Abstract

In this paper, we revisit the Image-to-Image (I2I) translation problem with transition consistency, namely the consistency defined on the conditional data mapping between each data pairs. Explicitly parameterizing each data mappings with a transition variable $t$, i.e., $x \xrightarrow[]{t} y$, we discover that existing I2I translation models mainly focus on maintaining consistency on results, e.g., image reconstruction or attribute prediction, named result consistency in our paper. This restricts their generalization ability to generate satisfactory results with unseen transitions in the test phase.

Consequently, we propose to enforce both result consistency and transition consistency for I2I translation, to benefit the problem with a closer consistency between the input and output. To benefit the generalization ability of the translation model, we propose transition encoding to facilitate explicit regularization of these two kinds of consistencies on unseen transitions. We further generalize such explicitly regularized consistencies to distribution-level, thus facilitating a generalized overall consistency constraints for I2I translation problems. With the above design, our proposed model, named Transition Encoding GAN (TEGAN), can possess superb generalization ability to generate realistic and semantically consistent translation results with unseen transitions in the test phase. It also provides a unified understanding on the existing GAN-based I2I transition models with our explicitly modeling of the data mapping, i.e., transition. Experiments on four different I2I translation tasks demonstrate the efficacy and generality of TEGAN.

1. Introduction

Image-to-Image (I2I) translation \cite{15} targets to map an image from a source domain $X$ to a target domain $Y$, with these two domains differentiated by different sets of attributes. Under this definition, various computer vision problems can be formulated as I2I translation problems, such as face editing \cite{17}, style transfer \cite{7}, image inpainting \cite{21, 22}. Recently, following the development of generative adversarial networks (GANs) \cite{9, 20}, I2I translation problems have received significant attention.

In previous I2I translation works, conditional GANs \cite{9} are widely utilized to make translated images with specific attributes as expected. For example, in face editing tasks, RelGAN \cite{19} utilizes a conditional GAN-based framework to translate an image with an original expression (attributes) to a new image with a different expression. In other words, a successful trained I2I network must make the attributes of translated results consistent with attributes expected. To achieve this goal, previous works either adopt corresponding attribute classifiers to control the attributes of translated results \cite{10, 3, 28, 1, 19, 8, 27}, or design specific loss function, e.g., cycle-GAN \cite{21} to control the consistency between translated results and target domain $Y$. Those previous works focus on keeping consistency on translated results, which we shortly call result consistency.

When looking closer at the I2I translation works focusing on maintaining result consistencies \cite{10, 3, 28, 1, 19, 8, 27}, we argue that other than the result consistency in I2I translation, there is another consistency existing while ignored in previous works, which is named as transition consistency. The definition of transition consistency is as follows: if a translated output $\hat{y}$ and $y$ sharing the same attributes (i.e. result consistency) are close, they should also share a common conditional transformation. To formulate the transition consistency existing in I2I translation, we explicitly parameterize data transformation between a source data $x$ and a target data $y$ by introducing a transition variable $t$, namely $t \triangleq t(x, y)$. In this way, an I2I translation problem is reformulated to maintain two consistencies in learning processes, i.e., result consistency and transition consistency.
consistency. Considering both result consistency and transition consistency to solve I2I translation problems can benefit us from at least two aspects:

First, comparing to previous works only modeling the result consistency [10, 3, 28, 1, 19, 8, 27], our method models two consistencies simultaneously, i.e., result consistency and transition consistency. Explicitly modeling the transition between $x$ and $y$ and conducting learning under the transition consistency make the trained network model the inherent structure among closely related I2I translation tasks, which is proved in comparison to previous representative works in Fig. 1(a). Since our method explicitly parameterizes the transition process and controls the consistency on both transition and results, our method outperforms previous I2I translation works focusing only on result consistency, which has been proved in our experiments.

Second, to generalize our targeted result consistency and transition consistency to unseen transitions, we propose to explicitly model the unseen transitions $\hat{t}$. In this way, the transition consistency defined on the observed transitions can be naturally extended to unseen transitions and its generated results. An example is presented in Fig. 1(b). To overcome the limitation of simply regularizing results consistency on observed transitions in existing methods, we explicitly model unseen transitions $\hat{t}$ with an transition encoding module [16], which can flexibly manipulate $\hat{t}$ to create reasonable new transitions for generation. We define loss terms to enforce consistency between these explicitly mimicked triplet data. We then further generalized the overall consistency defined on the triplet data to the distribution-level via joint distribution matching [6]. We name our method Transition Encoding GAN (TEGAN).

Our contributions are summarized as follows:

- We propose TEGAN to model the result consistency and transition consistency simultaneously. To the best of our knowledge, this is the first time to explicitly model transition consistency in I2I translation problems. Comparing to previous works focusing on result consistency only, TEGAN can work in higher granularity, model dynamic relation between attributes, and generate more reasonable results with better qualities.

- TEGAN presents a general framework for I2I translation. As discussed in Sec. 4, existing GAN-based I2I translation works can be connected and interpreted as special cases of our TEGAN.

- We conduct extensive experiments on various I2I translation problems, including face editing, image inpainting, style transfer. The experimental results demonstrate the efficacy and generality of TEGAN.

2. Related Works

GANs based I2I translation. I2I translation aims to transform an input image from one domain X to another domain Y with different attribute(s) [11]. Conditional Generative Adversarial Networks (cGANs) have been widely adopted to tackle I2I translation tasks and have achieved impressive results. Specifically, some methods simply leverage example images for training, and formulate the problem as $G(x) = \hat{y} \approx y$ where $\hat{y}$ is the generated output and $\hat{y} \sim p(y|x)$. For example, Pix2Pix [15] utilized paired data, i.e., $\{x, y\}$ for training; CycleGAN [27] taking advantages of cycle-consistency to train with unpaired data, i.e., $x \in X$ and $y \in Y$. Other works adopt attribute annotations of the target domain, denoted as $z_y$, to make the generated results present desired property, with the problem formulated as $G(x, z_y) = \hat{y} \approx y$, where $\hat{y} \sim p(y|x, z_y)$ [10, 8, 28, 1, 19, 8].

These previous methods focus on maintaining consistency defined on results, namely result consistency, for I2I translation. Such a design would restrict their translation capacity in two aspects. First, these models count on labeled data for training, either labeled with pair-wise images [18] or extra attribute annotations of each images [4]. In this way, they can be largely restricted by the data collection difficulty [5]. Second, since the result consistency can be simply defined on the training examples, such consistency can not be generalized to unseen transformation, e.g., $\hat{x} \to \hat{y}$, to achieve reasonable results in the test phase. We present further discussions on these previous methods in Sec. 4.

3. Method

In this section, we firstly reformulate I2I translation problems with transition consistency, then we illustrate how to generalize TEGAN to unseen transitions; finally, we explain how to train TEGAN by designed loss terms.
3.1. Looking back at I2I formulation with transition consistency

Let $x \in X$ and $y \in Y$ be the images of source domain $X$ and target domain $Y$, with the mapping between each data pair characterized with a transition variable $t \triangleq t(x, y) \in T$, i.e., $x \mapsto y$. After introducing the transition variable $t$, I2I translation problems aim to learn a mapping $G(x, y) = G(x)\mapsto y$, which is depicted in Fig. 3.

After reformulating I2I translation problems, the two consistencies in I2I translation should be formulated as: 1) result consistency: consistency defined on the output images, i.e., $G(x, t) = \hat{y} \approx y$; 2) transition consistency: the necessary condition that $y$ shares a common transition $t$ as the target image $\hat{y} = G(x)$, i.e.,:

$$
(x, y) \mapsto \hat{t}_y, \quad (x, \hat{y}) \mapsto \hat{t}_\hat{y}, \quad \hat{t}_y \approx \hat{t}_\hat{y} = t, \quad (1)
$$

where $\hat{t}_y$ and $\hat{t}_\hat{y}$ denote the transition of $\{x, y\}$ and $\{x, \hat{y}\}$, respectively. Eq. (1) explicitly formulates the transition consistency: given a source data $x$, if its translated result $\hat{y}$ is close to $y$, then the transition from $x$ to $\hat{y}$, denoted as $\hat{t}_y$, should be close to the transition from $x$ to $y$, denoted as $t_y$.

3.2. Generalize transition consistency to unseen data

Eq. (1) regularizes transition consistency for the observed transitions, i.e., $x \mapsto y$ in the training data. The generality for unseen transitions, denoted as $t$, depends on the quantity and diversity of training data.

To generalize the transition consistency to unseen transitions, we propose to explicitly model unseen transitions $\hat{t}$ with transition encoding $\{10\}$, i.e., $x \times y \mapsto t$. Specifically, we learn a distribution of $t$ with a $q$ function. The learned distribution characterizes a latent manifold $T$, where the observed transitions lied on. Random samples on this manifold are reasonable transitions $\hat{t}$, that lead to new generation outputs, i.e., $\hat{y}_i = G(x, \hat{t}_i)$.

4. TEGAN: translate with transition encoding

To generalize the transition consistency to unseen transitions, denoted as $\hat{t}$, we further generalize the transition consistency defined in observed transitions to unseen transitions with the designed operations in Sec. 3.2. Our TEGAN is a general generative I2I translation framework that can cover the existing models and also pose better generalization ability by regularize transition consistency for unseen transitions. Right: Our TEGAN is general to be applied to various I2I translation tasks with either paired or unpaired data, e.g. (a) supervised outdoor scene editing, (b) unsupervised style transfer.
3.3. Consistency between unseen transition and its corresponding translation result

Although the consistency between sampled \( i \) and its translation result can be explicitly regularized through Eq. (3), such consistency can not be generalized to the transitions not sampled in the training phase. It is also infeasible to sample all transitions for explicit regularization. Consequently, we further enforce a joint distribution matching on the triplet data, to generalize the consistency between \( x, \tilde{t} \) and \( \tilde{y} \) to distribution-level [26].

\[
p_\phi(x, \tilde{t}, \tilde{y}) \approx p_\phi(x, t', y') \approx p_\phi(x, \tilde{t}, \tilde{y}) \approx p_\phi(x, t, y) \tag{4}
\]

With Eq. (4), the transition consistency constraints defined in Eq. (3) are further generalized to transitions that have not been explicitly generated (still unseen transitions) during training. This thus further benefits the model’s generalization ability, facilitating it with even better translation performance in the test phase. We present Fig. [3] to illustrates our overall idea to simultaneously tackle these issues.

3.4. Our Model: Transition Encoding GAN

Based on the above design, we present Transition Encoding GAN (TEGAN), which instantiate our idea with a stochastic encoder, i.e., \( q(t|x, y) \equiv E(x, y) \). Note that, the general idea of our TEGAN can be applied to both the supervised setting and the unsupervised setting. Here, we discuss TEGAN in the unsupervised setting. Details of the supervised case of TEGAN are given in the supplementary.

As shown in Fig. [2] jointly modeling the result consistency and the transition consistency, our TEGAN is designed with two interactive generative processes. One is the vanilla translation process; the other is our proposed transition encoding process. The two generative process interact with each other through our design of generalized transition consistency on unseen transitions, i.e., Eq. (3) and Eq. (4).

3.4.1 Generation and reconstruction of image

Our loss terms for result consistency in the vanilla translation process, i.e., \( \hat{y} = G(x, t) \), includes:

The adversarial loss for image generation.

\[
\min_G \max_{D_{\text{Real}}} \mathcal{L}_{\text{Real, img}} = \mathbb{E}_{y \sim p(y)} [\log D_{\text{Real}}(y)] - \mathbb{E}_{(x, t)} [\log(1 - D_{\text{Real}}(G(x, t)))]
\tag{5}
\]

where the generator \( G \) aims to generate images that looks real, and the discriminator \( D_{\text{Real}} \) aims to distinguish between the real images and the generated images.

The losses for image reconstruction. Based on the essence of our transition variable, we adopt the following loss terms for image reconstruction.

The cycle-reconstruction loss is

\[
\min_G \mathcal{L}_{\text{Recons,newimg}} = \mathbb{E}_{(x, t)}[\| G(G(x, t), -t) - x \|_1]
\tag{6}
\]

The self-reconstruction loss is

\[
\min_G \mathcal{L}_{\text{Recons, img, self}} = \mathbb{E}_{x \sim p(x)}[\| G(x, 0) - x \|_1].
\tag{7}
\]

3.4.2 Generation and reconstruction of transition

Our losses for the transition encoding process include

The adversarial loss for transition generation. We adopt another discriminator on \( t \), namely \( D_t \), to learn a distribution for \( t \) with our encoder, and consequently facilitate the generation of new transitions, i.e., \( \tilde{t} \sim q(t|x, y) \equiv E(x, y) \).

\[
\min_E \max_{D_t} \mathcal{L}_{\text{Real, newtrans}} = \mathbb{E}_{x \sim p(x), y \sim p(y)}[\log D_{\text{Real}}(t)]
\tag{8}
\]

The loss for transition reconstruction. The loss for transition consistency on observed transitions, i.e., Eq. (1), is

\[
\begin{align*}
\min_E \mathcal{L}_{\text{Recons,trans}} & = \mathbb{E}_{(x, t, y) \sim p(x, t, y)}[\| E(x, t) - E(x, y) \|_1] \\
& + \mathbb{E}_{x \sim p(x), t \sim q(t|x, y)}[\| E(x, G(x, t)) - E(x, \tilde{t}) \|_1] \\
& + \mathbb{E}_{x \sim p(x)}[\| E(x, x) - 0 \|_1]
\end{align*}
\tag{9}
\]

3.4.3 Consistency losses for unseen translations

The above two generation process cooperatively interact through our design of Eq. (3) and Eq. (4), i.e., enforce transition consistency on the generative facilitated with the unseen transitions. The corresponding losses are

The adversarial loss for new generation:

\[
\begin{align*}
\min_G \max_{D_{\text{Real}}} \mathcal{L}_{\text{Real, newimg}} & = \mathbb{E}_{y \sim p(y)} [\log D_{\text{Real}}(y)] \\
& + \mathbb{E}_{x \sim p(x), \tilde{t} \sim q(t|x, y)} [\log(1 - D_{\text{Real}}(G(x, \tilde{t})))]
\end{align*}
\tag{10}
\]

The reconstruction of sampled transitions: The reconstruction loss defined for transition consistency on the randomly sampled transitions, i.e., Eq. (3) is

\[
\begin{align*}
\min_E \mathcal{L}_{\text{Recons,newtrans}} & = \mathbb{E}_{x \sim p(x), t' \sim q(t|x, y)}[\| E(x, G(x, t')) - E(x, \tilde{t}) \|_1] \\
& + \mathbb{E}_{x \sim p(x), \tilde{t} \sim q(t|x, y)}[\| E(x, G(x, \tilde{t})) - E(x, \tilde{t}) \|_1]
\end{align*}
\tag{11}
\]

The adversarial loss for triplet matching: We adopt a discriminator \( D_{\text{Match}} \) that takes triplet inputs, i.e., \( (x, t, y) \) to achieve our joint distribution matching design in Eq. (4).

\[
\begin{align*}
\min_G \max_{D_{\text{Match}}} \mathcal{L}_{\text{Match}} & = \mathbb{E}_{(x, t, y) \sim p(x, t, y)}[\log D_{\text{Match}}(x, t, y)] \\
& + \mathbb{E}_{(x, t) \sim p(x, t)}[\log (1 - D_{\text{Match}}(x, t, G(x, t)))] \\
& + \mathbb{E}_{x \sim p(x), t' \sim q(t|x, y)} [\log (1 - D_{\text{Match}}(x, \tilde{t}, G(x, \tilde{t})))]
\end{align*}
\tag{12}
\]
We further benefit triplet matching by incorporating wrong triplets \((x, t, y)\) and \((x, t, y_x)\), as in [19]. Our adversarial loss for triplet matching is then obtained as \(\mathcal{L}_{\text{Match}}^D\) and \(\mathcal{L}_{\text{Match}}^G\). (Pseudo-code given in the supplementary)

**Full objective.** Finally, the full objective of our TEGAN is

\[
G', E' = \arg \min_{G,E} \max_{D} \mathcal{L}_{\text{Real, img}} + \mathcal{L}_{\text{Real, newimg}} + \mathcal{L}_{\text{Real, newtrans}} + \lambda (\mathcal{L}_{\text{Match}}^D + \mathcal{L}_{\text{Match}}^G) + \lambda_1 (\mathcal{L}_{\text{Recons, img, cyc}} + \mathcal{L}_{\text{Recons, img, celt}}) + \lambda_2 (\mathcal{L}_{\text{Recons, trans}} + \mathcal{L}_{\text{Recons, newtrans}})
\]

(13)

where \(\lambda, \lambda_1, \lambda_2\) are hyper-parameters that control the relative importance of each term respectively.

### 3.5. Model training

Fig. 4 presents the network design of our TEGAN. Specifically, modeling bi-directional triplet matching, our TEGAN is trained with two phases. In the (a) phase, we do encoding and generation, within which the generation and reconstruction on results and transitions are both constrained, i.e., the losses presented in Sec. 3.4.1 and Sec. 3.4.2, respectively. In the (b), generation and encoding phase, we do reconstruction on the sampled transitions, i.e. Eq. (3). These two training phase are connected through the common discriminator \(D_{\text{Match}}\) for the joint distribution matching of the triplet data, i.e., Eq. (4).

### 4. Discussion

Our proposed TEGAN provides a general generative framework for I2I translation. The aforementioned models can all be covered or explained with our TEGAN. We summarize their connections and comparisons in Tab. 1.

Table 1. Comparison and connections of different I2I translation methods with our proposed TEGAN.

| Methods | Transition | Transition manipulation | Consistency |
|---------|------------|-------------------------|-------------|
| DPE [5] | ✓          | ✓                       | #           |
| CycleGAN [9] | ✓         | ✓                       | #           |
| TEGAN [10] | ✓         | ✓                       | #           |
| StarGAN [3] | ✓         | ✓                       | #           |
| RelGAN [19] | ✓         | ✓                       | #           |
| DLOW [8] | ✓         | ✓                       | #           |
| BicycleGAN [28] | ✓     | ✓                       | #           |
| AugCGAN [11] | ✓     | ✓                       | #           |
| TEGAN (ours) | ✓     | ✓                       | #           |

reconstruction constraint in [19] can be explained with transition as \(G(x, t = 0) \approx x\). Without explicitly model \(t(x, y)\), these methods can only regularize consistency on the observed transitions, limiting their generation capacity when translating with unseen transitions.

RelGAN [19]. DLOW [8], BicycleGAN [28] and AugCGAN [11] seeks to study consistency on unseen transitions. However, the generalization ability of their model could be inferior to our TEGAN in mainly two aspects. RelGAN and DLOW introduce unseen transitions via synchronized interpolation on the observed \(y_0\) and its corresponding image pair, i.e., \(G(x, \alpha \cdot t) \approx (1 - \alpha) \cdot x + \alpha \cdot y\). They can simply obtain \(\tilde{t}\) via interpolation, i.e., the simple linear case of \(\tilde{t} = \alpha \cdot t\) in the transition encoding of TEGAN (Fig 3), making their manipulation inflexible. In addition, the interpolated images may not be realistic itself, thus leading to unreasonable transitions \(\tilde{t} = \tilde{t}(x, \tilde{y})\) that fails to capture the intrinsic relations among the data, e.g. relations between attribute annotations in face editing tasks [19]. This can disorder the transition consistency defined on unseen transitions. BicycleGAN [28] and AugCGAN [11] flexibly manipulate \(t\) by encoding \(z_y\), and enforce result consistency via attribute prediction. However, such regularization can only work on the explicitly sampled transitions.

Our TEGAN possesses better generalization ability through jointly considering the result consistency and transition consistency for unseen transitions, and further generalizing both of them to distribution-level via joint distribution matching, i.e. our Eq. (5) and Eq. (4), respectively.
CelebA-HQ is a High-Quality images version of the CelebFaces Attributes (CelebA) dataset. The dataset consists of 30,000 high-quality face images in 1024 × 1024 resolution, each annotated with 40 binary attributes concerning hair colors, gender, age, etc.

Setting: We resize and center-crop these images to 256 × 256, and adopt 10 easily identifiable attributes, including hair color (black, blond, brown), dress-up (eyeglasses, bangs, mustache, pale skin, smiling), gender(male/female) and age(young/old), for experiment. Among all these images, we adopt a 90/10 split for training and testing [19]. We compare with three state-of-the-art face editing methods: AttGAN [10], StarGAN [3] and RelGAN [19].

Single attribute/seen transition: Fig. 5 presents example comparison results of TEGAN and other baseline methods regarding single facial attribute editing. Our results are more semantically meaningful than AttGAN and RelGAN, while the results of StarGAN are the worst with notable artifacts in most images. In the first set(the first four columns) of comparison, all the results of baseline methods trigger the smiling attribute (Mouth changes to be open), when smiling is not the desired changed attribute. This is not semantically consistent with our desired changes. In the second set of results, our results present more evident visual effect of attributes compared with RelGAN and AttGAN. This phenomenon is most evident in the comparison between the editing results of young and old (marked with red rectangle in Fig. 5), where our TEGAN presents obviously meaningful semantics, especially the color of hairs. Overall, the results of our TEGAN is more realistic and semantically meaningful. We present more results in the supplementary, as well as the comparisons on binary attribute interpolations.

Multiple attributes/unseen transition: We further conduct experiments on multiple attributes editing task, where modeling unseen transitions are more important.

Qualitative results: As shown in Fig. 6, our TEGAN generates more realistic and semantically meaningful results among the four methods. To be specific, the results of StarGAN present clear deficiencies, especially the one marked with red rectangle in the first row. It is due to the mismatch between StarGAN’s domain classifier and attribute annotations. RelGAN obtains more blurry results in this setting since it simply models the unseen transitions with linear operations, i.e., \( \tilde{t} = \alpha \cdot t \), which can not cover other flexible combinations of attributes that do not present in the training set. Although AttGAN achieves reasonable results, because these attributes are independently modeled, its generated results can not present meaningful fusion of changes according to the inherent relation between the attributes, e.g., mustache still exists when we change the gender from male to female.(See the second image in the third row) In contrast, results of TEGAN are more realistic and highly semantically meaningful, because of the ability of modeling unseen transitions in non-linear space. For example, TEGAN changes a male’s bangs when editing him to female in the

5. Experiments

We test the performance of TEGAN on four I2I translation tasks, including face editing, outdoor scene editing, multi-domain style transfer, and image inpainting. In these tasks, our transition possesses different semantic definition, according to the desired attribute change in each of them. Our translation results are expected to be both realistic and transition consistent, i.e., presenting the property defined by \( t \).

5.1. Face editing

We first test TEGAN on face editing, where the transition \( t \) specifies the change of facial attributes, i.e., \( x^t \leftarrow a_x \rightarrow y \), where \( a_x \) and \( a_y \) denote the attribute annotation of face images \( y \) and \( x \), respectively.

CelebA-HQ dataset: The CelebA-HQ is a High-Quality version of the CelebFaces Attributes (CelebA) dataset. The dataset consists of 30,000 high-quality face images in

[1] https://github.com/switchablenorms/CelebAMask-HQ

Figure 5. Comparisons of TEGAN with AttGAN, StarGAN and RelGAN on binary facial attribute editing on CelebA-HQ dataset. More semantically meaningful results are marked with red rectangles. Zoom in for better resolution.

Figure 6. Comparisons results on multiple facial attribute editing on CelebA-HQ dataset. More obvious comparisons are marked with red rectangles. Zoom in for better resolution.
Table 2. Fréchet Inception Distance (FID ↓) between input images and transformed images by three representative previous methods and TEGAN. Lower value means better.

| Metrics   | StarGAN | RelGAN | AttGAN | TEGAN |
|-----------|---------|--------|--------|-------|
| FID       | 125.94  | 71.78  | 72.22  | 66.93 |

We adopt Pix2Pix and BicycleGAN\(^2\) on supervised I2I translation on the Transient Attribute Dataset. More obvious results are marked with red rectangles. Zoom in for better resolution.

last image of the last row.

Quantitative results: We qualitatively evaluate and compare the generation results. Table 2 presents the comparison regarding Fréchet Inception Distance (FID)\(^1\) (lower is better) between the input image \(x\) and the translated image \(G(x, t)\). We can see TEGAN achieves the best results among all the methods, which indicates the translated images with high quality. Please refer to the supplementary for more results on multiple attributes editing and translated images by sampling unseen transitions.

5.2. Outdoor scenes editing

We test the performance of TEGAN on supervised attribute transfer on the Transient Attribute Dataset\(^1\).

Transient Attribute Dataset: It is a dataset containing scene images from 101 outdoor webcams. The images are captured over long time spans which exhibit drastic changes in appearance. The dataset contains 8571 images in total, each annotated with 40 transient attributes, such as sunny, bright and ice.

Setting: We select 20 visually evident attributes to define the transition and randomly select one image taken from the same camera for each single image to construct \(8571 \times 1\) paired images for experiment. 8076 of these images are randomly selected for training, and the rest data are for testing. We adopt Pix2Pix and BicycleGAN\(^2\), two state-of-the-art supervised I2I translation methods, for comparison.

Qualitative evaluation: We present the qualitative comparison results in Fig. 7. Specifically, the results of baseline methods are less realistic than our TEGAN. For example, in the image marked with red rectangle in the last row, Pix2Pix fails to preserve the semantic properties declared in the ground truth images, i.e. the river and houses are missing. In terms of the results of BicycleGAN in the first row, BicycleGAN is influenced by deformation in the target image \((y)\), with the details around the border of the image missing. In terms of semantic meanings presenting in the translated images, our results are more reasonable and semantically consistent with the declared transitions. In the second row of the figure, we expect the translated images would be a scene in winter with ice and snow. However, the translated images by Pix2Pix and RelGAN present more green color, which is not semantically consistent with the winter attribute. Overall, the results of our TEGAN present more realism and consistently semantic meanings. More results are given in the supplementary.

Quantitative evaluation: We further compute the SSIM\(^2\) and PSNR\(^2\) scores of the generated images by each method for quantitative evaluation. The evaluated image pairs are (1) input \(x\) and \(\hat{x} = G(x, 0)\) (the reconstruction of \(x\) with the 0 transition); (2) input \(x\) and \(\hat{y} = G(x, t)\) (the translated image of \(x\) in domain \(Y\)). From Table. 3, it is clear that our TEGAN achieves the highest score in all settings w.r.t. the two metrics.

Table 3. Comparison of the SSIM(↑) and PSNR(↑) of the translated images for the Transient Attribute Dataset. The best results are highlighted in bold. The larger the better.

| Metrics   | Image Pairs | Pix2Pix | BicycleGAN | TEGAN (ours) |
|-----------|-------------|---------|-------------|--------------|
| SSIM      | \(x\) and \(\hat{x} = G(x, 0)\) | -       | 0.52        | 0.82         |
|           | \(x\) and \(\hat{y} = G(x, t)\) | 0.25    | 0.41        | 0.52         |
| PSNR      | \(x\) and \(\hat{y} = G(x, t)\) | 8.5     | 10.41       | 12.36        |

\(^1\)https://github.com/junyanz/BicycleGAN

Figure 7. Comparisons of TEGAN with Pix2Pix\(^1\) and BicycleGAN\(^2\) on supervised I2I translation on the Transient Attribute Dataset. More obvious results are marked with red rectangles. Zoom in for better resolution.

Figure 8. Comparisons of TEGAN with DLOW\(^3\) on multi-domain style transfer. Images marked with red rectangles from DLOW present flaws as a fusion of different styles(purple color from monet). Zoom in for better resolution.
5.3. Multi-domain style transfer

Due to the flexibility of the semantics conveyed by transition, our TEGAN is handy to be generalized to conduct image-to-image translation with multiple target domains \( Y^{(1,2,...,n)} \), namely Multi-Domain I2I translation, with \( t \) specifies the target domain index of each transformation, \( i.e., x \rightarrow y^t \). Here, we adopt multi-domain style transfer as an example to testify TEGAN’s ability in multi-domain I2I translation tasks. The translation results are expected to be realistic and present obviously distinguishable visual characteristics for each target domain.

**Photo2Art dataset:** There are 4 commonly adopted photo→artistic painting datasets for style transfer task, \( i.e., photo\rightarrow caza\,mne, photo\rightarrow monet, photo\rightarrow ukiyoe \) and photo→vangogh. For each photo example \( x \), we randomly sample one art painting as \( y \) and consequently constitute one triplet data sample, where \( t \) specifies the style index of the target output. With all the photo images in these datasets, we finally construct a Photo2Art dataset containing a total number of 25148 photo→art image pairs.

**Setting:** We randomly select 500 samples among the whole dataset for testing, and use the rest of the data for training. We adopt DLOW \(^3\) as the baseline method.

**Qualitative evaluation:** Fig. 8 presents example comparisons regarding style transfer to each single domain. The results of our TEGAN present distinguishable stylization for each input, with each generated image presents recognizable style of the target domain. The results of DLOW fail to present clear distinction between the stylization of different domains. For example, the images in the second row present similar colors, and the same effect is shown in the fourth row. Besides, compared with the example images in the first row, some of results from DLOW also present evident flaw as a fusion of different styles. As shown in the figure, the images marked with red rectangles present some purple color pixels which is the representative color of monet style. The reason is that DLOW facilitates model generalization by providing interpolated intermediate transitions for simple mix-up \(^2\) of the existing training data, and constrain the transition consistency by preserving distance proportion, \( i.e., \) the domainness variable \( z \). In this way, the mix-up design may disorder the model from generating images with clear target property, \( i.e., z = 1 \). This shows the priority of our transition encoding.

5.4. Image inpainting

We additionally apply our TEGAN to image inpainting task, where \( t \) conveys the change of relative position between the realistic image region and the missing region, \( i.e., \) clues of the masking regions, in each masked input. The translation results should be realistic and plausible to filling the missing regions. More importantly, less evident boundaries are expected in our TEGAN, due to the enforced consistency between the region of mask and region of the inpainted pixels in the generated output.

**Celeba-HQ dataset:** We adopt images in the aforementioned CelebA-HQ dataset to conduct image inpainting task. We resize the original 1024×1024 images to 256×256 and erase an 128 × 128 squared region with an randomly sampled spatial location. Thus we get the inpainting dataset for TEGAN, where \( x \) is the marked incomplete image, \( y \) is the originally complete image, and \( t \) specifies the masked region in this data sample. We randomly select 1,000 samples among the whole dataset for testing; the rest of the data for training.

**Setting:** Following the state-of-the-art generative inpainting methods, we equip our TEGAN with an additional local critic, \( i.e., \) discriminator, to pay more concern on the quality of the inpainted regions. To clarify the semantics of transition in inpainting task, we expand \( t \) with one extra dimension to specify the translation direction in TEGAN, \( i.e., +1 \) indicates the positive direction to inpaint the marked images to complete ones, while \(-1 \) means the inverse direction, \( i.e., \) masking the complete image to the corresponding incomplete ones. We adopt the model of generative image Inpainting with Contextual Attention \(^3\) \( [23] \) (named as CA-Inpainting for simplicity), for a comparison.

**Qualitative evaluation:** Example results of the inpainting task is presented in Fig. 9. All these results are direct outputs of each model without any post-processing step. Considering the overall image quality, our TEGAN achieves comparable results with the CA-inpainting, the state-of-the-art method specialised in inpainting. However, the generated images of CA-inpainting suffers from sharp changes along the boundary regions of the mask. This is shown in the marked regions with blue rectangles in Fig. 9. For example, in the first row of figure, CA-Inpainting just generate parts of the eyebrows. The similar phenomenons are also shown in the other example images. Considering these

\(^3\) https://github.com/JiahuiYu/generative_inpainting
boundary cases, the generated results of our TEGAN is much better, since its inpainted regions present a harmonious integration with the given context images. The reason behind our advantage is that our TEGAN poses better generalization ability to inpainting with unseen masking regions through our generative transition encoding design.

6. Conclusion

This paper lodges a generative transition mechanism, TEGAN, to model the transition consistency among generated I2I translations. Our TEGAN not only covers existing GAN-based I2I translation models, but also shows superior generalization ability benefited from our consistent generative transition design on unseen transitions.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron C. Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In Jennifer G. Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 195–204. PMLR, 2018.

[2] A. Bulat and G. Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In CVPR, pages 109–117, 2018.

[3] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 8789–8797. IEEE Computer Society, 2018.

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8188–8197, 2020.

[5] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5154–5163, 2020.

[6] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. arXiv preprint arXiv:1605.09782, 2016.

[7] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 2414–2423. IEEE Computer Society, 2016.

[8] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. DLOW: domain flow for adaptation and generalization. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 2477–2486. Computer Vision Foundation / IEEE, 2019.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, 2014.

[10] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. IEEE Trans. Image Process., 28(11):5464–5478, 2019.

[11] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pages 327–340, 2001.

[12] 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 NIPS, pages 6626–6637, 2017.

[13] Bingwen Hu, Zhedong Zheng, Ping Liu, Wankou Yang, and Mingwu Ren. Unsupervised eyeglasses removal in the wild. IEEE Transactions on Cybernetics, 2020.

[14] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III, volume 11207 of Lecture Notes in Computer Science, pages 179–196. Springer, 2018.

[15] Philip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 5967–5976. IEEE Computer Society, 2017.

[16] Diederik P. Kingma and Max Welling. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NIPS, pages 6626–6637, 2014.

[17] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Temporally consistent image manipulation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 513–521, 2019.

[18] Pingbo Pan, Ping Liu, Yan Yan, Tianbao Yang, and Yi Yang. Relgan: Multi-domain image-to-image translation via 3d imitative-contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 2477–2486. Computer Vision Foundation / IEEE, 2019.

[19] Yu-Jing Lin, Po-Wei Wu, Che-Han Chang, Edward Y. Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 5913–5921. IEEE, 2019.

[20] Pingbo Pan, Ping Liu, Yan Yan, Tianbao Yang, and Yi Yang. Adversarial localized energy network for structured prediction. In AAAI 2020, pages 5347–5354. AAAI Press, 2020.
[21] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 2536–2544. IEEE Computer Society, 2016.

[22] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In ICCV 2019, pages 10560–10569. IEEE, 2019.

[23] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 5505–5514. IEEE Computer Society, 2018.

[24] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018. Conference Track Proceedings. OpenReview.net, 2018.

[25] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In ECCV, 2018.

[26] Shuai Zheng, Zhenfeng Zhu, Xingxing Zhang, Zhizhe Liu, Jian Cheng, and Yao Zhao. Distribution-induced bidirectional generative adversarial network for graph representation learning. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 7222–7231. IEEE, 2020.

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 2242–2251. IEEE Computer Society, 2017.

[28] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 465–476, 2017.