; Zhang, Kan et al., 2018) for the image-to-image translation including the most recent works. All models in Table 1 are trained with unpaired datasets. Compared to other methods, SuperstarGAN is specialized in representing large-scale domains with a unified single model.

The CycleGAN (Zhu et al., 2017) achieves remarkable results in image-to-image translation in an unsupervised manner, making it possible to learn a model without paired training datasets because CycleGAN trains the property of cycle consistency (Zhou et al., 2016). The CycleGAN consists of two generators (G_{AB} and G_{BA}) and two discriminators (D_A and D_B), where each generator and discriminator pair manages a domain (A or B). Given an image (X_A) with a specific domain A, the G_{AB} modifies the image to display the other domain B ($G_{AB}(X_A) = \hat{X}_B$). Thus, if a given image is modified by the two generators in CycleGAN, the modified image must be the same as the original image, i.e., $G_{BA}(G_{AB}(X_A)) = X_A$. The CycleGAN uses this property as a training loss, called cycle-consistency loss. The two discriminators decide whether the translated images are real and display the corresponding domain feature.

However, CycleGAN has a fundamental limitation in that it can only be used for image-to-image translation between two domains. Thus, for image-to-image translation with multiple domains, each pair of two domains must be trained with each CycleGAN model. For example, image-to-image translation with CycleGAN would require 4950 separate models for a dataset of 100 domains, which is infeasible considering the amount of time it takes to train a single CycleGAN model.

2.4. StarGAN

To address the limitation of CycleGAN, StarGAN (Choi et al., 2018) employs the ACGAN idea of using conditional inputs for the generator and an auxiliary classifier in the discriminator. As a result, StarGAN effectively performs image-to-image translation in multiple domains using only a single generator and discriminator pair. Furthermore, as training a single model can learn global features for multiple domains, the training process is more efficient than using many individual models.

Specifically, to achieve image-to-image translation for multiple domains with a single generator, StarGAN modifies components of the general CycleGAN structure; the model uses a target domain vector as an additional input to the generator. The target domain is randomly determined in the training phase, which enables the generator flexibly translate to various domains. The discriminator in StarGAN uses the discriminator of ACGAN, which has an auxiliary classifier that determines whether the translated image by the generator is modified adequately into the target domain, which is used for the target domain vector of the generator. Using this modification of StarGAN, the image-to-image translation in multiple domains becomes possible with a single generator.

The training of StarGAN is similar to the cycle-consistency training in CycleGAN. Given an input image X with a corresponding domain T, the generator translates the image with a target domain of T', i.e., $G(X, T')$. The translated image is reconstructed into the original domain T by the same generator with a different input label, i.e., $G(G(X, T'), T)$. This image with twice translations must be the same as the original image. Therefore, a loss with L_1 norm is adopted to compare $G(G(X, T'), T)$ and X. Through this reconstruction process, the translated image is guaranteed to be modified only with the domain-related features while maintaining the overall features of the original images. The training process of StarGAN can be represented as follows:

$$\hat{G} = \arg\min_G \{L_{adv}(r, D(G(\boldsymbol{X}, \boldsymbol{T}))) + \lambda_{cls} \cdot L_{cls}(\boldsymbol{T}, C_D(G(\boldsymbol{X}, \boldsymbol{T}))) + \lambda_{rec} \cdot L_{rec}(\boldsymbol{X}, G(G(\boldsymbol{X}, \boldsymbol{T}), \boldsymbol{T}'))\},$$

$$\widehat{D} = \arg\min_D \{L_{adv}(f, D(G(\boldsymbol{X}, \boldsymbol{T}))) + L_{adv}(r, D(\boldsymbol{X})) + \lambda_{cls} \cdot L_{cls}(\boldsymbol{T}, C_D(\boldsymbol{X}))\},$$

where λ_{cls} and λ_{rec} denote the hyperparameter sets controlling the ratio of adversarial loss, classification loss, and reconstruction loss; C_D is the auxiliary classifier of the discriminator.

3. Methods

This paper proposes SuperstarGAN, an improved StarGAN for large-scale domains. As the conventional StarGAN employs the ACGAN discriminator, it also has the limitation of ACGAN in that the auxiliary classifier overfits the training set. The overfitted classifier can hardly transfer helpful information for the training of the generator, resulting in a limitation in the training of domain information. This limitation is mainly caused by the structural problem of the ACGAN discriminator, in which the classifier and discriminator are integrated. Additionally, due to this limitation, StarGAN generally fails to be trained with large-scale domains. Furthermore, the quality of generated images deteriorates because the auxiliary classifier is not sophisticated enough to accurately train domain information.

To address the limitations of StarGAN, the ControlGAN framework is introduced in SuperstarGAN. As illustrated in Fig. 1, SuperstarGAN consists of three components: the generator, discriminator, and classifier. Compared to StarGAN, the main modification of SuperstarGAN is the independent classifier. As

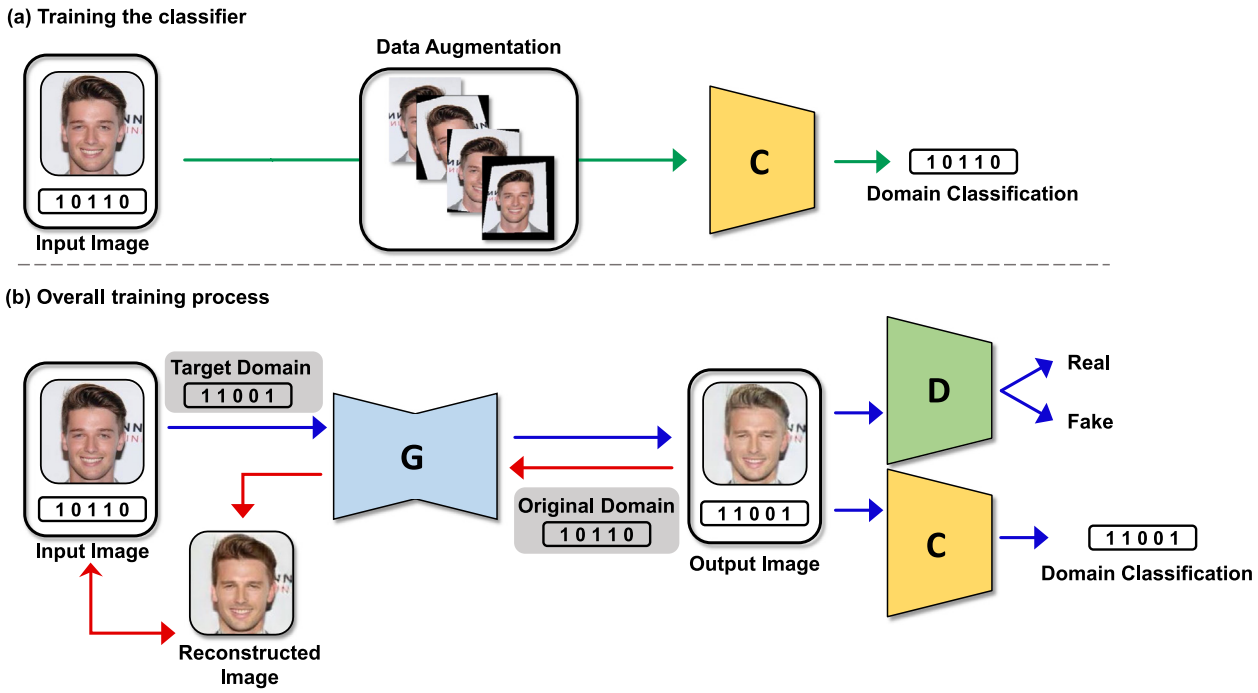

Fig. 1. Overall structure of SuperstarGAN: (a) The classifier is trained with real images with data augmentation. (b) The generator produces a fake image with an input image and a target domain label; then, the generated image is evaluated by the discriminator and classifier in terms of genuinity and class-accordance, respectively. The generator reconstructs the fake image with the original domain label. Then, the reconstructed image is compared with the original image. The discriminator tries to distinguish whether the image is real or fake. The classifier classifies domain of generated image.

demonstrated in ControlGAN (Lee & Seok, 2019), the ACGAN discriminator is not trained with DA techniques in general since the performance decreases. Conversely, the independent classifier can be trained with DA techniques without affecting the GAN training. This advantage of the ControlGAN framework enables the generator to be more effectively trained with the domain information by the fine classifier with DA. The classifier does not suffer from overfitting and can capture domain-invariant and domain-specific features. Consequently, SuperstarGAN has the capacity to learn mappings among large-scale domains and express small feature changes.

3.1. Generator

The SuperstarGAN generator produces a target-domain image with a source image and a target-domain vector. For the generator, the objective function is the same as that of StarGAN:

$$\hat{G} = \arg\min_{G} \{L_{adv}(r, D(G(\boldsymbol{X}, T))) + \lambda_{cls} \cdot L_{cls}(T, C_D(G(\boldsymbol{X}, T))) + \lambda_{rec} \cdot L_{rec}(\boldsymbol{X}, G(G(\boldsymbol{X}, T), T'))\}.$$

There are three different types of loss functions that the generator employs. The first is called adversarial loss, which deceives the discriminator into determining that a fake image is real. Minimizing this term means that the generator can create more realistic fake images. The second loss function is the classification loss, which is used to ensure that the generated image has the desired characteristics of the target domain. This is done by feeding the generated image into the classifier and comparing the correct label to the probability that the classifier returns. The more significant the difference between these two values, the higher the classification loss. By adopting this term, generated images will be classified as belonging to target domains by the classifier. Finally, there is reconstruction loss, which helps to preserve the identity features of the input image; the two aforementioned loss functions cannot help the generated images maintain features other than the target feature. The reconstruction loss is used to keep the other features the same as the input image. This can be done by translating the image twice, once with an arbitrary domain label and once with the original domain label.

3.2. Discriminator

Distinct from StarGAN, SuperstarGAN employs the vanilla discriminator, which only determines whether the input image is real or fake. This is because, in SuperstarGAN, the classifier is detached from the discriminator. Hence, the discriminator is only optimized with adversarial loss as follows:

$$\widehat{D} = \arg\min_{D} \{L_{adv}(f, D(G(\boldsymbol{X}, T))) + L_{adv}(r, D(\boldsymbol{X}))\}.$$

3.3. Classifier

The SuperstarGAN classifier is trained with DA in order to handle the overfitting problem that the StarGAN discriminator presents. The StarGAN classifier is improved with this modification, resulting in more accurate domain representations and improved learning for the generator. The training of the classifier can be represented as follows:

$$\hat{C} = \arg\min_{C} \{L_{cls}(T, C(\boldsymbol{X}_{aug}))\}.$$

We designed a classification loss function that compares the target label ($\boldsymbol{L}$) with the values obtained from the classifier. As for L_{cls}, the categorical cross-entropy is used for multiclass datasets, and the binary cross-entropy is used for multilabel datasets. The classifier updates its parameters with the classification loss using modified real images with DA. When the classifier receives the images, the classifier returns the probabilities of belonging to each label. Then, the cross-entropy is calculated with the probabilities and target labels. The parameters of the classifier are updated to minimize the cross-entropy. This process to train the classifier is essentially the same as that of conventional deep learning classifiers with DA.

Table 2
Datasets used in the experiments. The average number of samples per label is rounded.

Name	Num. of samples	Num. of labels	Average Num. of samples per label	Label category	Training iterations
CelebA	200,599	40	5015	Multi-label	1,000,000
AFHQ	14,630	7	2090	Multi-class	300,000
14Painters	9000	14	643	Multi-class	350,000

3.4. Training specifications

To compare the performance of image-to-image translation models, we adopted the same training process as StarGAN (He et al., 2016; Li & Wand, 2016; Ulyanov et al., 2016), but several modifications are specified in detail. In the training of SuperstarGAN, we used the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.0$ and $\beta_2 = 0.9$ for the generator and discriminator, and $\beta_1 = 0.9$ and $\beta_2 = 0.9$ for the classifier. The number of iterations was set at 1.0×10^6. The learning rates for the generator, discriminator, and classifier were set at 0.0001, 0.0001, and 0.00012, respectively. The learning rates of each component linearly decayed toward zero after 1.0×10^5 iterations were left.

For training the generator, λ_{cls} and λ_{rec} were set to 0.25 and 1.3, respectively. The hinge loss (Lim & Ye, 2017) adapted to WGAN-GP (Arjovsky et al., 2017; Gulrajani et al., 2017) objective function was applied as the adversarial loss of the discriminator training. The λ_{gp} for the gradient penalty was set to 10. The spectral normalization (Miyato et al., 2018) was employed as the weight normalization. For the DA used in the classifier, we use two transformations: the random horizontal flip with a probability of 0.5 and the random rotation at an angle of -20 to 20.

The resolution of the real images was reduced to 128×128. Thus, the synthetic images by generators have the same resolution. For a fair comparison, we adopted the same architecture as that of StarGAN: The generator is composed of 18 convolutional layers with 64 kernels, where the first and the last three layers are regular convolutional layers, and the other intermediate layers are six residual blocks with two convolutional layers each. There are two downsampling and upsampling layers; thereby, the residual blocks are applied to 32×32 feature maps. The discriminator consists of eight convolutional layers. In the first seven convolutional layers, the stride was set at two, resulting in downsamplings. The last layer corresponds to the output layer of the discriminator, which determines whether the input images are real or fake.

4. Experiments and results

4.1. Datasets

The CelebA (Liu et al., 2015) is a large-scale face dataset of celebrity images comprising approximately 200,000 facial images, each of which has 40 binary attributes, such as hair color and facial expression. For additional experiments, we used two additional datasets. One is the high-quality animal face dataset called Animal Faces HQ (AFHQ), which has been used in the experiments of StarGANv2 (Choi et al., 2020). The dataset consists of three domains: dogs, cats, and wildlife. For this study, we self-labeled the wildlife domain into five additional domains: dots, fox, lion, tiger, and wolf. The dots class includes animals with dot patterns, such as the leopard. The total number of images in the dataset is 15,000, separated into training and testing sets for each domain. Another dataset is the 14Painters, which is a collection of paintings from Wikiart.org. We adopted the sorted version used in ComboGAN (Anoosheh et al., 2018), where 9283 paintings belong to 14 different artists. The artists contained in the dataset are as follows: Zdzislaw Beksinski, Eugene Boudin, David Burliuk, Paul Cezanne, Marc Chagall, Jean-Baptiste-Camille Corot, Eyvind Earle, Paul Gauguin, Childe Hassam, Isaac Levitan, Claude Monet, Pablo Picasso, Ukiyo-e (style, not person), and Vincent Van Gogh. The specifications of the datasets used in the experiments are shown in Table 2.

4.2. Baseline models

The StarGAN was selected as a baseline of the experiments since the proposed model is a modification of StarGAN. Additionally, three baselines were considered in order to demonstrate that the improvement of SuperstarGAN has not resulted from other modifications than the proposed framework: First, SuperstarGAN without DA shows that DA is crucial for avoiding the overfitting problem while performance is improved by simply separating the classifier from the discriminator. Second, StarGAN+SN+Hinge baseline introduced spectral normalization (Miyato et al., 2018) and hinge loss for the structures of StarGAN and a loss function for the StarGAN discriminator. This baseline shows that the improved performance of SuperstarGAN is not due to spectral normalization or hinge loss. Third, StarGAN+DA used DA to train the vanilla StarGAN. We adopted this baseline model to demonstrate that using DA on the vanilla StarGAN hinders the training.

4.3. Experimental results

4.3.1. Qualitative evaluation

The SuperstarGAN and following baseline models were trained on the CelebA dataset for comparison. To evaluate the ability of large-scale domains, we designed the experiment in which models simultaneously learn the mapping of all 40 CelebA attribute labels, whereas a few labels were selected in the experiments of the StarGAN study (Choi et al., 2018). In the training phase, each model was trained by changing a single attribute of the source image.

As Fig. 2 demonstrates, the SuperstarGAN was able to learn the mappings for 40 labels. The generated images from SuperstarGAN express the target domain features precisely. Even though training only occurred with a single attribute transfer, SuperstarGAN was still able to perform multiple facial attribute changes effectively. However, this was not the case for the images from the baseline models, as many features were not properly expressed. In terms of visual quality, SuperstarGAN is superior to other baseline models. Furthermore, SuperstarGAN is able to translate images while still preserving the identity features of the source image. In contrast, facial features and the outline for images from the baseline models are often blurred. Additionally, in many cases, the baseline models cannot provide image translation at all as they failed to maintain the overall structure of the source image.

We hypothesized that a well-trained independent classifier can improve performance as generator training becomes precise. In other words, the independent classifier allows the generator and discriminator to concentrate on generating and detecting realistic fake samples, which are their fundamental objectives. Therefore, the classification performance of the fine classifier enables the generator to learn to map for large-scale target domains and generate realistic samples that precisely follows the target domains.

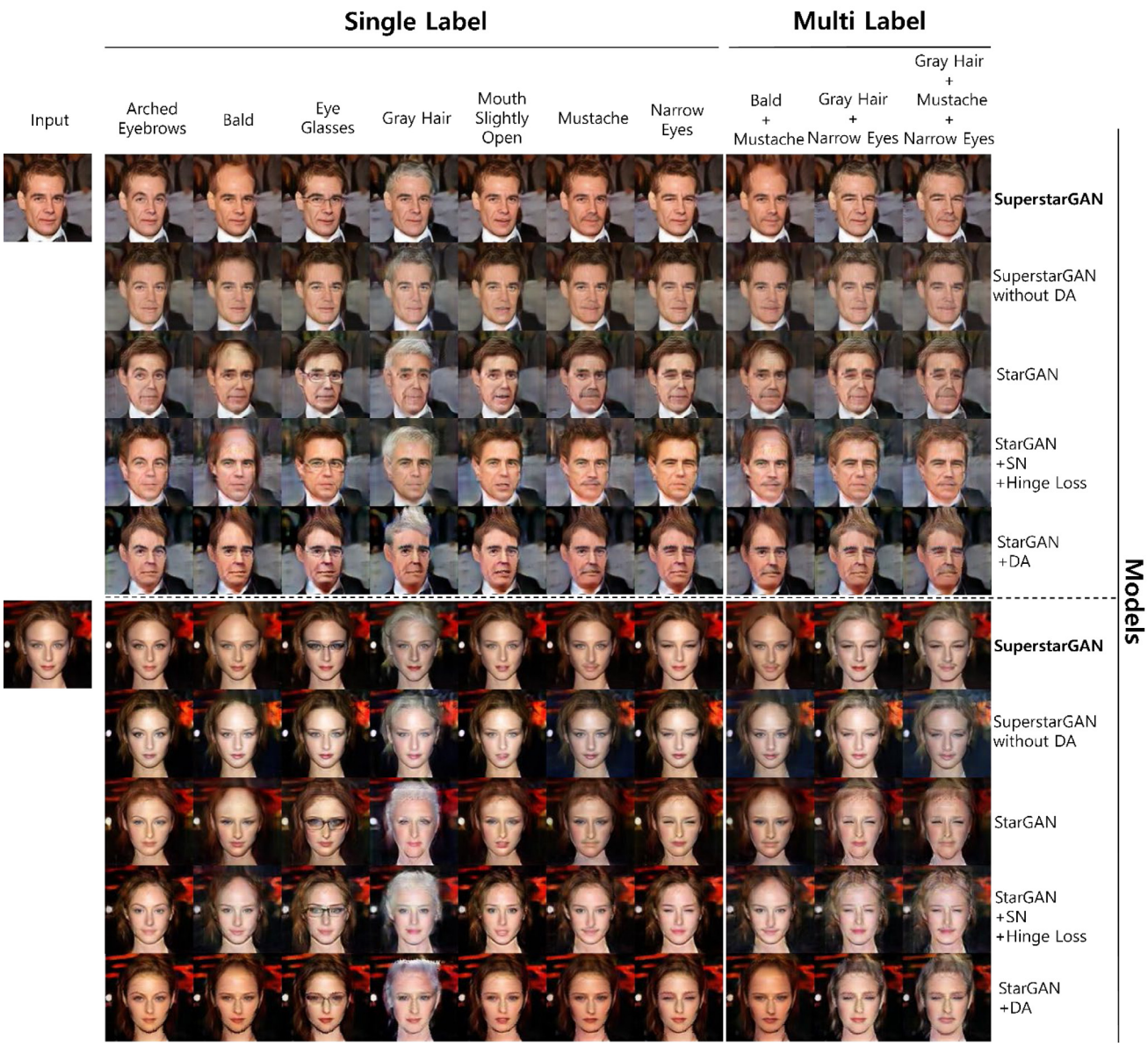

Fig. 2. Image-to-image translation in large-scale domains with the CelebA dataset: Comparison of the SuperstarGAN and four other baselines. SN: spectral normalization, DA: data augmentation.

Table 3
Frechet Inception distance (FID) and learned perceptual image patch similarity (LPIPS) of the SuperstarGAN and baselines.

Method	FID↓	LPIPS↓
SuperstarGAN without DA	29.1	0.134
StarGAN	27.7	0.181
StarGAN+SN+Hinge	34.0	0.188
StarGAN+DA	38.3	0.201
SuperstarGAN	**22.7**	**0.104**

4.3.2. Quantitative evaluation

To evaluate the quantitative performance of the models, we adopted two evaluation metrics: FID (Heusel et al., 2017) and LPIPS (Zhang, Isola et al., 2018). The FID is a widely used metric for measuring feature distance between real and generated images. Specifically, the distance between two distributions was measured using feature vectors extracted from Inception-v3 (Szegedy et al., 2016) pretrained with ImageNet. In this comparison, we calculated FID with 10,000 real images from the CelebA dataset and 10,000 generated images by each model. The generated images were composed of 250 real images, which were translated into all 40 domains of CelebA. The source images of the generated samples did not include the real images to be compared against. A low FID score indicates a high similarity between generated images and real images.

The LPIPS measures similarity based on human perception using AlexNet (Krizhevsky et al., 2017) pretrained with ImageNet. In this study, to score LPIPS, we compared 2000 pairs of the generated images and the corresponding source images. We calculated the average values after comparing the paired images. A low LPIPS score signifies that the two images are perceptually similar.

Table 3 exhibits the FID and LPIPS of SuperstarGAN and the other baselines. As a result, SuperstarGAN showed an FID of 22.7 and an LPIPS of 0.104, corresponding to decreased FID and LPIPS by 18.1% and 42.5%, respectively, compared to StarGAN. As expected, StarGAN+DA showed inferior performance compared to StarGAN, indicating that DA can decrease the performance

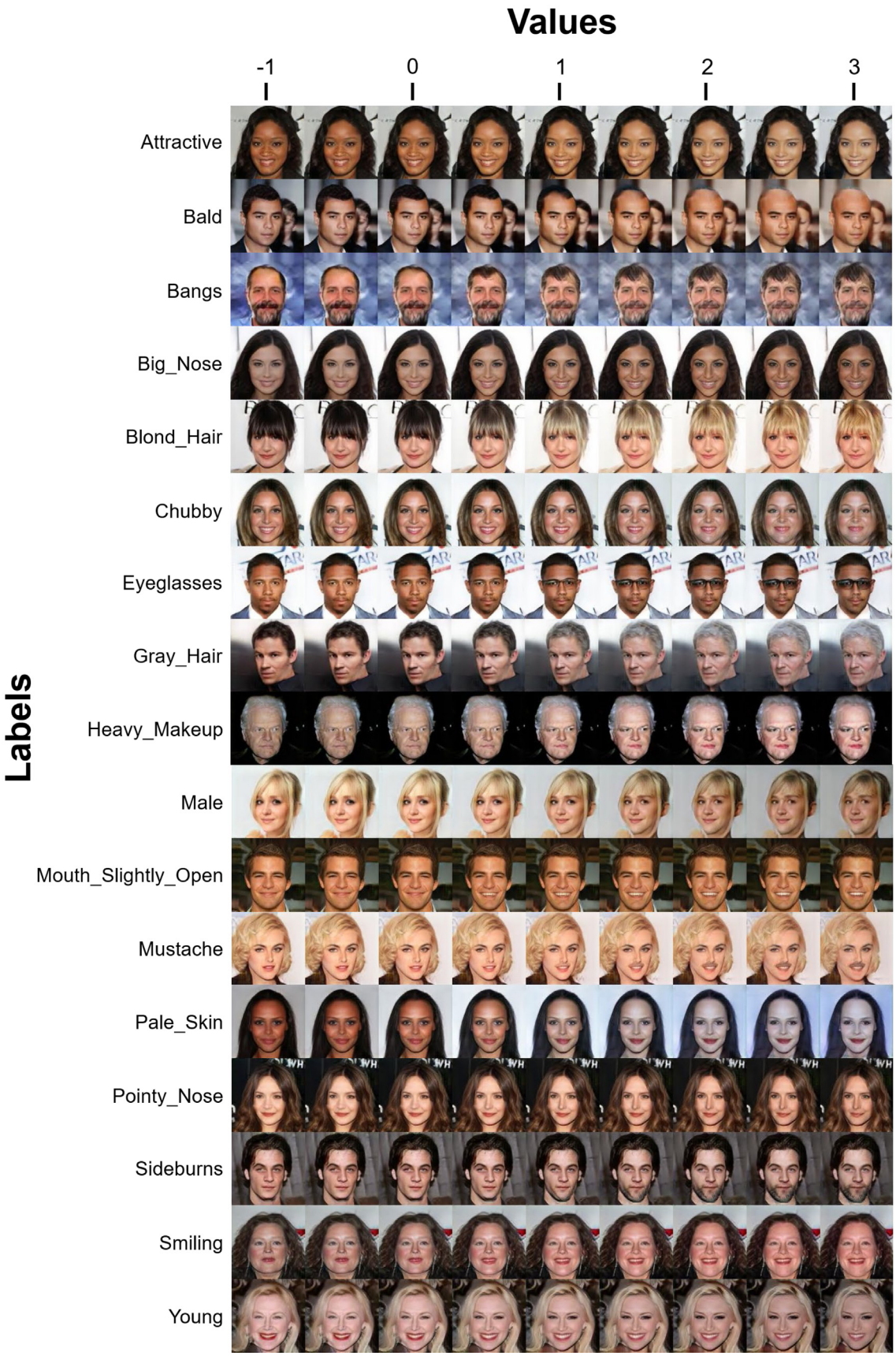

Fig. 3. Generated images with various target domains using interpolated and extrapolated label values.

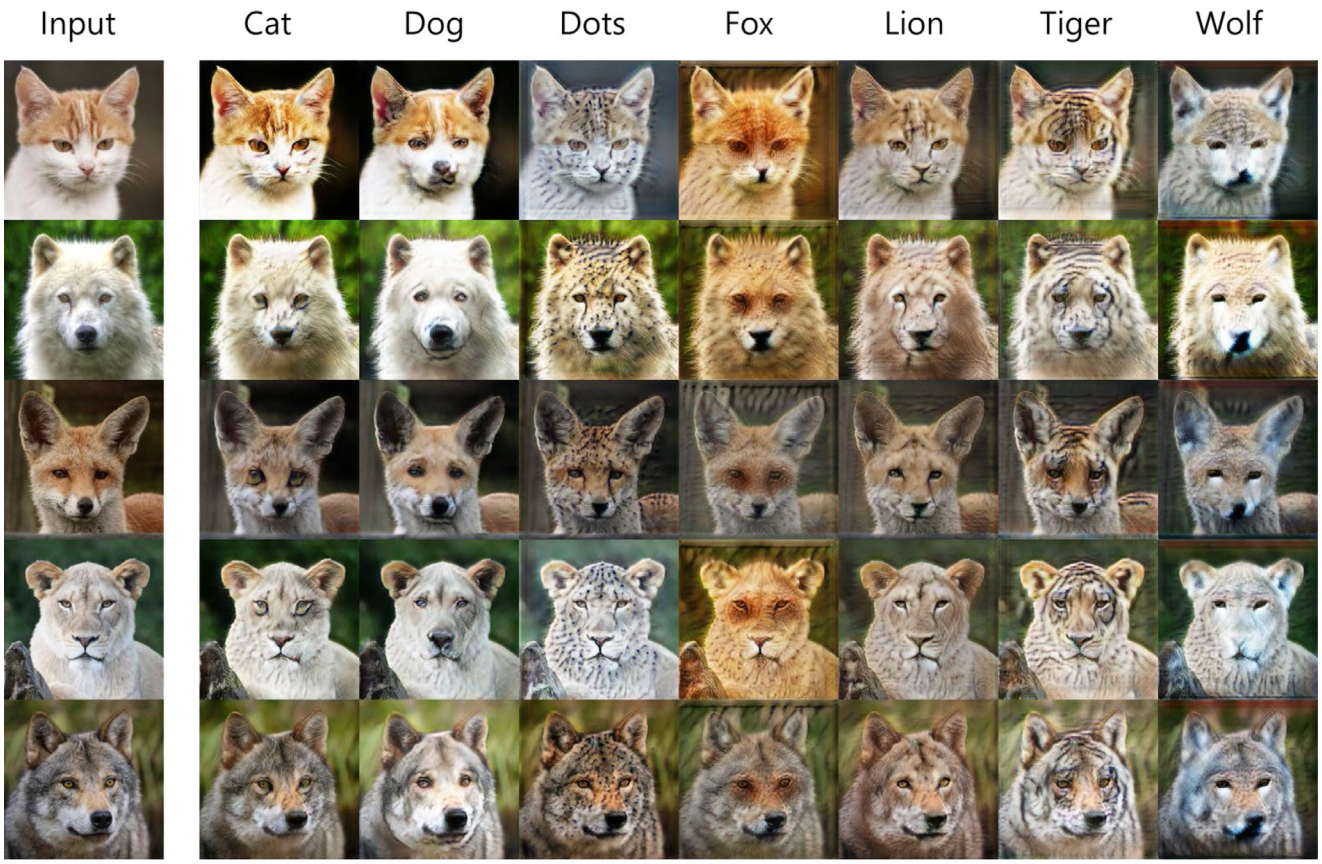

Fig. 4. Animal facial style translations using SuperstarGAN with the AFHQ dataset.

when it is used with StarGAN architecture. Additionally, SuperstarGAN outperformed StarGAN+SN+Hinge, which implies that the improved performance of SuperstarGAN was not due to spectral normalization and hinge loss. These results indicate that SuperstarGAN outperformed the other baselines.

We numerically demonstrated that SuperstarGAN outperformed other baselines in generating realistic samples. Additionally, as we employed images translated into all 40 labels for scoring FID, such results indicate that SuperstarGAN has a high capacity for various transformations. Also, these results signify that generated images from SuperstarGAN were more perceptually similar to the real images than those of the other baselines. In addition, LPIPS, which assesses the GAN model based on human perception, verified that SuperstarGAN can generate visually plausible images.

4.3.3. Interpolation and extrapolation with domain features

We conducted experiments with both interpolated and extrapolated label values. We found that SuperstarGAN, when trained only with binary label values (zero and one), was able to adjust the degree of target domain features. Specifically, the intermediate features could be expressed using interpolated values between zero and one. Furthermore, the proposed method was able to emphasize the target domain features using high extrapolated values. Additionally, we found that even opposite features could be generated using negative values.

The results in Fig. 3 show that SuperstarGAN can control the degree to which facial features are expressed by adjusting the size of the input label value. Target domain features become increasingly apparent as the label value increases. We also observed an intermediate feature expression image where a half-bang image was generated with a "Bang" label value of 0.5. In addition, interesting results were demonstrated when the label value was negative (i.e., -1). For example, when the "Big Nose" label had a value of -1 as the conditional input, the corresponding output had a small nose. While there was no "Small Nose" label in CelebA dataset, this implies that the proposed method allows learning an untrained opposite feature implicitly.

4.4. Experimental on AFHQ and 14Painters

To demonstrate that SuperstarGAN generally works regardless of datasets, we trained SuperstarGAN using two additional datasets, AFHQ and 14Painters. We adopted the same model architecture as that used on the CelebA dataset. The AFHQ and the 14Painters are multiclass datasets; therefore, categorical cross-entropy loss was used to train the classifier. In this experiment, we set λ_{cls} at 0.3 and λ_{rec} at 0.05. To avoid mode collapse, we decreased the number of iterations from 1,000,000 to 300,000 and 350,000 in the training on the AFHQ and the 14Painters, respectively. As shown in Table 2, relatively small datasets in terms of the number of labels require fewer training iterations. Additionally, we experimentally demonstrated that SuperstarGAN can learn mappings among multi-domain with a small size of training samples in each label. While SuperstarGAN uses 5015 CelebA images per label, there are only 2090 and 643 samples per label in AFHQ and 14Painters, respectively. We believe that this is because of the proposed independent classifier trained with data augmentation; even if a small number of samples are given, fine classifier allows the models to learn domain features. Figs. 4 and 5 illustrate that SuperstarGAN also performs properly on several datasets, which explains the generality of the SuperstarGAN performance.

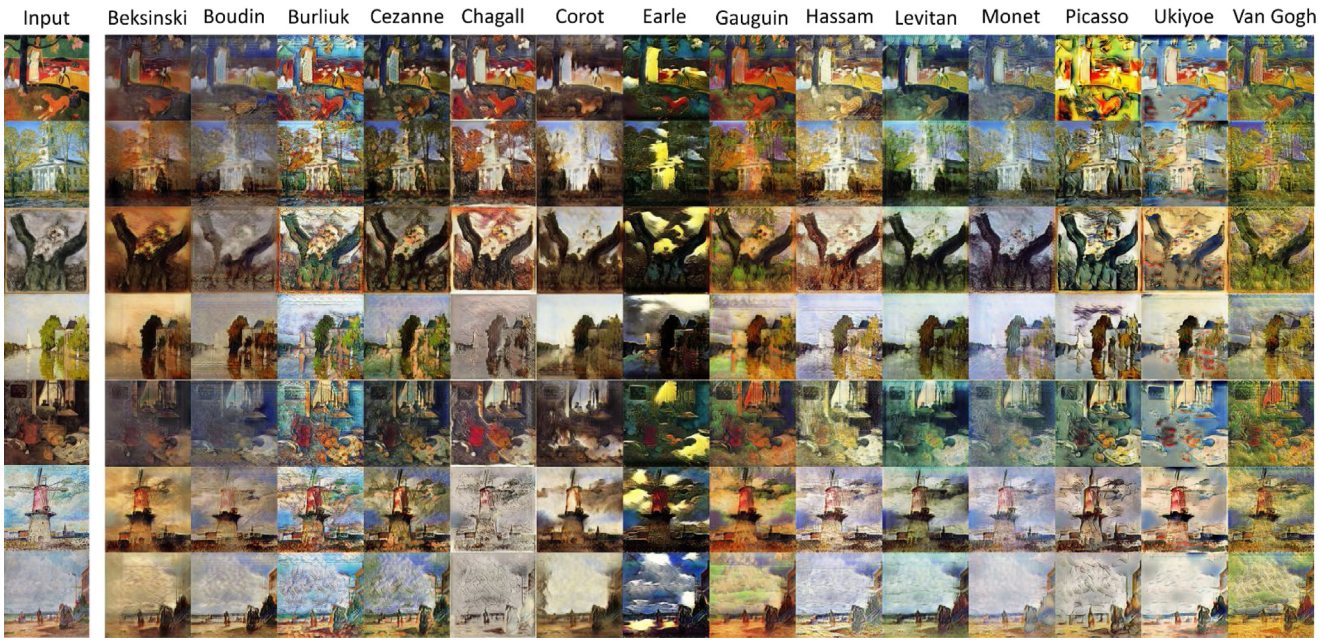

Fig. 5. Style transfer results on the 14Painters dataset.

5. Conclusion

In this paper, we proposed SuperstarGAN, an improved StarGAN for large-scale domains. While the conventional StarGAN has a limitation of training with large-scale domains due to the ACGAN discriminator, SuperstarGAN introduced the ControlGAN framework to address the limitation. In the experiments, we demonstrated that SuperstarGAN is superior to the baselines in terms of visual quality and evaluation metrics. It was confirmed that SuperstarGAN can generate high-quality images that follow the given conditions. Evaluated with the CelebA dataset, the FID and LPIPS of SuperstarGAN were 22.7 and 0.104, respectively, corresponding to decreased FID and LPIPS by 18.1% and 42.5%, respectively, compared to StarGAN. By adjusting SuperstarGAN to two additional datasets, we demonstrated the generality of the proposed method. Furthermore, an additional experiment with interpolated and extrapolated label values revealed that SuperstarGAN effectively generates diverse samples, including exaggerated images and opposite images.

CRediT authorship contribution statement

Kanghyeok Ko: Conceptualization, Software, Formal analysis, Investigation, Writing – original draft. **Taesun Yeom:** Software, Investigation, Formal analysis, Investigation. **Minhyeok Lee:** Conceptualization, Investigation, Writing – original draft, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2022 as well as the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1F1A1050977).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2023.02.042.

References

Anoosheh, A., Agustsson, E., Timofte, R., & Van Gool, L. (2018). Combogan: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., & Yang, B. (2020). Medgan: Medical image translation using GANs. *Computerized Medical Imaging and Graphics*, *79*, Article 101684. http://dx.doi.org/10.1016/j.compmedimag.2019.101684.

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. http://dx.doi.org/10.48550/arXiv.1809.11096, arXiv preprint arXiv:1809.11096.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, 29.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, 30.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2019). Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, *28*(11), 5464–5478. http://dx.doi.org/10.1109/TIP.2019.2916751.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Hu, B., Zheng, Z., Liu, P., Yang, W., & Ren, M. (2020). Unsupervised eyeglasses removal in the wild. *IEEE Transactions on Cybernetics*, *51*(9), 4373–4385. http://dx.doi.org/10.1109/TCYB.2020.2995496.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. http://dx.doi.org/10.48550/arXiv.1710.10196, arXiv preprint arXiv:1710.10196.

Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*.

Kim, W., Kim, S., Lee, M., & Seok, J. (2022). Inverse design of nanophotonic devices using generative adversarial networks. *Engineering Applications of Artificial Intelligence*, *115*, Article 105259. http://dx.doi.org/10.1016/j.engappai.2022.105259.

Kim, K., Park, S., Jeon, E., Kim, T., & Kim, D. (2022). A style-aware discriminator for controllable image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. http://dx.doi.org/10.1145/3065386.

Lee, M., & Seok, J. (2019). Controllable generative adversarial network. *IEEE Access*, *7*, 28158–28169.

Li, C., & Wand, M. (2016). Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*.

Lim, J. H., & Ye, J. C. (2017). Geometric gan. http://dx.doi.org/10.48550/arXiv.1705.02894, arXiv preprint arXiv:1705.02894.

Liu, Y. (2021). Improved generative adversarial network and its application in image oil painting style transfer. *Image and Vision Computing*, *105*, Article 104087. http://dx.doi.org/10.1016/j.imavis.2020.104087.

Liu, M.-Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. *Advances in Neural Information Processing Systems*, (30).

Liu, R., Ge, Y., Choi, C. L., Wang, X., & Li, H. (2021). Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International conference on learning representations*.

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*.

Ojha, U., Li, Y., Lu, J., Efros, A. A., Lee, Y. J., Shechtman, E., & Zhang, R. (2021). Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Pang, Y., Lin, J., Qin, T., & Chen, Z. (2021). Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia*, http://dx.doi.org/10.1109/TMM.2021.3109419.

Park, M., Lee, M., & Yu, S. (2022). HRGAN: A generative adversarial network producing higher-resolution images than training sets. *Sensors*, *22*(4), 1435. http://dx.doi.org/10.3390/s22041435.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Srivastava, A., Chanda, S., & Pal, U. (2022). AGA-gan: Attribute guided attention generative adversarial network with U-net for face hallucination. *Image and Vision Computing*, *126*, Article 104534. http://dx.doi.org/10.1016/j.imavis.2022.104534.

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

Taigman, Y., Polyak, A., & Wolf, L. (2016). Unsupervised cross-domain image generation. http://dx.doi.org/10.48550/arXiv.1611.02200, arXiv preprint arXiv:1611.02200.

Toshpulatov, M., Lee, W., & Lee, S. (2021). Generative adversarial networks and their application to 3D face generation: A survey. *Image and Vision Computing*, *108*, Article 104119. http://dx.doi.org/10.1016/j.imavis.2021.104119.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. http://dx.doi.org/10.48550/arXiv.1607.08022, arXiv preprint arXiv:1607.08022.

Xie, S., Ho, Q., & Zhang, K. Unsupervised image-to-image translation with density changing regularization. Advances in Neural Information Processing Systems.

Yang, S., Jiang, L., Liu, Z., & Loy, C. C. (2022). Unsupervised image-to-image translation with generative prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Yuan, Q.-L., & Zhang, H.-L. (2022). RAMT-GAN: Realistic and accurate makeup transfer with generative adversarial network. *Image and Vision Computing*, *120*, Article 104400. http://dx.doi.org/10.1016/j.imavis.2022.104400.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhang, G., Kan, M., Shan, S., & Chen, X. (2018). Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision*.

Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., & Efros, A. A. (2016). Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*.

Zhuang, P., Koyejo, O. O., & Schwing, A. (2020). Enjoy your editing: Controllable GANs for image editing via latent space navigation. In *International conference on learning representations*.