3D-Aware Multi-Class Image-to-Image Translation with NeRFs

Senmao Li[1] Joost van de Weijer[2] Yaxing Wang[1*]
Fahad Shahbaz Khan[3,4] Meiqin Liu[5] Jian Yang[1]
[1]VCIP,CS, Nankai University, [2]Universitat Autònoma de Barcelona
[3]Mohamed bin Zayed University of AI, [4]Linkoping University, [5]Beijing Jiaotong University
senmaonk@gmail.com {yaxing,csjyang}@nankai.edu.cn joost@cvc.uab.es
fahad.khan@liu.se mqliu@bjtu.edu.cn

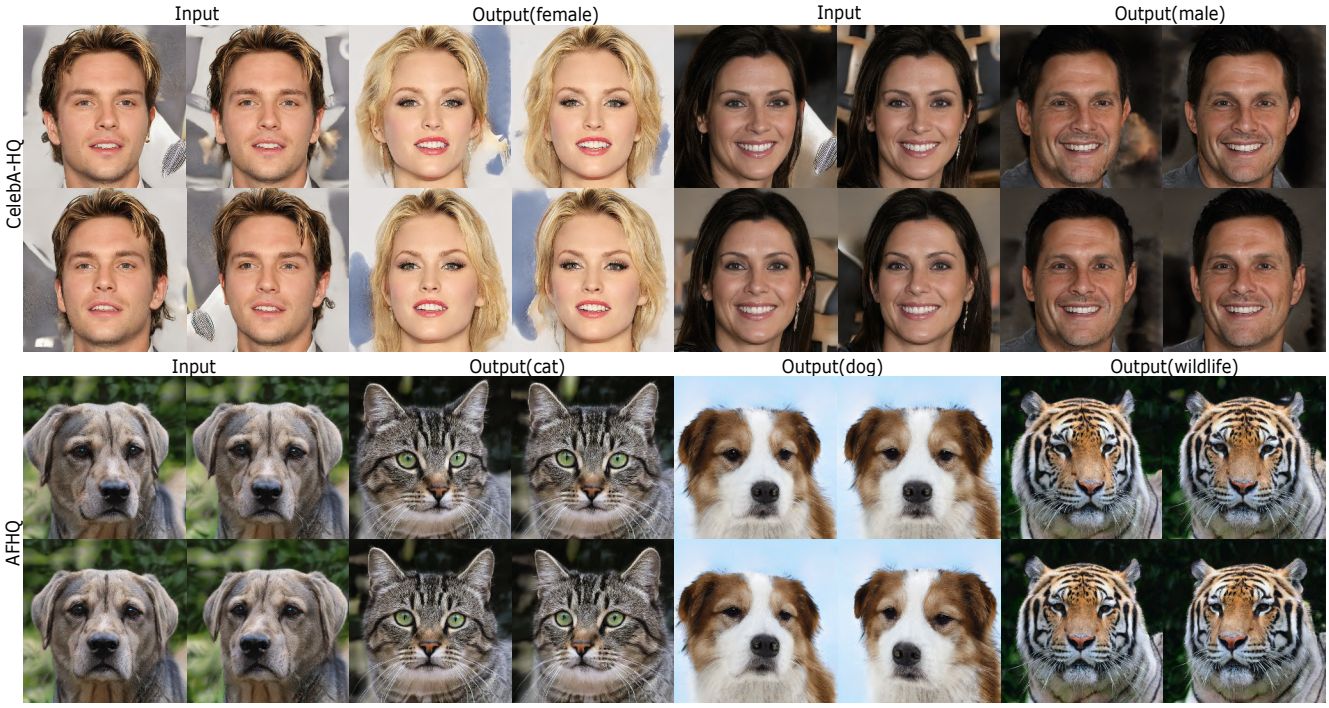

Figure 1. 3D-aware I2I translation: given a view-consistent 3D scene (the input), our method maps it into a high-quality target-specific image. Our approach produces consistent results across viewpoints.

Abstract

Recent advances in 3D-aware generative models (3D-aware GANs) combined with Neural Radiance Fields (NeRF) have achieved impressive results. However no prior works investigate 3D-aware GANs for 3D consistent multi-class image-to-image (3D-aware I2I) translation. Naively using 2D-I2I translation methods suffers from unrealistic shape/identity change. To perform 3D-aware multi-class I2I translation, we decouple this learning process into a multi-class 3D-aware GAN step and a 3D-aware I2I translation step. In the first step, we propose two novel techniques: a new conditional architecture and an effective training strategy. In the second step, based on the well-trained multi-class 3D-aware GAN architecture, that preserves view-consistency, we construct a 3D-aware I2I translation system. To further reduce the view-consistency problems, we propose several new techniques, including a U-net-like adaptor network design, a hierarchical representation constrain and a relative regularization loss. In extensive experiments on two datasets, quantitative and qualitative results demonstrate that we successfully perform 3D-aware I2I translation with multi-view consistency. Code is

*The corresponding author.

available in 3DI2I.

1. Introduction

Neural Radiance Fields (NeRF) have increasingly gained attention with their outstanding capacity to synthesize high-quality view-consistent images [31,39,66]. Benefiting from the adversarial mechanism [11], StyleNeRF [12] and concurrent works [4, 8, 44, 69] have successfully synthesized high-quality view-consistent, detailed 3D scenes by combining NeRF with StyleGAN-like generator design [22]. This recent progress in 3D-aware image synthesis has not yet been extended to 3D-aware I2I translation, where the aim is to translate in a 3D-consistent manner from a source scene to a target scene of another class (see Figure 1).

A naive strategy is to use well-designed 2D-I2I translation methods [15, 16, 26, 28, 46, 63, 65, 70]. These methods, however, suffer from unrealistic shape/identity changes when changing the viewpoint, which are especially notable when looking at a video. Main target class characteristics, such as hairs, ears, and noses, are not geometrically realistic, leading to unrealistic results which are especially disturbing when applying I2I to translate videos. Also, these methods typically underestimate the viewpoint change and result in target videos with less viewpoint change than the source video. Another direction is to apply video-to-video synthesis methods [2, 3, 6, 30, 53]. These approaches, however, either rely heavily on labeled data or multi-view frames for each object. In this work, we assume that we only have access to single-view RGB data.

To perform 3D-aware I2I translation, we extend the theory developed for 2D-I2I with recent developments in 3D-aware image synthesis. We decouple the learning process into a multi-class 3D-aware generative model step and a 3D-aware I2I translation step. The former can synthesize view-consistent 3D scenes given a scene label, thereby addressing the 3D inconsistency problems we discussed for 2D-I2I. We will use this 3D-aware generative model to initialize our 3D-aware I2I model. It therefore inherits the capacity of synthesizing 3D consistent images. To train effectively a multi-class 3D-aware generative model (see Figure 2(b)), we provide a new training strategy consisting of: (1) training an unconditional 3D-aware generative model (i.e., StyleNeRF) and (2) partially initializing the multi-class 3D-aware generative model (i.e., multi-class StyleNeRF) with the weights learned from StyleNeRF. In the 3D-aware I2I translation step, we design a 3D-aware I2I translation architecture (Figure 2(f)) adapted from the trained multi-class StyleNeRF network. To be specific, we use the main network of the pretrained discriminator (Figure 2(b)) to initialize the encoder E of the 3D-aware I2I translation model (Figure 2(f)), and correspondingly, the pretrained generator (Figure 2(b)) to initialize the 3D-aware I2I generator (Figure 2(f)). This initialization inherits the capacity of being sensitive to the view information.

Directly using the constructed 3D-aware I2I translation model (Figure 2(f)), there still exists some view-consistency problem. This is because of the lack of multi-view consistency regularization, and the usage of the single-view image. Therefore, to address these problems we introduce several techniques, including a U-net-like adaptor network design, a hierarchical representation constrain and a relative regularization loss.

In sum, our work makes the following **contributions**:

- We are the first to explore 3D-aware multi-class I2I translation, which allows generating 3D consistent videos.
- We decouple 3D-aware I2I translation into two steps. First, we propose a multi-class StyleNeRF. To train this multi-class StyleNeRF effectively, we provide a new training strategy. The second step is the proposal of a 3D-aware I2I translation architecture.
- To further address the view-inconsistency problem of 3D-aware I2I translation, we propose several techniques: a U-net-like adaptor, a hierarchical representation constraint and a relative regularization loss.
- On extensive experiments, we considerably outperform existing 2D-I2I systems with our 3D-aware I2I method when evaluating temporal consistency.

2. Related Works

Neural Implicit Fields. Using neural implicit fields to represent 3D scenes has shown unprecedented quality. [37, 38, 43, 45, 48, 51] use 3D supervision to predict neural implicit fields. Recently, NeRF has shown powerful performance to neural implicit representations. NeRF and its variants [31,39,66] utilize a volume rendering technique for reconstructing a 3D scene as a combination of neural radiance and density fields to synthesize novel views.

3D-aware GANs Recent approaches [5, 9, 13, 19, 35, 40–42, 52, 62, 68] learn neural implicit representations without 3D or multi-view supervisions. Combined with the adversarial loss, these methods typically randomly sample viewpoints, render photorealistic 2D images, and finally optimize their 3D representations. StyleNeRF [12] and concurrent works [4,8,44,69] have successfully synthesized high-quality view-consistent, detailed 3D scenes with StyleGAN-like generator design [22]. In this paper, we investigate 3D-aware image-to-image (3D-aware I2I) translation, where the aim is to translate in a 3D-consistent manner from a source scene to a target scene of another class. We combine transfer learning of GANs [55, 60].

I2I translation. I2I translation with GAN [16, 57, 59, 61] has increasingly gained attention in computer vision. Based

on the differences of the I2I translation task, recent works focus on paired I2I translation [10, 16, 71], unpaired I2I translation [1, 18, 24, 27, 32, 36, 46, 50, 56, 58, 63, 64, 70], diverse I2I translation [24,32,36,46,64,70] and scalable I2I translation [7, 29, 65]. However, none of these approaches addresses the problem of 3D-aware I2I. For the 3D scenes represented by neural implicit fields, directly using these methods suffers from view-inconsistency.

3. Method

Problem setting. Our goal is to achieve 3D consistent multi-class I2I translation trained on single-view data only. The system is designed to translate a viewpoint-video consisting of multiple images (source domain) into a new, photorealistic viewpoint-video scene of a target class. Furthermore, the system should be able to handle *multi-class* target domains. We decouple our learning into a multi-class 3D-aware generative model step and a multi-class 3D-aware I2I translation step.

3.1. Multi-class 3D-aware generative model

Let $\mathcal{I}_{\mathcal{RGB}} \in \mathbb{R}^{H\times W\times 3}$ be in the image domain. In this work, we aim to map a source image into a target sample conditioned on the target domain label $\boldsymbol{l} \in \{1,\dots,L\}$ and a random noise vector $\mathbf{z} \in \mathbb{R}^{\mathbf{Z}}$. Let vector $\boldsymbol{x}$ and $\boldsymbol{d}$ be 3D location and 2D viewing direction, respectively.

Unconditional 3D-aware generative model. StyleNeRF [12] introduces a 5D function (3D location $\boldsymbol{x}$ and 2D viewing direction $\boldsymbol{d}$) to predict the volume density σ and RGB color $\boldsymbol{c}$. Both σ and $\boldsymbol{c}$ are further used to render an image. As shown on Figure 2(a) StyleNeRF consists of four subnetworks: a mapping network M, a fully connected layer F, a generator G and a discriminator D. The mapping network M takes random noise $\boldsymbol{z}$ as input, and outputs latent code $\boldsymbol{w}$, which is further fed into both the fully connected layer F and generator G. Given the 3D location $\boldsymbol{x}$, the 2D viewing direction $\boldsymbol{d}$ and latent code $\boldsymbol{w}$, StyleNeRF renders the feature map $\boldsymbol{f}$:

$$\begin{aligned}
\boldsymbol{f}(\boldsymbol{r}) &= \int_0^\infty p(t)\boldsymbol{c}(\boldsymbol{r}(t),\boldsymbol{d})dt \\
p(t) &= \exp\left(-\int_0^t \sigma(\boldsymbol{r}(s))ds\right)\cdot \sigma_{\boldsymbol{w}}(\boldsymbol{r}(t)) \\
\boldsymbol{c},\sigma &= F(\boldsymbol{x},\boldsymbol{d},\boldsymbol{w}),
\end{aligned} \tag{1}$$

where $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$ ($\boldsymbol{o}$ is the camera origin) is a camera ray for each feature representation position. Generator G takes as an input the representation $\boldsymbol{f}$ and the latent code $\boldsymbol{w}$, and outputs view-consistent photo-realistic novel result $\hat{I}_{RGB}$. The discriminator D is to distinguish real images I_{RGB} from generated images $\hat{I}_{RGB}$.

The fully objective of StyleNeRF is as following:

$$\begin{aligned}
\mathcal{L}_G &= \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z},\boldsymbol{p}\sim\mathcal{P}}\left[v(D(G(F(\boldsymbol{z},\boldsymbol{x},\boldsymbol{d}),M(\boldsymbol{z}))))\right] \\
&+ \mathbb{E}_{I_{RGB}\sim p_{\text{data}}}\left[v(-D(I_{RGB}) + \lambda\|\nabla D(I_{RGB})\|^2)\right] \\
&+ \beta\cdot\mathcal{L}_{\text{NeRF-path}}
\end{aligned} \tag{2}$$

where $v(u) = -\log(1+\exp(-u))$, and p_{data} is the data distribution. $\mathcal{L}_{\text{NeRF-path}}$ is NeRF path regularization used in StyleNeRF. We also set $\beta = 0.2$ and $\lambda = 0.5$ following StyleNeRF.

Conditional 3D-aware generative model. Figure 2(b) shows the proposed multi-class 3D-aware generative model (i.e., multi-class StyleNeRF). Compared to the StyleNeRF architecture (Figure 2(a)), we introduce two mapping networks: M_1 and M_2. The mapping network M_1 outputs the latent code $\boldsymbol{w}_1$. While the mapping network M_2 takes as input the concatenated noise $\boldsymbol{z}$ and class embedding $e_{\boldsymbol{l}\text{-}th}$, and outputs the latent code $\boldsymbol{w}_2$. The second mapping network M_2 aims to guide the generator G to synthesize a class-specific image. Here we do not feed the latent code $\boldsymbol{w}_2$ into NeRF's fully connected layer F, since we expect F to learn a class-agnostic feature representation, which contributes to perform multi-class 3D-aware I2I translation.

To be able to train multi-class StyleNeRF we adapt the loss function. We require D to address multiple adversarial classification tasks simultaneously, as in [33]. Specifically, given output $D \in \mathbb{R}^L$, we locate the $\boldsymbol{l}\text{-}th$ class response. Using the response for the $\boldsymbol{l}\text{-}th$ class, we compute the adversarial loss and back-propagate gradients:

$$\begin{aligned}
\mathcal{L}_G^l &= \mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z},\boldsymbol{x}\sim\mathcal{P}_x,\boldsymbol{d}\sim\mathcal{P}_d}\left[v(D(G(\hat{I}_{RGB}))_{\boldsymbol{l}\text{-}th}\right] \\
&+ \mathbb{E}_{I_{RGB}\sim p_{\text{data}}}\left[v(-D(I_{RGB})_{\boldsymbol{l}\text{-}th} + \lambda\|\nabla D(I_{RGB})_{l_{th}}\|^2)\right] \\
&+ \beta\cdot\mathcal{L}_{\text{NeRF-path}}.
\end{aligned} \tag{3}$$

We initialize the multi-class StyleNeRF with the weights learned with the unconditional StyleNeRF (E.q. 2), since the training from scratch fails to convergence. Results of this are show in Figs. 7. To be specific, we directly copy the weights from the one learned from StyleNeRF for M_1, F and G with the same parameter size. For the mapping network M_2, we duplicate the weight from M except for the first layer, which is trained from scratch because of the different parameter sizes. The discriminator is similarly initialized except for the last layer, which is a new convolution layer with L output channels. Using the proposed initialization method, we successfully generate class-specific photo-realistic high-resolution result.

3.2. 3D-aware I2I translation

Figure 2 (f) shows the 3D-aware I2I translation network at inference time. It consists of the encoder E, the generator G and two mapping networks M_1 and M_2. Inspired

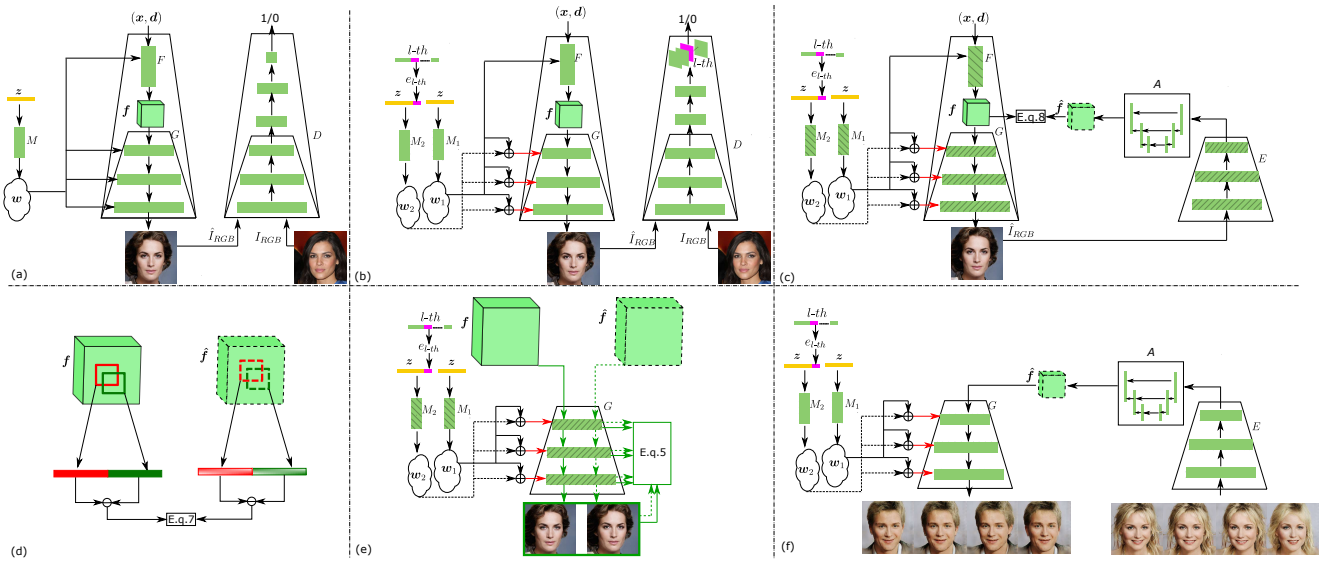

Figure 2. Overview of our method. (a) We first train a 3D-aware generative mode (i.e., StyleNeRF) with single-view photos. (b) We extend StyleNerf to multi-class StyleNerf. We introduce an effective training strategy: initializing multi-class StyleNeRF with StyleNeRF. (c) The training of the proposed 3D-aware I2I translation. It consists of the encoder E, the adaptor A, the generator G and two mapping networks M_1 and M_2. We freeze all networks except for training the adaptor A. The encoder is initialized by the main networks of the pretrained discriminator. We introduce several techniques to address the view-consistency problems: including a U-net-like adaptor A, (d) relative regularization loss and (e) hierarchical representation constrain. (f) Usage of proposed model at inference time.

by DeepI2I [61], we use the pretrained discriminator (Figure 2(b)) to initialize the encoder E of the 3D-aware I2I translation model (Figure 2(f)), and correspondingly, the pretrained generator (Figure 2(b)) to initialize the 3D-aware I2I generator. To align the encoder with the generator, [61] introduces a Resnet-like adaptor network to communicate the encoder and decoder. The adaptor is trained without any real data. However, directly using these techniques for 3D-aware I2I translation still suffers from some view-consistency problems. Therefore, in the following, we introduce several designs to address this problem: a U-net-like adaptor network design, a hierarchical representation constrain and a relative regularization loss.

U-net-like adaptor. As shown in Figure 2(c), to overcome 3D-inconsistency in the results, we propose a U-net-like adaptor A. This design contributes to preserve the spatial structure of the input feature. This has been used before for semantic segmentation tasks and label to image translation [17]. In this paper, we experimentally demonstrate that the U-net-like adaptor is effective to reduce the inconsistency.

Hierarchical representation constrain. As shown in Figure 2(e), given the noise $\boldsymbol{z}$, 3D location $\boldsymbol{x}$ and 2D viewing direction $\boldsymbol{d}$ the fully connected layer F renders the 3D-consistent feature map $\boldsymbol{f} = F(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{w}_1) = F(\boldsymbol{x}, \boldsymbol{d}, M1(\boldsymbol{z}))$. We further extract the hierarchical representation $\{G(\boldsymbol{f}, \boldsymbol{w}_1, \boldsymbol{w}_2)_k\}$ as well as the synthesized image $\hat{I}_{RGB} = G(\boldsymbol{f}, \boldsymbol{w}_1, \boldsymbol{w}_2)$. Here $G(\boldsymbol{f}, \boldsymbol{w}_1, \boldsymbol{w}_2)_k$ is the $k\text{-}th (k = m, \ldots, n, (n > m))$ ResBlock [1] output of the generator G. We then take the generated image $\hat{I}_{RGB}$ as input for the encoder E: $E(\hat{I}_{RGB})$, which is fed into the adaptor network A, that is $\hat{\boldsymbol{f}} = A(E(\hat{I}_{RGB}))$. In this step, our loss is

$$\mathcal{L}_A = \left\| \boldsymbol{f} - \hat{\boldsymbol{f}} \right\|_1. \tag{4}$$

For the intermediate layers, we propose a hierarchical representation constrain. Given the output $\hat{\boldsymbol{f}}$ and the latent codes (i.e., $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$) [2], we similarly collect the hierarchical feature $\left\{G(\hat{\boldsymbol{f}}, \boldsymbol{w}_1, \boldsymbol{w}_2)_k\right\}$. The objective is

$$\mathcal{L}_H = \sum_k \left\| G(\boldsymbol{f}, \boldsymbol{w}_1, \boldsymbol{w}_2)_k - G(\hat{\boldsymbol{f}}, \boldsymbol{w}_1, \boldsymbol{w}_2)_k \right\|_1. \tag{5}$$

In this step, we freeze every network except for the U-net-like adaptor which is learned. Note that we do not access to any real data to train the adaptor, since we utilize the generated image with from the trained generator (Figure 2(b)).

Relative regularization loss. We expect to input the consistency of the translated 3D scene with single-image regularization [3] instead of the images from the consecutive views. We propose a relative regularization loss based on neighboring patches. We assume that neighboring patches

[1] After each ResBlock the feature resolution is half of the previous one in the encoder, and two times in generator. In the generator, the last output is image.

[2] Both $\boldsymbol{w}1$ and $\boldsymbol{w}2$ are the ones used when generating image $\hat{I}_{RGB}$

[3] More precisely, that is the feature map in this paper.

Method \ Dataset	CelebA-HQ		AFHQ	
	TC↓	FID↓	TC↓	FID↓
*MUNIT	30.240	31.4	28.497	41.5
*DRIT	35.452	52.1	25.341	95.6
*MSGAN	31.641	33.1	34.236	61.4
StarGANv2	10.250	**13.6**	3.025	16.1
Ours (3D)	**3.743**	22.3	**2.067**	**15.3**
	TC↓	(unc)FID↓	TC↓	(unc)FID↓
†Liu *et al.* [34]	13.315	17.8	3.462	20.0
StarGANv2	10.250	12.2	3.025	**9.9**
†Kunhee *et al.* [23]	10.462	**6.7**	3.241	10.0
Ours (3D)	**3.743**	18.7	**2.067**	11.4

Table 1. Comparison with baselines on TC and FID metrics.* denotes that we used the results provided by StarGANv2. † means that we used the pre-trained networks provided by authors.

are equivalent to that on corresponding patches of two consecutive views. For example, when inputting multi-view consistent scene images, the position of eyes are consistently moving. The fully connected layers (i.e., NeRF mode) F renders the view-consistent feature map $\boldsymbol{f}$, which finally decides the view-consistent reconstructed 3D scene. Thus, we expect the output $\hat{\boldsymbol{f}}$ of the adaptor A to obtain the view-consistent property of the feature map $\boldsymbol{f}$.

We randomly sample one vector from the feature map $\boldsymbol{f}$ (e.g., red square in (Figure 2(d))), denoted as $\boldsymbol{f}^{\eta}$. Then we sample the *eight* nearest neighboring vectors of $\boldsymbol{f}^{\eta}$ (dark green square in Figure 2(d))), denoted by $\boldsymbol{f}^{\eta,\varepsilon}$ where $\varepsilon = 1, \cdots, 8$ is the neighbor index. Similarly, we sample vectors $\hat{\boldsymbol{f}}^{\eta}$ and $\hat{\boldsymbol{f}}^{\eta,\varepsilon}$ from the feature map $\hat{\boldsymbol{f}}$ (red and dark green dash square in Figure 2(d))). We then compute the patch difference:

$$d_{\boldsymbol{f}}^{\eta,\varepsilon} = \boldsymbol{f}^{\eta} \ominus \boldsymbol{f}^{\eta,\varepsilon}, \; d_{\hat{\boldsymbol{f}}}^{\eta,\varepsilon} = \hat{\boldsymbol{f}}^{\eta} \ominus \hat{\boldsymbol{f}}^{\eta,\varepsilon}, \tag{6}$$

where $\ominus$ represents vector subtraction. In order to preserve the consistency, we force these patch differences to be small:

$$\mathcal{L}_R = \left\| d_{\boldsymbol{f}}^{\eta,\varepsilon} - d_{\hat{\boldsymbol{f}}}^{\eta,\varepsilon} \right\|_1. \tag{7}$$

The underlying intuition is straightforward: the difference vectors of the same location should be most relevant in the latent space compared to other random pairs.

The final objective is

$$\mathcal{L} = \mathcal{L}_H + \mathcal{L}_A + \mathcal{L}_R. \tag{8}$$

4. Experiments

4.1. Experimental setup

Training details. We use the trained StyleNeRF to partially initialize our multi-class StyleNeRF architecture. We adapt the structure of the multi-class StyleNeRF to the 3D-aware I2I architecture. The proposed method is implemented in Pytorch [47]. We use Adam [25] with a batch size of 64, using a learning rate of 0.0002. We use 2× Quadro RTX 3090 GPUs (24 GB VRAM) to conduct all our experiments. We show the network details and more results on Supp. Mat..

Ini.	Ada.	Hrc.	Rrl.	TC↓	FID↓
Y	N	N	N	2.612	23.8
Y	Y	N	N	2.324	23.1
Y	Y	Y	N	2.204	16.1
Y	Y	Y	Y	**2.067**	**15.3**

Table 2. Impact of several components in the performance on AFHQ. The second row is the case where the 3D-aware I2I translation model is initialized by weights learned from the multi-class StylyNeRF. Then it is trained with a Resnet-based adaptor and L_1 loss between the representations $\boldsymbol{f}$ and $\hat{\boldsymbol{f}}$. The proposed techniques continuously improve the consistency and performance. Ini.: initialization method for multi-class StyleNeRF, Ada.: U-net-like adaptor, Hrc.: Hierarchical representation constrain, Rrl: Relative regularization loss.

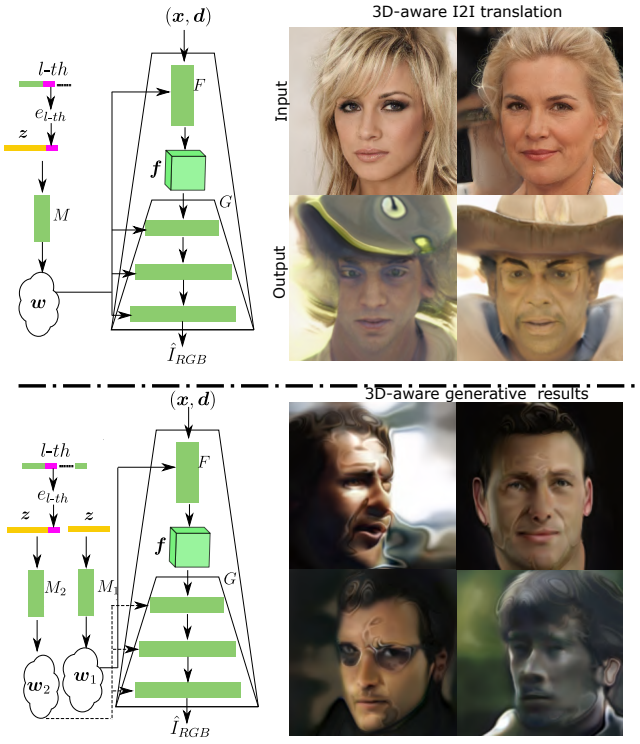

Figure 3. (Top) Using a single mapping network which takes as input the concatenated class embedding and the noise. We find it fails to generate target-specific realistic image. (Bottom) we use two mapping networks without concatenating their outputs like the proposed method. This design fails to generate 3D-aware results.

Datasets. Our experiments are conducted on the Animal Faces (AFHQ) [7] and CelebA-HQ [21] datasets. AFHQ contains 3 classes, each one has about 5000 images. In CelebA-HQ, we use gender as a class, with ∼10k(10057) male and ∼18k(17943) female images in the training set. In this paper, all images are resized to 256×256.

Figure 4. Comparative results between the proposed method and StarGANv2. We observe that StarGANv2 suffers from underestimating viewpoint changes when changing the input viewpoint (first column). It also leads to identity change (third and fourth columns), and a geometrically unrealistic ear (last two columns).

Figure 5. The generated images of (top) $G(\boldsymbol{f}, \boldsymbol{w}_1, \boldsymbol{w}_2)$ and (bottom) $G(\hat{\boldsymbol{f}}, \boldsymbol{w}_1, \boldsymbol{w}_2)$, which show that we correctly align the outputs of both the NeRF mode F and the adaptor A.

Baselines. We compare to MUNIT [15], DRIT [28], MSGAN [20], StarGANv2 [7], [23] and [34], all of which perform image-to-image translation.

Evaluation Measures. We employ the widely used metric for evaluation, namely Fréchet Inception Distance (FID) [14]. We also propose a new measure in which we combine two metrics, one which measures the consistency between neighboring frames (which we want to be low), and another that measures the diversity over the whole video (which we would like to be high). We adopt a modified *temporal loss* (TL) [54]. This temporal loss computes the Frobenius difference between two frames to evaluate the video consistency. Only considering this measure would lead to high scores when neighboring frames in the generated video are all the same. For successful 3D-aware I2I translation, we expect the system to be sensitive to view changes in the source video and therefore combine low consecutive frame changes with high diversity over the video. Therefore, we propose to compute LPIPS [67] for each video (vLPIPS), which indicates the diversity of the generated video sequence. To evaluate both the consistency and the sensitiveness of the generated video, we propose a new temporal consistency metric (TC):

$$TC = TL/vLPIPS. \tag{9}$$

Due to the small changes between two consecutive views, for each video we use frame interval 1, 2 and 4 in between to evaluate view-consistency. Note that a lower TC value is better.

4.2. Quantitative and qualitative results.

We evaluate the performance of the proposed method on both the AFHQ animal and CelebA human face dataset. As reported in Table 1, in terms of TC the proposed method achieves the best score on two datasets. For example, we

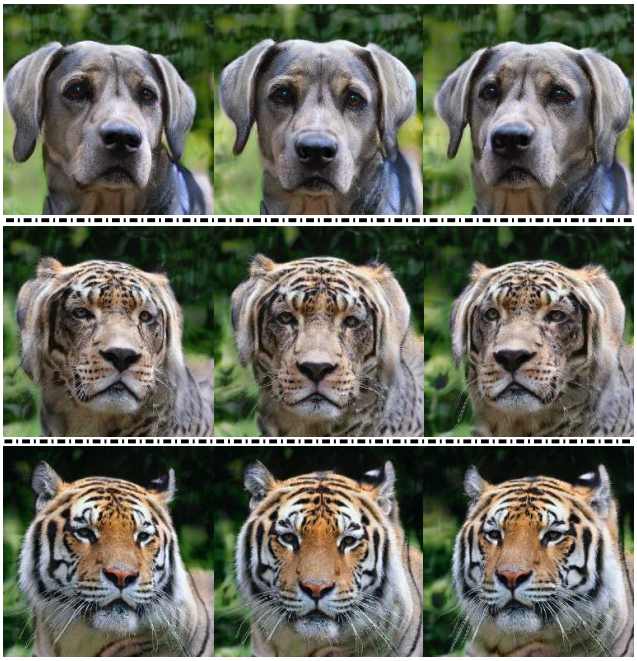

Figure 6. Interpolation between the dog and wildlife classes.

have 3.743 TC on CelebA-HQ, which is better than StarGANv2 (10.250 TC). This indicates that our method dramatically improves consistency. As reported in Table 1 (up), across both datasets, the proposed method consistently outperforms the baselines with significant gains in terms of FID and LPIPS, except for StarGANv2 which obtains superior results. However, on AFHQ we achieve better FID score than StarGANv2. Kunhee *et al.* [23] reports the unconditional FID (*(unc)FID*) value which is computed between synthesized images and training samples instead of each class. As reported in Table 1 (bottom), We are able to achieve completing results on uncFID metrics. Note that while 2D I2I translation (e.g., StarGANv2) can obtain high-quality for each image, they cannot synthesize images of the same scene with 3D consistency, and suffers from unrealistic shape/identity changes when changing the viewpoint, which are especially notable when looking at a video.

In Figures 1,4, we perform 3D-aware I2I translation. When changing the input viewpoint (Figure 4 (first two columns)), the outputs of StarGANv2 do not maintain the correct head pose, and underestimate the pose changes with respect to the frontal view. To estimate that this is actually the case, we also compute the diversity (i.e., vLPIPS) in a single video sequence. For example, both StarGANv2 and our method are 0.032 and 0.101 on CelebA-HQ. This confirms that the diversity (due to pose changes) is lowest for StarGANv2. More clearly showing the limitations of standard I2I methods for 3D-aware I2I, we observe that StarGANv2 suffers from unrealistic changes when changing the viewpoint. For example, when translating the class *cat* to *wildlife*, the generated images changes from *wolf* to *leopard* when varying the viewpoint (Figure 4 (third and fourth columns)). Also, the main target class characteristics, such as ears, are not geometrically realistic, leading to unrealistic 3D scene videos. Our method, however, eliminates these shortcomings and performs efficient high-resolution image translation with high 3D-consistency, which preserves the input image pose and changes the style of the output images. We show high-resolution images (1024×1024) on Supp. Mat..

4.3. Ablation study

Conditional 3D-aware generative architecture In this experiment, we verify our network design by comparing it with two alternative network designs. As shown in Figure 3(up), we explore a naive strategy: using one mapping which takes as input the concatenated class embedding and the noise. In this way, the fully connected network F outputs the *class-specific* latent code $\boldsymbol{w}$, which is fed into the fully connected network F to output the*class-specific* representation $\boldsymbol{f}$. Here, both the latent code $\boldsymbol{w}$ and the representation $\boldsymbol{f}$ are decided by the same class. However, when handling 3D-aware multi-class I2I translation task, the feature representation $\hat{\boldsymbol{f}}$ is combined with the latent code $\boldsymbol{w}$ from varying class embeddings, which leads to unrealistic image generation (Figure. 3(up)).

As shown in Figure 3(bottom), we utilize two mapping networks without concatenating their outputs like the proposed method. This design guarantees that the output of the fully connected layers F are *class-agnostic*. We experimentally observe that this model fails to handle 3D-aware generation.

Effective training strategy for multi-class 3D-aware generative model. We evaluate the proposed training strategy on AFHQ and CelebA-HQ datasets. We initialize the proposed multi-class 3D I2I architecture from scratch and the proposed method, respectively. As shown on Figure 7 (up), the model trained from scratch synthesizes unrealistic faces on CelebA-HQ dataset, and low quality cats on AFHQ. This is due to the style-based conditional generator which is hard to be optimized and causes mode collapse directly [49]. The proposed training strategy, however, manages to synthesize photo-realistic high-resolution images with high multi-view consistency. This training strategy first performs unconditional learning, which leads to satisfactory generative ability. Thus, we relax the difficulty of directly training the conditional model.

Alignment and interpolation. Figure 5 exhibits the outputs of the generator when taking as input the feature representation $\boldsymbol{f}$ and $\hat{\boldsymbol{f}}$. This confirms that the proposed method successfully aligns the outputs of the fully connected layers F and the adaptor A. Figure 6 reports interpolation by freezing the input images while interpolating the class em-

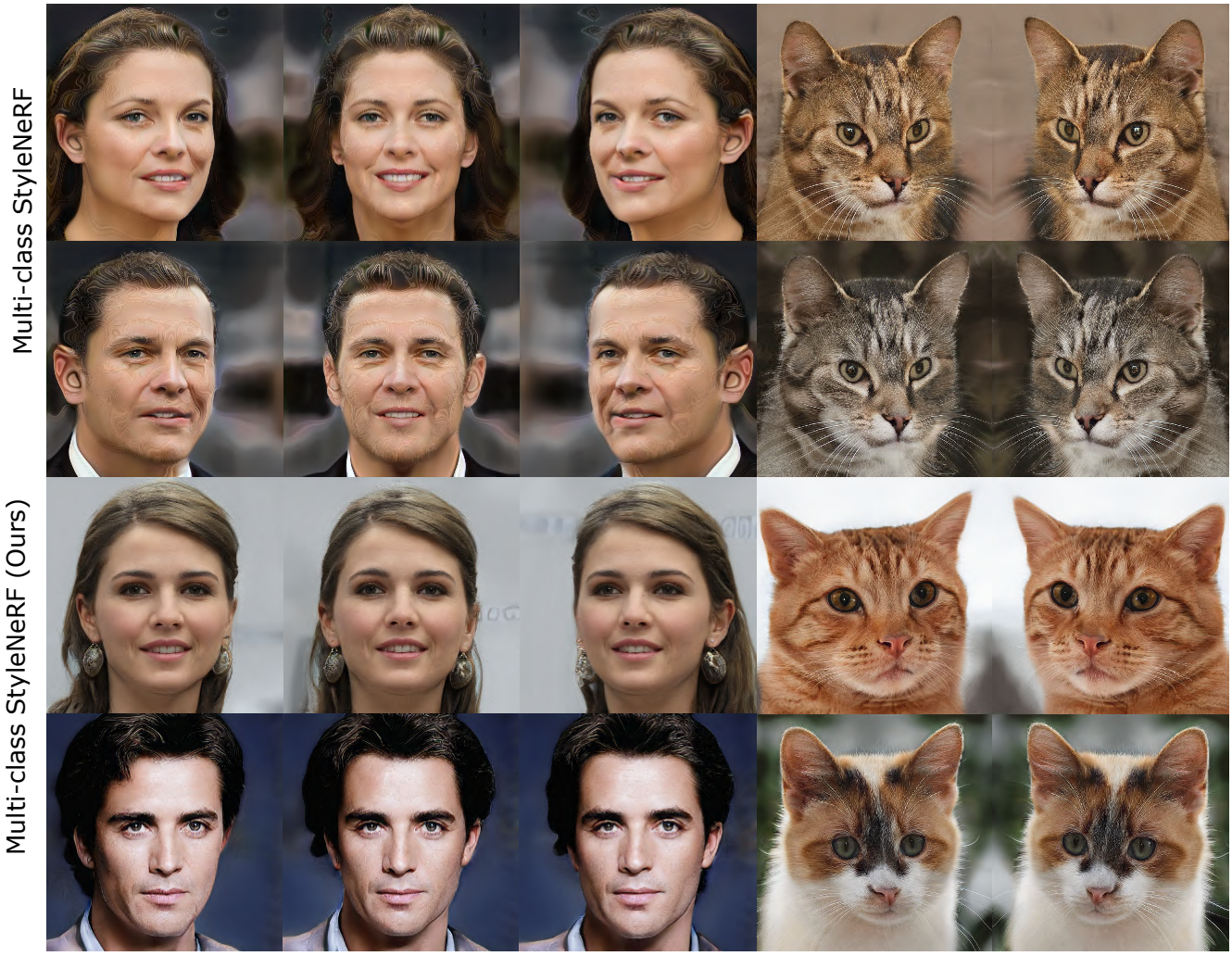

Figure 7. Qualitative results of multi-class StyleNeRF training from scratch (up) and from the proposed strategy (bottom).

bedding between two classes. Our model still manages to preserve the view-consistency, and generate high quantity images with even given never seen class embeddings.

Techniques for improving the view-consistency. We perform an ablation study on the impact of several design elements on the overall performance of the system, which includes the proposed initialization 3D-aware I2I translation model (Ini.), U-net-like adaptor (Ada.), hierarchical representation constrain (Hrc.) and relative regularization loss (Rrl.). We evaluate these four factors in Table 2. The results show that only using the proposed initialization (the second row of the Table 2) has already improved the view-consistency comparing to StarGANv2 (Table 1). Utilizing either U-net-like adaptor (Ada.) or hierarchical representation constrain (Hrc.) further leads to performance gains. Finally we are able to get the best score when further adding relative regularization loss (Rrl.) to the 3D-aware I2I translation model.

5. Conclusion

In this paper we first explore 3D-aware I2I translation. We decouple the learning process into a multi-class 3D-aware generative model step and a 3D-aware I2I translation step. In the first step, we propose a new multi-class StyleNeRF architecture, and an effective training strategy. We design the 3D-aware I2I translation model with the well-optimized multi-class StyleNeRF model. It inherits the capacity of synthesizing 3D consistent images. In the second step, we propose several techniques to further reduce the view-consistency of the 3D-aware I2I translation.

Acknowledgement. We acknowledge the support from the Key Laboratory of Advanced Information Science and Network Technology of Beijing (XDXX2202), and the project supported by Youth Foundation (62202243). We acknowledge the Spanish Government funding for projects PID2019-104174GB-I00, TED2021-132513B-I00.

References

[1] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14154–14163, 2021. 3

[2] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *Proceedings of the European conference on computer vision (ECCV)*, pages 119–135, 2018. 2

[3] Dina Bashkirova, Ben Usman, and Kate Saenko. Unsupervised video-to-video translation. *arXiv preprint arXiv:1806.03698*, 2018. 2

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2

[5] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2

[6] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. Mocycle-gan: Unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 647–655, 2019. 2

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 3, 5, 6

[8] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10673–10683, 2022. 2

[9] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017. 2

[10] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *NeurIPS*, pages 1294–1305, 2018. 3

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2

[12] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 2, 3

[13] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7498–7507, 2020. 2

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 6

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018. 2, 6

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 3

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 4

[18] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. Memory-guided unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6558–6567, 2021. 3

[19] Danilo Jimenez Rezende, SM Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. *Advances in neural information processing systems*, 29, 2016. 2

[20] Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7799–7808, 2020. 6

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 5

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2

[23] Kunhee Kim, Sanghun Park, Eunyeong Jeon, Taehun Kim, and Daijin Kim. A style-aware discriminator for controllable image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18239–18248, 2022. 5, 6, 7

[24] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 3

[25] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014. 5

[26] Minsu Ko, Eunju Cha, Sungjoo Suh, Huijin Lee, Jae-Joon Han, Jinwoo Shin, and Bohyung Han. Self-supervised dense consistency regularization for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18301–18310, June 2022. 2

[27] Héctor Laria, Yaxing Wang, Joost van de Weijer, and Bogdan Raducanu. Hyper-gan: Transferring unconditional to conditional gans with hypernetworks. *arXiv preprint arXiv:2112.02219*, 2021. 3

[28] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 2, 6

[29] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *IJCV*, pages 1–16, 2020. 3

[30] Kangning Liu, Shuhang Gu, Andrés Romero, and Radu Timofte. Unsupervised multimodal video-to-video translation via self-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1030–1040, 2021. 2

[31] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 2

[32] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, pages 700–708, 2017. 3

[33] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *CVPR*, pages 10551–10560, 2019. 3

[34] Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10785–10794, 2021. 5, 6

[35] Sebastian Lunz, Yingzhen Li, Andrew Fitzgibbon, and Nate Kushman. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv preprint arXiv:2002.12674*, 2020. 2

[36] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *NeurIPS*, pages 3693–3703, 2018. 3

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[38] Mateusz Michalkiewicz, Jhony K. Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*, 2020. 2

[40] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2

[41] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *2021 International Conference on 3D Vision (3DV)*, pages 951–961. IEEE, 2021. 2

[42] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2

[43] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *arXiv preprint arXiv:1912.07372*, 2019. 2

[44] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 2

[45] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[46] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for conditional image synthesis. In *ECCV*, 2020. 2, 3

[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[48] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *ArXiv*, abs/2003.04618, 2020. 2

[49] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 7

[50] Xuning Shao and Weidong Zhang. Spatchgan: A statistical feature based discriminator for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6546–6555, 2021. 3

[51] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, pages 1119–1130, 2019. 2

[52] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2022. 2

[53] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018. 2

[54] Wenjing Wang, Shuai Yang, Jizheng Xu, and Jiaying Liu. Consistent video style transfer via relaxation and regularization. *IEEE Transactions on Image Processing*, 29:9125–9139, 2020. 6

[55] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *CVPR*, 2020. 2

[56] Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, and Luis Herranz. SDIT: Scalable and diverse cross-domain image translation. In *ACM MM*, 2019. 3

[57] Yaxing Wang, Salman Khan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Semi-supervised learning for few-shot image-to-image translation. In *CVPR*, 2020. 2

[58] Yaxing Wang, Hector Laria Mantecon, Joost van de Weijer, Laura Lopez-Fuentes, and Bogdan Raducanu. Transferi2i: Transfer learning for image-to-image translation from small datasets, 2021. 3

[59] Yaxing Wang, Joost van de Weijer, and Luis Herranz. Mix and match networks: encoder-decoder alignment for zero-pair image translation. In *CVPR*, pages 5467–5476, 2018. 2

[60] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *ECCV*, pages 218–234, 2018. 2

[61] Yaxing Wang, Lu Yu, and Joost van de Weijer. Deepi2i: Enabling deep hierarchical image-to-image translation by transferring from gans. *NeurIPS*, 2020. 2, 4

[62] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18440–18449, 2022. 2

[63] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Unsupervised image-to-image translation with generative prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18332–18341, 2022. 2, 3

[64] Zili Yi, Hao Zhang, Ping Tan Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 3

[65] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *NeurIPS*, pages 2990–2999, 2019. 2, 3

[66] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[68] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18450–18459, 2022. 2

[69] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 2

[70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2, 3

[71] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, pages 465–476, 2017. 3