IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 28, 2021, accepted October 22, 2021, date of publication October 26, 2021, date of current version October 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122834

Controllable Image Dataset Construction Using Conditionally Transformed Inputs in Generative Adversarial Networks

FARKHOD MAKHMUDKHUJAEV[1], (Member, IEEE), JUNSEOK KWON[2], (Member, IEEE), AND IN KYU PARK[1], (Senior Member, IEEE)

[1]Department of Information and Communication Engineering, Inha University, Incheon 22212, South Korea
[2]School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: In Kyu Park (pik@inha.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant by the Korean Government through the MSIT under Grant NRF-2019R1A2C1006706, in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant by the Korean Government through the MSIT (Artificial Intelligence Convergence Research Center, Inha University) under Grant 2020-0-01389, and in part by Inha University Research Grant.

ABSTRACT In this paper, we tackle the well-known problem of dataset construction from the point of its generation using generative adversarial networks (GAN). As semantic information of the dataset should have a proper alignment with images, controlling the image generation process of GAN comes to the first position. Considering this, we focus on conditioning the generative process by solely utilizing conditional information to achieve reliable control over the image generation. Unlike the existing works that consider the input (noise or image) in conjunction with conditions, our work considers transforming the input directly to the conditional space by utilizing the given conditions only. By doing so, we reveal the relations between conditions to determine their distinct and reliable feature space without the impact of input information. To fully leverage the conditional information, we propose a novel architectural framework (i.e., conditional transformation) that aims to learn features only from a set of conditions for guiding a generative model by transforming the input to the generator. Such an approach enables controlling the generator by setting its inputs according to the specific conditions necessary for semantically correct image generation. Given that the framework operates at the initial stage of generation, it can be plugged into any existing generative models and trained in an end-to-end manner together with the generator. Extensive experiments on various tasks, such as novel image synthesis and image-to-image translation, demonstrate that the conditional transformation of inputs facilitates solid control over the image generation process and thus shows its applicability for use in dataset construction.

INDEX TERMS Dataset construction, conditional image generation, generative adversarial networks, conditional transformation.

I. INTRODUCTION

The dataset is a key element in teaching a learning-based method to understand real-world scenarios. However, datasets often lack a sufficient number of samples necessary for the training stage; whereas, training highly efficient deep learning algorithms require large-scale dataset covering up wide-range of variations. In this circumstance, dataset construction is an obvious way to overcome this problem. However, construction itself is a tedious process demanding not only time but also finance. Recently, generative adversarial networks (GAN) [1] have produced prominent results in image synthesis [2]–[7], super-resolution [8]–[11], image-to-image translation [12]–[16], image style transfer [17], [18], segmentation [19], and lossy compression [20]. Indeed, such successive applications reveal the potential of using GAN as a powerful base for image generation and further dataset construction. GAN [1] is trained using a zero-sum non-cooperative game between its generator and discriminator modules. Such a competitive process allows the generation of realistic and sharp images from lower- or higher-dimensional inputs (noise or image) by minimizing

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

the discrepancy between the learned and real distributions. Several works [12], [21]–[25] have demonstrated the possibility of augmenting GAN with side information to control the generation and make it consistent with the existing context. Additionally, it has been shown that introducing the side information as a conditioning factor improves GAN performance [23], [26].

Over the past years, several attempts have been made to control the image generation process of GAN. A concatenation of the input and condition at the initial stage of the generator module is the earliest approach; it extends the vanilla GAN to conditional GAN (cGAN) setup [21]. Such approach helps to create a deterministic relationship between the input and output and thus achieve control over the generation. A strategy of cGAN has also been successfully used in image-to-image translation tasks [12], [27]–[31], which seek to transform the source images into different domains by considering image-based conditions. A few works [32], [33] conditioned the generator by concatenating an initial-level information with a learned representation of conditions obtained through a linear layer.

By contrast, several other methods [3], [17], [26], [34]–[37] introduce conditional information (occasionally with noise) into the hidden layers of the generator through normalization technique [38] by replacing the non-adaptive parameters (i.e., scale and shift) with input-dependent ones. Notably, these parameters are learned based on conditions by utilizing the embedding functions. Alternative normalization techniques [39], [40] for inserting conditional context to the generation process are utilized in these representative works [5], [7], [14], [41].

Another practical approach used in existing works [42], [43] relies on statistical information (i.e., mean and variance) that captures the meaning of given conditions. This information is also concatenated with randomly sampled noise to be fed as input to the generator. In a similar manner, [44] obtained such statistical information based on the concatenated input of noise and conditions, and used to sample latent variable for image generations.

The aforementioned methods generate images based on the concatenated representation, which means two substantial pieces of information for the image generation process are simply and directly utilized in conjunction. Intuitively, a condition can be regarded as a controlling factor that defines the context, whereas the input (as noise or image) is responsible for the diversity and fidelity of the generated images. In this case, the generator that performs mapping operations should take the entire obligation of learning higher-order interactions between the input and condition and provide reliable features for subsequent layers. It has been reported that the generator conditioned on the given context ignores random noise information [27], [28], [45]. Moreover, there might be the case when the condition is not a single class label, but in the form of multiple class labels, which could increase the complexity of the learning process. An example is the generation of facial images with multiple attributes corresponding to various genders, ages, and expression classes.

In consideration of these statements, we consider conditioning the generator from a different perspective where conditional information can be utilized on its own. Motivated by this consideration, we propose a novel architectural addition to GAN called conditional transformation (CT) framework. Unlike previous works, this framework focuses on using only the given conditional information, such that the conditions control the generation process. A novelty of our framework is that it learns relations between conditions to determine their conditional feature space and utilizes this information to transform a given input as a function of specific conditions. As conditional transformation operates on the input layer, it can be used along with diverse GAN by effortlessly prepending to the most generator networks.

The contributions of this work are as follows:

- We present a simple and yet efficient framework that provides reliable control over the generation process for image synthesis and image-to-image translation tasks.
- The framework enables the transformation of the generator inputs to be specific to conditions (e.g., single/multiple class), thereby facilitating the generation of semantically desired images.
- Using the framework, we demonstrate the qualitative and quantitative improvements of state-of-the-art works in their respective tasks.
- Our framework effectively controls the condition-specific image generation and thus can be an alternative to massive data augmentation in construction of image dataset.

The rest of the paper is organized as follows. Section II provides a brief discussion on existing works that dealt with conditional image generation. Section III describes the proposed framework in a detailed manner. In Section IV, we justify the efficiency of the proposed framework through various experiments. We conclude this work by presenting our considerations over the conditional image generation in Section V.

II. RELATED WORKS

Conditional GAN (cGAN) [21] is a pioneering approach for conditioning generators with additional information. The conditioning process is performed by simply feeding side information c as an additional input where it is combined with noise z in joint hidden representation (Fig. 1a). Subsequently, several works [3], [26], [33], [35] started adapting advanced strategies, such as using linear and/or embedding layers, to condition the generator. LSGAN [33] utilized a linear layer to obtain compact representations of a large number of class vectors to concatenate with the input noise. The embedding layer is another extensively applied strategy [3], [26], [35]. In [3] and [35], embedding layers were utilized to obtain scaling and shifting parameters (i.e., γ and β, respectively) of label information for injection into a generator

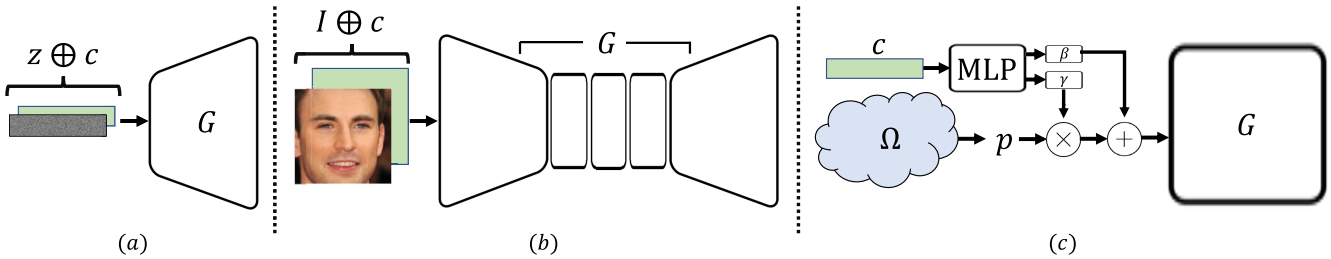


FIGURE 1. Conceptual comparison of methods for conditioning the generator network *G*. (a) Conventional approach using a concatenation of input noise *z* with conditions *c* for image synthesis tasks, (b) conventional approach using a concatenation of input images *I* with conditions *c* for image-to-image translation tasks, and (c) our proposal on conditionally transforming the input *p* sampled from random or image distribution Ω. The ⊕ sign stands for concatenation.

using conditional batch normalization layers (CBN) [17], [37]. Similarly, sBN [26] modulated the intermediate feature maps of the generator through batch normalization [38] using the same γ and β obtained by considering a bi-linear interaction between z and two trainable embedding functions of the class labels. Another commonly applied approach is adaptive instance normalization (AdaIN) [39], and spatially adaptive normalization (SPADE) [40]. These normalization techniques have been successfully applied in StyleGAN [5], StarGAN v2 [14], AMGAN [41], etc. In fact, AdaIN is mostly used in image generation tasks to normalize the layer-wise feature maps. For that purpose, AdaIN is placed at every layer of the generative network at the cost of computational complexity and the resources necessary for image generation. Additionally, such usage of AdaIN may lead to the water droplet artifacts in the novel image synthesis [7]. As a practical remedy, StackGAN [42] and its improved version [43] attempted to learn the mean and variance parameters of given conditions to sample the input to the generator. Although conditions are being utilized separately, they are concatenated with noise at the input level. Similarly, VCGAN [44] applied a strategy where mean μ and covariance Σ are estimated using linear layers for sampling the latent variable as an input to the generator given the condition along with noise.

In the domain of image-to-image translation, the following strategies for conditioning the generator have been applied. Earlier work [24], Invertible cGANs, were equipped with two independent encoders to invert given input image into latent representation and conditional information and further use their concatenation in the settings of cGANs [21]; meanwhile, the variations in attribute information were applied to generate a modified image. Inspired by such conditional positioning, the authors of [25] conditioned CycleGAN [46] for guiding image translation task. Particularly, at the input layer of G, the input image was concatenated with a conditional vector that was resized to match image dimensions. StarGAN [12] adopted this strategy in the channel-wise concatenation of an image with a condition but aims to train a single generator and discriminator (Fig. 1b). FEGAN [47] utilized a regression model to construct the attribute axis from which obtained latent vector with the target attributes for generator network.

III. CONDITIONAL TRANSFORMATION FRAMEWORK

In this section, we provide a substantial description of our proposed method. Specifically, we start from the problem formulation of conditional image generation and then move to explain in detail our proposed conditional transformation framework for the generator network while presenting the analysis of its workflow.

A. PROBLEM FORMULATION

To control the condition-specific image generation process of the generator G, the proposed framework CT considers noise/image $p \in P$ and condition $c \in A$ drawn from respective distributions. Specifically, p is a task-dependent input that is to be mapped into a domain specified by condition c of attribute space A. Our framework aims to provide a transformed p, denoted by p', which is expected to possess characteristics of given conditional information c. Unlike the approaches using these inputs in a composed form to achieve control, our proposed framework transforms the input based on only conditions (Fig. 1(c)). The framework comprises two steps: conditional feature-space learning and conditional transformation of learned features on the input. Our intention is to learn features that best represent the conditional vector and apply this information to the input. Through such transformation, each input belongs to a particular discriminative space corresponding to a specific condition, which generates the desired image.

B. CONDITIONAL FEATURE-SPACE LEARNING

The first step is to learn the conditional feature-space from given labels in c. In particular, this step defines the discriminative space corresponding to specific conditions only. To find a feature space for each of the given conditions, we introduce a network consisting of an MLP in the form of fully connected layers with non-linear activation functions. We initially set the number of units at the input layer to be close to the number of conditional classes and increase this number by a factor of two at every subsequent layer till the last layer while considering the dimensions of p. In particular, we select this option because using higher-order interactions enables us to learn complex relations between features and improve the representational power of conditional space for each c. Moreover, such a construction can be regarded as

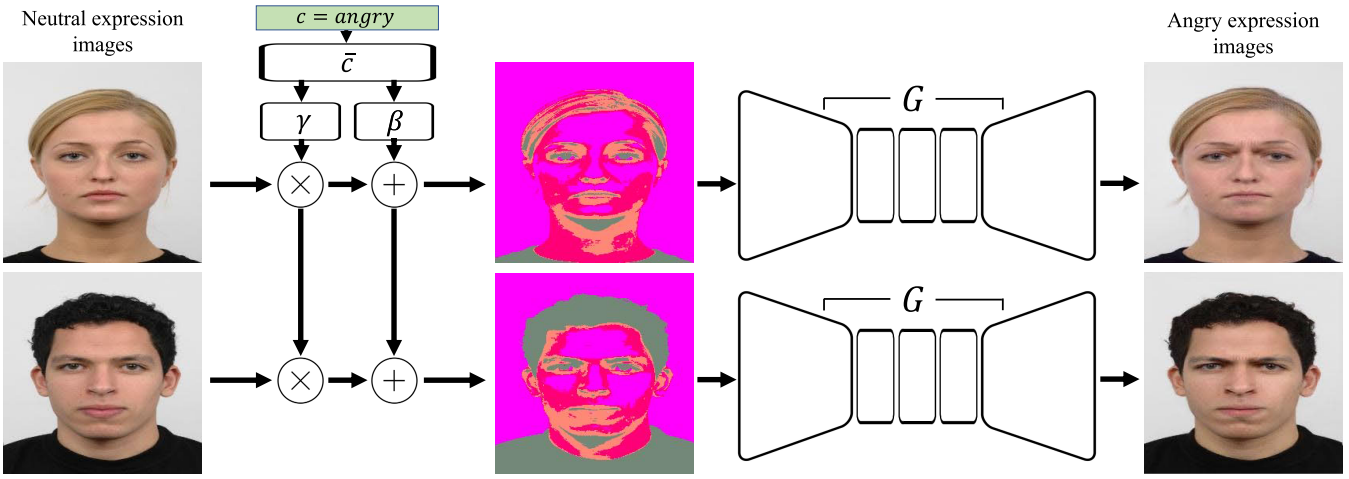


FIGURE 2. Illustration of the proposed framework for image-to-image translation task. Given input (neutral expression) images are being translated into angry expressions using our proposed framework.

that for an encoder because our aim is to learn feature-space encoding given a conditional vector. The learned conditional feature-space information can be expressed as follows:

$$\bar{c} = \sigma\left(\sum_i W_i * c_i + b\right) \tag{1}$$

where σ is a non-linear activation function (e.g., ReLU, Leaky ReLU), W are the learned weights of fully connected layers, b is a bias term, and c is a class label represented as one-hot vector through indicator function $\mathbb{1}_A(c)$.

$$c = \mathbb{1}_A(c) := \begin{cases} 1, & \text{if } c \in A, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

Simply, this function recasts the label to one dimensional vector containing 1 on element c and 0 elsewhere. Accordingly, c is a one-hot vector encoding a single-condition-based label. However, a dataset construction may require generating images that contain multi attributes. The structure of our framework allows controlling such a process in a straightforward manner. In this scenario, we can represent multi-condition-based labels as a concatenation of one-hot representations of single-class labels. For instance, to generate face image with conditions such as gender c_g, age c_a, and expression c_e, we can recast them separately using (2) and then form input as $c = c_g \oplus c_a \oplus c_e$, where $\oplus$ denotes concatenation operation, and apply (1) to obtain $\bar{c}$. This simple maneuver allows the proposed framework to learn the entanglement of various image attributes and effectively estimate the reliable conditional space $\bar{c}$ for a given set of labels.

C. CONDITIONAL TRANSFORMATION OF INPUT

Instead of concatenating the learned features of $\bar{c}$ directly to the input p, the second step of our framework transforms the input to the conditional feature-space information, such that the input to the generator is aligned with the desired conditions. We consider two independent linear functions $\gamma(\bar{c})$ and $\beta(\bar{c})$ to learn affine parameters, which respectively scale and shift p according to the conditional space features for generating images only from a corresponding single/multi conditions. We opt to implement them because multiplicative and additive modulations are typically applied operations [5], [17], [37]. Such a transformation can be expressed as follows:

$$p' = \gamma(\bar{c}) \odot p + \beta(\bar{c}) \tag{3}$$

where $\odot$ denotes element-wise multiplication. Here, channel-wise scaling factor $\gamma(\bar{c})$ and additive shifting $\beta(\bar{c})$ terms directly depend on $\bar{c}$ by

$$\gamma(\bar{c}) = \sum W_g * \bar{c}, \qquad \beta(\bar{c}) = \sum W_b * \bar{c} \tag{4}$$

where W_g and W_b are the learned weights of fully connected layers.

Subsequently, G learns the function of $f : p' \rightarrow x$ to generate an image based on conditionally transformed p'. Fig. 2 illustrates the conditional transformation process in the example of neutral-to-angry expression translations. Compared with existing works, the proposed framework can handle the conditioning of input according to not only a single-class label but also multi-class-based labels, thereby facilitating the learning of higher-order interactions and enabling a controllable and complex image generation.

D. LEARNING OBJECTIVES

Our framework aims to control the image generation process by conditionally transforming the inputs to the generator network. Because conditions are considered as a control factor, we need to ensure that framework is learning a proper condition-specific information of target distribution. This requires determining the objectives that provide certain information to supervise the framework. As framework operates in co-ordinance with generator, it can receive information through generator. The straightforward way is to provide a signal using class conditional classification. A cross-entropy

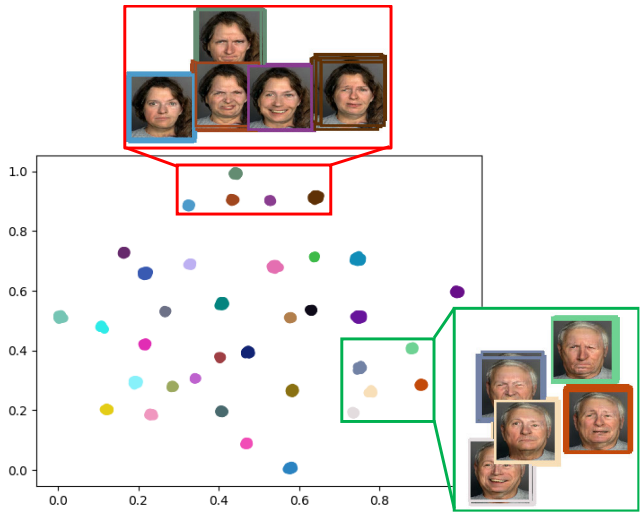


FIGURE 3. 2D scatter plot of the proposed conditionally transformed noises and their corresponding generated images in the FACES dataset. Each cluster belongs to a combination of diverse conditions and represented by a unique color.

loss function can be applied to train our proposal as well as generator to learn domain distribution characteristics.

Recent works [3], [14] started adopting loss functions that combines adversarial term like Hinge [3] and WGAN-GP [14] with conditional by considering label information via an inner product or changing the discriminator structure to have multiple linear output branches for each domain. As all of them consider conditional factor, our framework can be readily trained using these objectives. In the following section, we demonstrate that our proposed framework works well with diverse objective functions used in novel image synthesis and image-to-image translations.

IV. EXPERIMENTS

In this section, we present the performance of the proposed CT framework when used along with diverse existing GAN architectures for image synthesis and image-to-image translation tasks. For image synthesis, we test our proposed framework with DCGAN- [34] and ResNet-based [48] architectures. For image-to-image translations, we use the framework together with architectures of [12], [14], [30]. For these tests, we employ the following datasets that contain images with diverse labels: FACES [49], RaFD [50], Multi-PIE [51], CelebA [52] and its high quality variant CelebA-HQ [2], HWDB1.0 [53], and KITTI [54].

A. EVALUATION METRICS

We provide details on the evaluation metrics used in our experiments in this section. To evaluate the visual quality and diversity of the generated images quantitatively, we utilize the following metrics.

1) INCEPTION SCORE (IS) [55]

Using an Inception network [56] pre-trained on ImageNet, we calculate the statistics of generated images by considering the conditional label distribution $p(y|x)$ and marginal label distribution $p(y) = \int_x p(y|x)p(x)$ in the form of $\text{IS}(G) = \exp\left(\mathbb{E}_{x\sim G}\left[\text{KL}\left(p(y|x), p(y)\right)\right]\right)$.

2) FRECHÉT INCEPTION DISTANCE (FID) [57]

For the FID score, we extract specific layer features of the Inception network for real image x and generated image g and then (assuming that the features follow a multivariate Gaussian distribution) compute the distance as $\text{FID}(x, g) = \|\mu_x - \mu_g\|_2^2 + \text{TR}\left(\Sigma_x + \Sigma_g - 2\left(\Sigma_x\Sigma_g\right)^{\frac{1}{2}}\right)$, where μ and Σ are the empirical mean and covariance, respectively.

3) LEARNED PERCEPTUAL IMAGE PATCH SIMILARITY (LPIPS) [58]

To measure the diversity of generated images, we employ such metric [58]. For this purpose, we exploit ImageNet-pretrained AlexNet [59]. We perform diversity calculation by using L1 distance between extracted features of generated images.

B. STATISTICAL ANALYSIS OF THE FRAMEWORK

1) VISUALIZATION

Conditional generative models that produce condition-specific images should also ensure that these images solely fall within their own discriminative spaces. For example, given the condition "Female", the image should belong solely to this class. Our framework which comprises two steps for condition-specific transformation of inputs, maintains such a necessary aspect for the generation process. By determining the condition-specific space, the framework obtains affine parameters that best represent the given condition vector and thus transforms the input to correspond solely to this space. Fig. 3 presents the transformed noise space given all possible sets of multi conditions for the FACES dataset. We generate a 2D scatter plot by using the t-SNE approach [60]. To do so, we used our proposal to conditionally transform 3600 noise samples according to the 36 given unique classes (100 samples per class). Each of the classes represents a combination of several attributes (i.e., gender, age, expressions). We consider the combinations of the following attributes to make unique classes: two genders (male, female); three age-groups (young, middle, senior); six expressions (anger, disgust, fear, happiness, neutrality, surprise). All 3600 conditionally transformed feature-vectors are then mapped into low-dimensional representations via t-SNE. For this analysis, we plugged our proposal into a ResNet-based generator. Each cluster in 2D space represents one of the 36 unique classes. Notably, all transformed noises correspond to the distinctive conditional spaces and do not overlap with one another. Such an analysis confirms the solid and discriminative control maintained by our framework.

2) MoG SYNTHETIC DATA

We also analyze the performance of the proposed framework in different discriminator settings. To do so, we consider

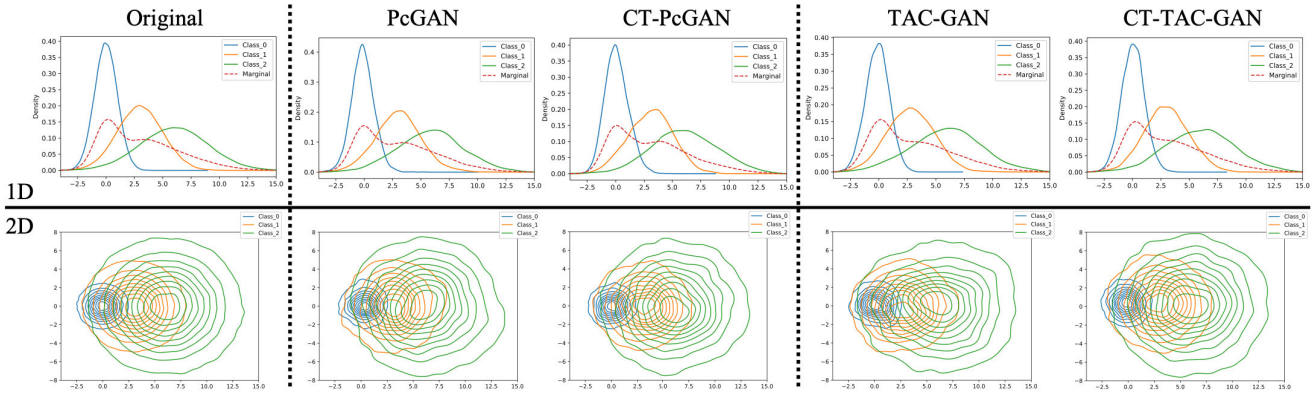


FIGURE 4. Comparison of sample quality on a 1D and 2D synthetic MOG dataset.

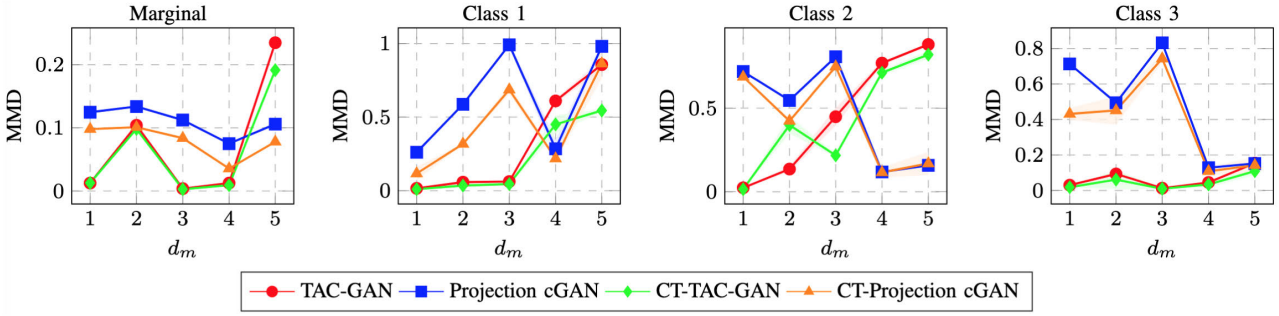


FIGURE 5. MMD distances evaluated on one-dimensional original and generated distributions. Here, d_m denotes distance between means of adjacent Gaussian components.

experiments using Mixture of Gaussian (MoG) synthetic data and discriminator approaches of PcGAN [3], and TAC-GAN [36]. Specifically, we compare the distribution matching ability of these discriminators using our proposal in the generator part. In case of [36], conditioning process has been achieved by concatenating the input with the output of embedding layers. Here, we replace such process by our CT-based approach. Similar to [36], we utilize samples drawn from a one-dimensional and two-dimensional MoG distribution with three Gaussian components. As experiment is condition-based, we ensure samples being labeled as one of the classes ranging in $[0 \sim 2]$. We fix the standard deviations of components to $\sigma_0 = 1$, $\sigma_1 = 2$, and $\sigma_2 = 3$.

In Fig. 4, we present original 1D and 2D Gaussian distributions, when $\mu_0 = 0$, $\mu_1 = 3$, and $\mu_2 = 6$, along with estimated ones produced by PcGAN [3], TAC-GAN [36] as well as CT-PcGAN and CT-TAC-GAN. We obtain these results by estimating the kernel density on the generated data distributions. According to the plots, both PcGAN and TAC-GAN can accurately learn the original distribution using our framework. Additionally, we also report the Maximum Mean Discrepancy (MMD) [61] in Fig. 5 which shows the distances between original distribution and the generated ones. For this evaluation, we follow [36] and train models using cross-entropy log loss. In both 1D and 2D evaluations, CT-based models achieve close to zero distances which means generated distributions are near to the original ones. We consider that such an analysis reveals the orthogonality of our proposal to different discriminator settings.

C. IMAGE SYNTHESIS

In this experiment, we demonstrate the effectiveness of the proposed CT framework against existing GAN architectures for image synthesis problems. For this, we conduct four different types of experiments, including single/multi-condition based image generation, multi-conditional image interpolation, and handwritten Chinese character generation.

1) SINGLE-CONDITION BASED IMAGE GENERATION

To demonstrate the capability of conditional transformation on controlling single-conditional image generation, we specifically experiment with this task as well. Within this testing, we also consider the possibility of using the proposed framework with different generator architectures. In the GAN literature, there are many types of architectures used for constructing generator networks. The most commonly used architectures are DCGAN-based [34] and ResNet-based ones [48]. According to this, we prepended our proposal on top of these generators and perform single-conditioned image synthesis. For this purpose, we used the Radboud Faces Database (RaFD) [50] providing facial images with eight expressions including angry, contemptuous, disgusted, fearful, happy, sad, neutral, and surprised. As for comparison, we consider generators that use condition with concatenation

TABLE 1. Quantitative comparison for single-/multi-condition based experiments on RaFD, FACES and Multi-PIE datasets.

Methods	*Single-Condition*		*Multi-Condition*			
	RaFD		FACES		Multi-PIE	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS ↑
Real Data	0.0	1.34	0.0	1.67	0.0	1.65
Concat	116.4	1.21	89.7	1.44	131.4	1.23
CBN	104.1	1.29	80.2	1.47	93.1	1.25
VCGAN	101.3	1.30	78.4	1.49	91.6	1.26
Ours$_{DC}$	110.6	1.17	84.9	1.45	98.3	1.24
Ours$_{Res}$	**86.8**	**1.32**	**69.1**	**1.53**	**76.2**	**1.30**

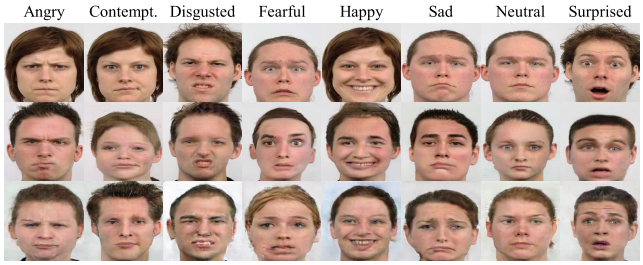


FIGURE 6. Single-conditioned image synthesis based on training of RaFD images. Top-to-bottom images are from RaFD dataset, ResNet and DCGAN-based generators respectively.

(Concat) [21], conditional batch normalization (CBN) [3], and VCGAN [44] based approaches.

We present quantitative comparison results in Table 1. As it can observed the proposed framework with ResNet-based generator (Ours$_{Res}$) provides better performance compared to all other methods under the consideration. Moreover, in terms of IS score, we could achieve closer result to the real data depicting the diversity of generated images. The generated images given in Fig. 6 also affirms the better quality of ResNet-based generator. Although DCGAN-based generator (Ours$_{DC}$) have lower performance than ResNet-based one, the proposed framework still achieves its goal on conditionally transforming the input noise, and exhibits its efficient usage with diverse generator architectures. Overall, such performance demonstrates that the proposed framework can apply its learning mechanism on a single condition and provide condition-specific information for transforming the noises and thereby, control the single-conditioned image generation. This comparison indicates that proposed CT framework can effectively operate regardless the generator architecture, however, the results are even better with more advanced ones like ResNet-based generator.

2) MULTI-CONDITION BASED IMAGE GENERATION

To evaluate a conditional transformation GAN on multi-conditional inputs, we considered two datasets: FACES [49] and Multi-PIE [51]. The FACES provides naturalistic facial images of young, middle-aged, and older females and males portraying six expressions such as anger, disgust, fear, happiness, neutrality, and surprise. Multi-PIE is another multi-labeled dataset consisting of facial images with variations in pose, illumination, along with neutral and smile expressions.

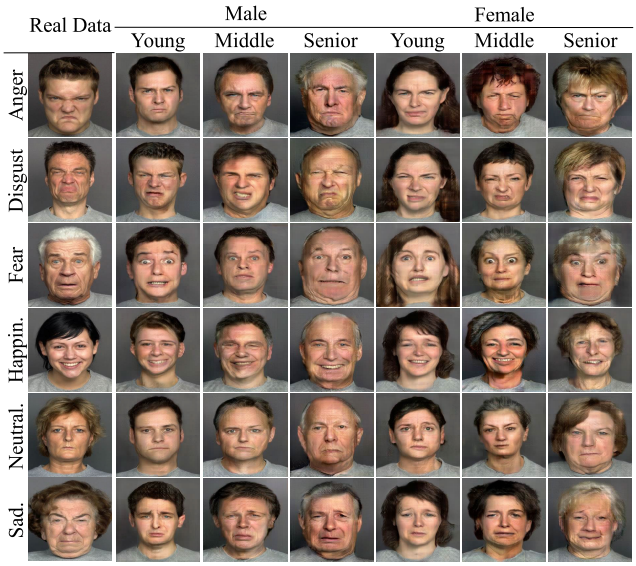


FIGURE 7. Multi-attribute facial images generated by the proposed method given conditions such as gender, age-group, and expression.

FIGURE 8. Results generated by using proposed framework given conditions such as pose, illumination, and expression. Odd rows: exemplar images from Multi-PIE dataset for neutral and smile expressions; Even rows: generated samples with diverse pose and illumination, as well as neutral and smile expressions.

We consider the FACES dataset to train generator along with the proposed framework so that it controls the synthesis of facial images having soft biometrics. Through learning complex relations between given multi conditions such as gender, age, and expression our proposal enables the transformation of noise accordingly which yields to generate images as shown in Fig. 7. Looking at the generated images, we could observe that for each combination of diverse gender, age, and expression conditions, the generator synthesizes images displaying the proper facial attributive details. It is also noteworthy to mention that by transforming the noise according to the given conditions, we achieve control over the content and have variations within content-specific images. A quantitative comparison provided in Table 1 also shows the better performance of generator with proposed framework in both scores under the consideration.

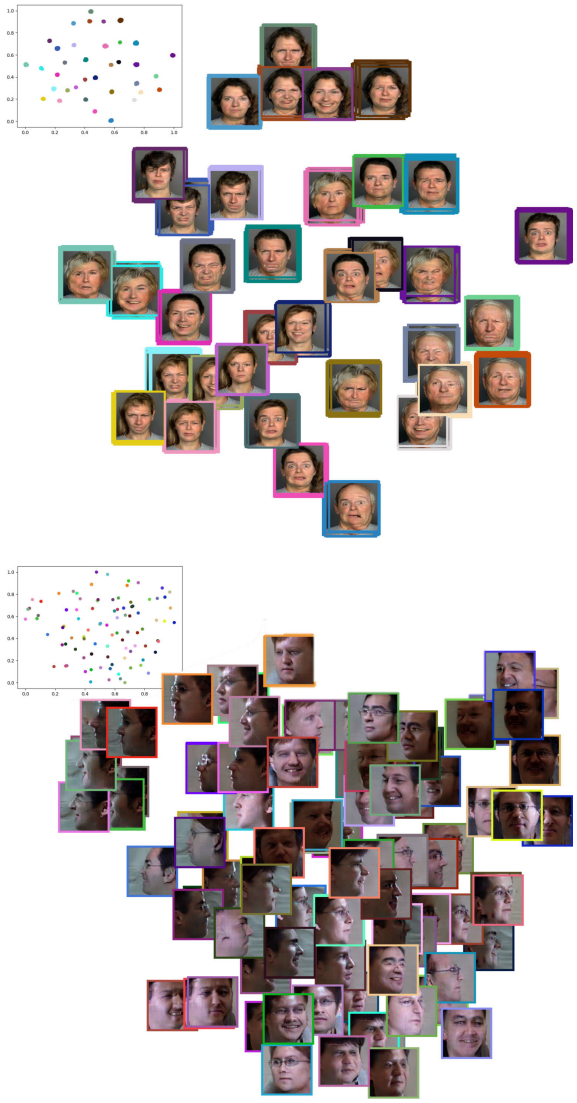


FIGURE 9. 2-D scatter plot of conditionally transformed noises and their corresponding generated images on FACES (top) and Multi-PIE (bottom) datasets. Each cluster belongs to the combination of diverse conditions and represented by a unique color.

Apart from generating faces with soft biometrics, we also considered facial image generation having different poses and illuminations. As the Multi-PIE dataset[1] additionally provides two expression labels, the generation process is more challenging since generated images should depict not only illumination variations but also face in particular pose having either neutral or smile expressions. In Fig. 8, we present the generated images along with real ones both having the same multi-conditional information. We demonstrate images in this way to point out that our framework providing condition-specific features for noise transformation guides the generator to produce images having the same properties as real ones. Besides, quantitative comparison in Table 1 also quantifies its effectiveness in terms of quality and diversity measures.

Besides, we demonstrate the transformed noise space given all possible set of multi conditions and their generated images in Fig. 9 for FACES and Multi-PIE datasets. As can be seen from the scatter plot all transformed noises and their images correspond to the distinctive conditional spaces, and thus showing the effect of conditionally transforming of noise inputs.

3) MULTI-CONDITIONAL IMAGE INTERPOLATION

In this part, we verify whether the learned conditional feature-space encompassing a combination of various conditions can be used to generate smoothly interpolated images. As conditional vectors c representing multiple conditions are sampled as an one-hot distribution, to demonstrate interpolation across different conditional spaces, we sampled two c_1 and c_2 and interpolated these one-hot vectors to obtain newly transformed noise inputs. Although the proposed framework has not seen such conditional input during the training, we observed that such information has not affected the smooth transition between conditions. By observing Fig. 10A, we could notice that there is a fine transition even though c_1 and c_2 are sampled to have no same conditions, which verifies the possibility having smooth shifting from one conditional space to another. When we manipulate only one condition out of three in contrast to the previous interpolation (see Fig. 10B), we see that how one face having particular gender and expression ages in a smooth way. Such a transition can be clearly seen in neck introducing aging wrinkles as well as how young hair-style moves to conventional one. Altogether, these presented images demonstrate that how noises transformed on one specific set of conditions smoothly transit to another set of conditions.

4) HANDWRITTEN CHINESE CHARACTERS

LSGAN [33] raises a discussion on the infeasibility of directly conditioning the generator input with a one-hot vector representing thousands of classes in terms of memory and computational cost. Such a scenario is particularly more challenging in case of dealing with the generation of handwritten Chinese characters. To control the generation with such label vector, LSGAN uses a simple linear mapping to reduce the high-dimensionality of conditional information for further concatenation with a noise while still having high-dimensional input. We consider such a challenging case is to be appropriate for demonstrating the applicability of our proposed framework. Thus, we train LSGAN by replacing their linear mapping layer to our proposal (i.e., prepending to the top of LSGAN) on a handwritten Chinese character dataset (HWDB1.0) [53] consisting of 3740 classes. In particular, for experimenting with LSGAN on synthesizing handwritten Chinese characters, we adopt an implementation of LSGAN. We randomly sample labels representing Chinese characters from [0, 3740] for obtaining their affine parameters as described in Section III that transform the input noise

[1]We used a cropped version of this dataset in our experiments: https://github.com/bluer555/CR-GAN.

FIGURE 10. Interpolation between conditionally transformed noises and their generated images.

FIGURE 11. Handwritten Chinese characters generated by LSGAN with our conditional transformation framework.

sampled from uniform distribution $\mathcal{U}[-1, 1]$ before feeding to the generator. Note that LSGAN has been constructed based on implementation of DCGAN [34], and considers least-squares loss function for adversarial training. We set the hyper-parameters in the same way as it was done in the original work. Specifically, the learning rate is set to 2×10^{-4} and $\beta_1 = 0.5$.

We provide several synthesized characters in Fig. 11 to demonstrate the efficacy of using our conditional transformation framework on this task. As can be observed, the generator can synthesize readable Chinese characters corresponding to the given conditions. This result also exhibits that we can efficiently utilize our proposal for transforming the noise to be aligned with a particular class of high-dimensional label vector. Similar to the observation of [33], we think that our approach to conditioning the generator can be readily used for data augmentation or construction needs.

D. IMAGE-TO-IMAGE TRANSLATION

In this section, we demonstrate the applicability of the proposed framework within GAN architectures for image-to-image translation. For this, we conduct two different types of experiments: facial domain transfer and multi-view to novel view synthesis.

1) FACIAL DOMAIN TRANSFER

As discussed, conditioning the input image with class labels guides the generator in effectively translating an image from one domain to another. One of the representative works is StarGAN [12], which applies a strategy for conditioning the generator by concatenating the labels to the input as additional image channels. In fact, such a simplistic approach has shown its efficiency in translating facial expressions as well as appearance characteristics. Since conditioning is performed at the initial stage, we can utilize StarGAN image generation with our CT framework (CT-StarGAN). To verify such applicability, we also replace the concatenation layer with our framework as performed with LSGAN. However, unlike in LSGAN, we use WGAN-GP loss [62] for adversarial training.

The implementation of our model (CT-StarGAN) is based on a publicly available implementation of StarGAN. The difference between original StarGAN and our CT-StarGAN is in the incorporation of labels into image translation process. CT-StarGAN uses proposed CT framework to transform input image according to the given conditions rather than concatenating labels as additional channels to the image, as performed in [12]. Note that the network architecture of StarGAN is adopted from [46]. We do not perform any other changes on architecture construction aside from the aforementioned label incorporation.

We maintain the hyper-parameters settings in the original work of StarGAN. The models are trained with the Adam optimizer [63], and momentum parameters are set as follows: $\beta_1 = 0.5$ and $\beta_2 = 0.999$. A linear decay is applied on the learning rate of 0.0001 in the same manner as that in [12] for RaFD [50] and CelebA [52] datasets.

We train CT-StarGAN and its original implementation in translating facial expressions by using the same RaFD dataset [50]. As mentioned above, RaFD contains eight different expression images; for comparison, we set the input domain as the neutral expression and vary the target domains among the other seven. Choi *et al.* [12] split the dataset into training and test sets at a ratio of 90% − 10%. We also split dataset images while ensuring that there is no person overlapping occurs between sets to avoid bias in translation model. This splitting allows us to perform reliable qualitative and quantitative evaluations.

To evaluate the performance of translated images in RaFD quantitatively, we perform expression classification in the same way as original work. To this purpose, we used ResNet-18 architecture [48] following the work of [12]. Network training is conducted using the training set (90%) that has no overlapping with the test set (10%) to carry

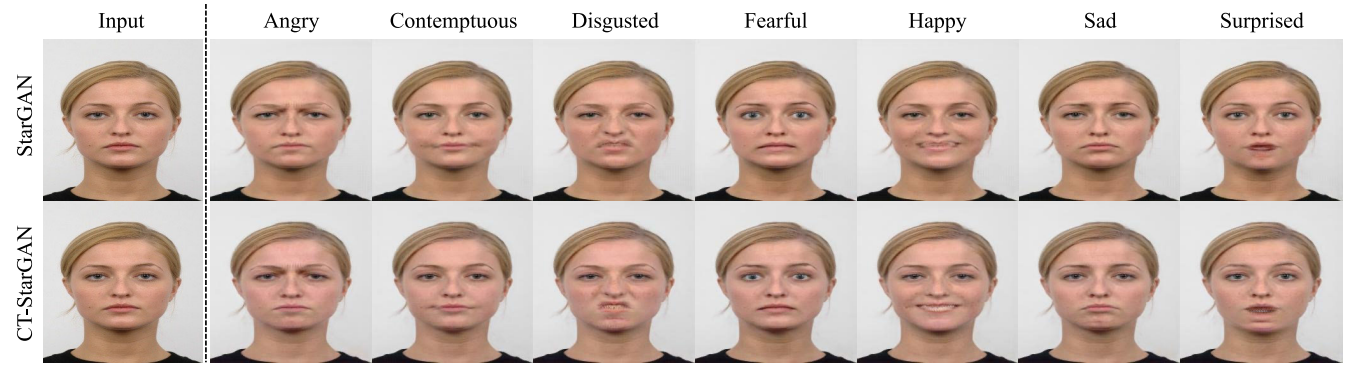


FIGURE 12. Facial expression translation results on RaFD dataset. For a given input face image with neutral expression, both translation models produced results for seven different expressions.

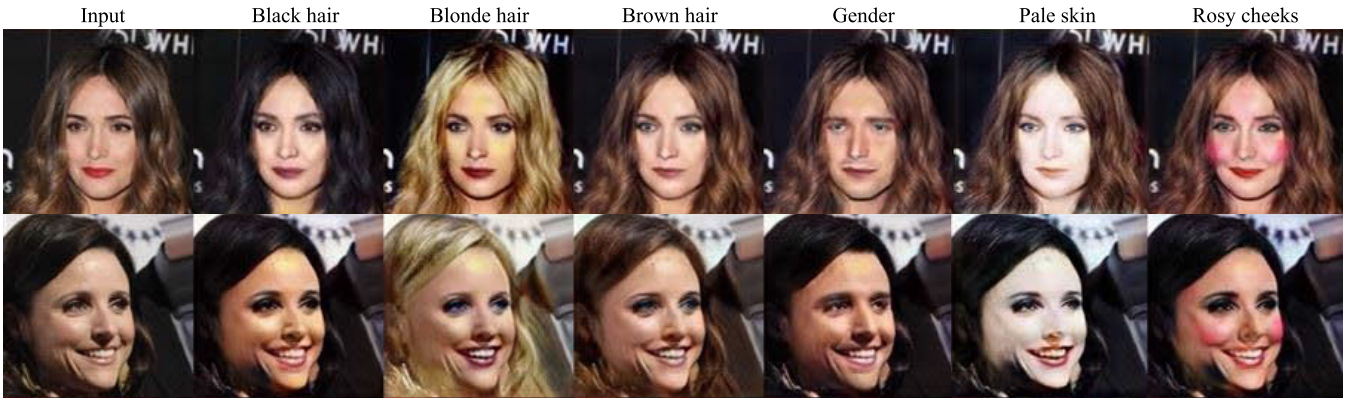


FIGURE 13. Facial attribute translation results on CelebA dataset by using our proposed method.

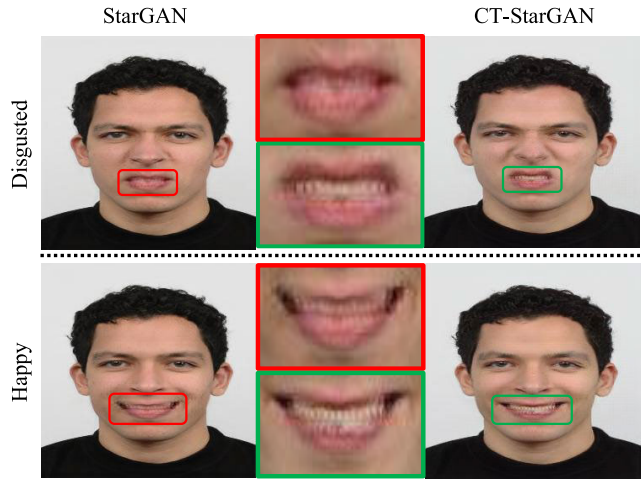


FIGURE 14. Qualitative comparison of `neutral→disgusted` and `neutral→happy` translations done by StarGAN and CT-StarGAN.

TABLE 2. Classification accuracy (%) and SER-FIQ scores on real and translated images of RaFD dataset.

Image set	StarGAN	CT-StarGAN	Real images
Accuracy	91.27	**95.24**	97.96
Score	0.709	**0.718**	0.805

TABLE 3. Quantitative comparison on latent- and reference-guided synthesis for CelebA-HQ dataset.

Methods	Latent		Reference	
	FID	LPIPS	FID	LPIPS
MUNIT [64]	31.4	0.363	107.1	0.176
DRIT [65]	52.1	0.178	53.3	0.311
MSGAN [66]	33.1	0.389	39.6	0.312
StarGAN v2 [14]	13.8	0.453	23.8	0.388
CT-StarGAN v2	**13.1**	**0.458**	**23.0**	**0.391**
Real images	14.8	-	12.9	-

out person-independent classification. We utilize the original test set for an expression classification of real (untranslated) images. After translating the images of test set by using StarGAN and CT-StarGAN models, we perform classification on generated (translated) images and report the results.

The qualitative results presented in Fig. 12 demonstrate that replacing concatenation with our framework does not diminish the image quality. Moreover, we observe an improved mapping quality of one expression to another using our proposal. For instance, Fig. 14 shows two translated images for disgusted and happy expressions done by both methods. It is noticeable that CT-StarGAN can add details on the mouth region of the face. Such performance by StarGAN was also discussed in [67], which pointed out that the method can accept only one domain as input, and when such

FIGURE 15. StarGAN v2-based translations using proposed framework for latent-guided synthesis. The top row presents input images, whereas the rest of the rows are generated.

FIGURE 16. StarGAN v2-based translations using proposed framework for reference-guided synthesis. Top row: reference images, left column: source images; rest of the images are generated.

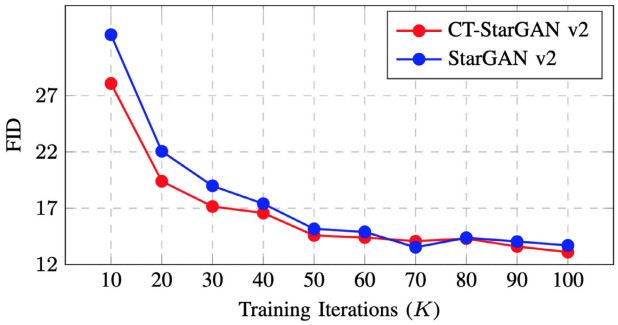


FIGURE 17. FID values at the different iteration steps on training latent-guided synthesis for CelebA-HQ dataset.

source does not contain information (e.g., teeth) on neutral expression, it cannot properly translate to the aforementioned expression. We think due to such translation, StarGAN cannot achieve better performance in terms of expression classification shown in Table 2. Note that for expression classification, we follow the strategy done by [12]. The analysis reveals that using our proposed framework can improve the shortcomings of StarGAN. We also consider SER-FIQ [68] as additional estimations on generated images. SER-FIQ [68] is a face quality assessment method that can be used for predicting the suitability of face images for face recognition. Table 2 includes SER-FIQ scores for real and generated images by StarGAN and CT-StarGAN. It is notable that even in the face image suitability test by SER-FIQ, images generated by the proposed framework obtained a higher score.

Considering that CT-StarGAN has demonstrated its applicability for translating facial expressions in the lab-controlled environments, we take into account its application for translating appearance-based attributes in-the-wild circumstances. To this end, we train CT-StarGAN on the CelebA dataset [52] by following the settings of [12]. In contrast to the previous test, this one aims to translate the input to the domains of the following appearance attributes: hair color (including black, blond, brown), gender, pale skin, and rosy cheeks. Fig. 13 presents the results of CT-StarGAN for this test. We observe that the model can generate plausible quality images with appropriate target appearance attributes. Such translation results exhibit the potential of proposed framework in learning the feature-space and its suitability for manipulating even facial appearances of in-the wild face images.

In addition, we apply our framework within StarGAN v2 [14] that is a recent variant of StarGAN [12]. The work of [14] focuses on scalability and diversity of generated images. For this purpose, they introduce a mapping network to incorporate style information of domain. In our experiment, we replace this network by our framework. As StarGAN, we maintain all parameter settings same as in [14]. The qualitative results are shown in Fig. 15 and 16. As can be observed, our proposed framework is able to generate sharp images with consistent domain and its style information. In quantitative manner, our results can

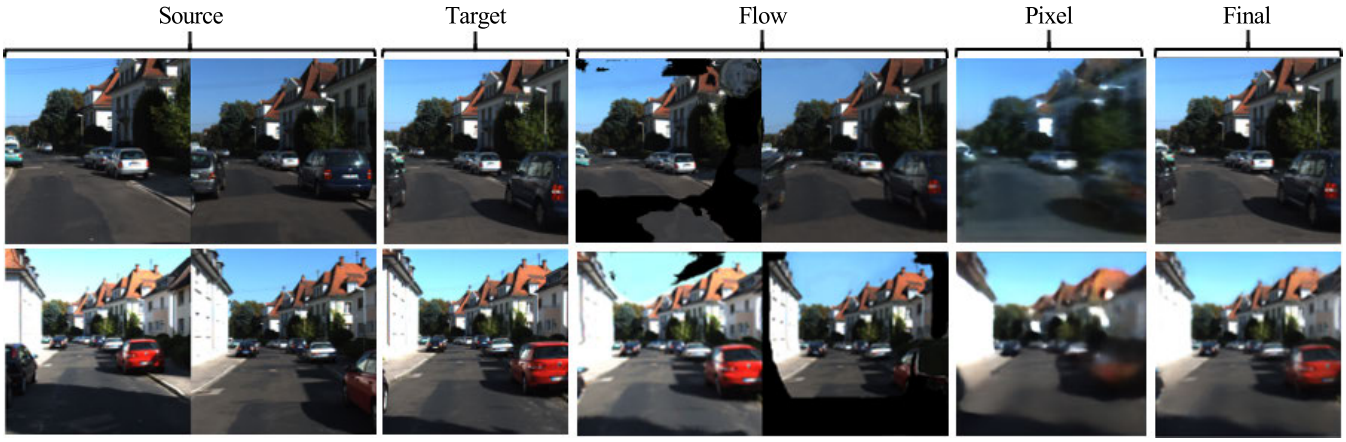


FIGURE 18. Synthesized scenes on KITTI dataset [54] using our framework within model of [30].

be explored in Fig. 17 and Table 3. As shown in Fig. 17, our results provides better score in terms of FID over the iterations, whereas the final scores presented in Table 3 using FID and LPIPS metrics demonstrates an improvement of the StarGAN v2 model. Such results also advocate applicability of proposed framework for this task, and facilitates to achieve better scoring values.

2) MULTI-VIEW TO NOVEL VIEW SYNTHESIS

The work of [30] have shown the possibility of generating novel views based on multi-view images with target camera poses using pixel and flow generators. Similar to StarGAN [12], this work also concatenates pose information to the source images and feeds to the encoder. Following our experiments using StarGAN, we apply our framework for novel view synthesis along with the model of [30]. Here, we aim to generate a target image by considering a target camera pose and N (image, camera-pose) pairs. To achieve such an image generation, we perform an experiment using well-known KITTI dataset [54] used for simultaneous localization and mapping (SLAM) evaluation. Similar to [30], we use a $6DoF$ vector as a continuous camera pose representation, and maintain settings for constructing the training and test splits.

We present a few examples of novel views generated by using our framework within the model of [30] in Fig. 18. Given the conditional information in any form, our framework can produce features transforming the images to be consistent with associated context (camera pose) information. Notably, our addition has no impact on the quality of images, and thus, results maintain structural consistency and realism.

V. CONCLUSION

In this paper, we proposed a novel architectural addition for a generator network, namely, conditional transformation (CT) framework. The technical contribution of our work consists of controlling the image generation process by conditionally transforming the input (e.g., noise or image) to the generator such that it corresponds to the given conditions. With this framework, conditional modulations are not required in the intermediate layers of the generator as the framework can be prepended to the top of any GAN generator and yield the desired output. We tested the applicability of our proposal along with diverse generator networks in image synthesis and image-to-image translation tasks. In both tasks, we demonstrated that the proposed framework can be used for conditional transformation of noises and images and guide the generator in producing condition-specific images. We conclude that through such transformation, features on conditional vectors can be learned, allowing us to explicitly control complex conditional image generation. In turn, such control can be readily helpful for diverse GAN architectures in constructing image datasets necessary for many applications.

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.*, 2014, pp. 2672–2680.
[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.
[3] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–23.
[4] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–35.
[5] 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.
[6] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2726–2737, Nov. 2019.
[7] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.
[8] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 284–293.
[9] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4570–4580.
[10] J. Cai, Z. Meng, and C. M. Ho, "Residual channel attention generative adversarial network for image super-resolution and noise reduction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 454–455.

[11] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7769–7778.

[12] 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.

[13] X. Ning, D. Gou, X. Dong, W. Tian, L. Yu, and C. Wang, "Conditional generative adversarial networks based on the principle of homologycontinuity for face aging," *Concurrency Comput., Pract. Exper.*, p. e5792, Apr. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.5792

[14] 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.

[15] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman, "Lifespan age transformation synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 739–755.

[16] Q. Wang, H. Fan, G. Sun, W. Ren, and Y. Tang, "Recurrent generative adversarial network for face completion," *IEEE Trans. Multimedia*, vol. 23, pp. 429–442, 2020.

[17] M. K. V. Dumoulin and J. Shlens, "A learned representation for artistic style," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–26.

[18] H. Wang, Y. Li, Y. Wang, H. Hu, and M.-H. Yang, "Collaborative distillation for ultra-resolution universal style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1860–1869.

[19] S. Qiu, Y. Zhao, J. Jiao, Y. Wei, and S. Wei, "Referring image segmentation by generative adversarial learning," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1333–1344, May 2019.

[20] M. Tschannen, E. Agustsson, and M. Lucic, "Deep generative models for distribution-preserving lossy compression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5929–5940.

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

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

[23] A. Van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with PixelCNN decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4790–4798.

[24] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional GANs for image editing," 2016, *arXiv:1611.06355*. [Online]. Available: http://arxiv.org/abs/1611.06355

[25] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cycleGAN," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 282–297.

[26] T. Chen, M. Lucic, N. Houlsby, and S. Gelly, "On self modulation for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–18.

[27] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[28] 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.

[29] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5400–5409.

[30] S.-H. Sun, M. Huh, Y.-H. Liao, N. Zhang, and J. J. Lim, "Multi-view to novel view: Synthesizing novel views with self-learned confidence," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 155–171.

[31] X. Ning, F. Nan, S. Xu, L. Yu, and L. Zhang, "Multi-view frontal face image generation: A survey," *Concurrency Comput., Pract. Exper.*, p. e6147, Dec. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.6147

[32] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[33] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[34] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[35] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.

[36] M. Gong, Y. Xu, C. Li, K. Zhang, and K. Batmanghelich, "Twin auxiliary classifiers GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1330–1339.

[37] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6594–6604.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[39] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[40] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.

[41] J. Despois, F. Flament, and M. Perrot, "AgingmapGAN (AMGAN): High-resolution controllable face aging with spatially-aware conditional GANs," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2020, pp. 613–628.

[42] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[43] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2018.

[44] M. Hu, D. Zhou, and Y. He, "Variational conditional GAN for fine-grained controllable image generation," in *Proc. Asian Conf. Mach. Learn.*, 2019, pp. 109–124.

[45] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–14.

[46] 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.

[47] X. Ning, S. Xu, W. Li, and S. Nie, "FEGAN: Flexible and efficient face editing with pre-trained generator," *IEEE Access*, vol. 8, pp. 65340–65350, 2020.

[48] 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.

[49] N. C. Ebner, M. Riediger, and U. Lindenberger, "FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behav. Res. Methods*, vol. 42, no. 1, pp. 351–362, Feb. 2010.

[50] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognit. Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[51] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[52] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[53] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "CASIA online and offline Chinese handwriting databases," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 37–41.

[54] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

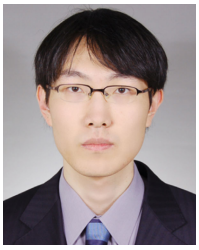
[55] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[57] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.
[60] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[61] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.
[62] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.
[63] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114
[64] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.
[65] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.
[66] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, and M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1429–1437.
[67] L. Shen, W. Zhu, X. Wang, L. Xing, J. Pauly, B. Turkbey, S. Harmon, P. L. Choyke, and B. J. Wood, "Representational disentanglement for multi-domain image completion," Tech. Rep., 2019. [Online]. Available: https://openreview.net/forum?id=rkg_wREYDS
[68] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5651–5660.

FARKHOD MAKHMUDKHUJAEV (Member, IEEE) received the B.S. and M.S. degrees from the Tashkent University of Information Technologies, Uzbekistan, in 2012 and 2014, respectively, and the Ph.D. degree in computer science and engineering from Kyung Hee University, South Korea, in 2019. He is currently conducting his postdoctoral research at the Artificial Intelligence Convergence Research Center, Inha University. His current research interests include image synthesis using generative adversarial networks and facial attribute analysis and recognition.



JUNSEOK KWON (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, South Korea, in 2006, 2008, and 2013, respectively, under the supervision of Prof. Kyoung Mu Lee. He was a Postdoctoral Researcher with the Computer Vision Laboratory, ETH Zurich, from 2013 to 2014, supervised by Prof. Luc Van Gool. He is currently an Associate Professor with the School of Computer Science and Engineering, Chung-Ang University, South Korea. He is working in the field of object tracking to capture the dynamics of cities. His research interests include visual tracking, visual surveillance, and Monte Carlo sampling method and its variants.

IN KYU PARK (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University (SNU), in 1995, 1997, and 2001, respectively. From September 2001 to March 2004, he was a member of Technical Staff at the Samsung Advanced Institute of Technology (SAIT). Since March 2004, he has been with the Department of Information and Communication Engineering, Inha University, where he is currently a Full Professor. From January 2007 to February 2008, he was an Exchange Scholar at Mitsubishi Electric Research Laboratories (MERL). From September 2014 to August 2015, he was a Visiting Associate Professor at the MIT Media Lab. From July 2018 to June 2019, he was a Visiting Scholar at the Center for Visual Computing, University of California, San Diego. His research interests include the joint area of computer vision and graphics, including 3-D shape reconstruction from multiple views, image-based rendering, computational photography, deep learning, and GPGPU for image processing and computer vision. He is a member of ACM.