Dr.3D: Adapting 3D GANs to Artistic Drawings

Supplemental Document

Wonjoon Jin
POSTECH
South Korea
jinwj1996@postech.ac.kr

Nuri Ryu
POSTECH
South Korea
ryunuri@postech.ac.kr

Geonung Kim
POSTECH
South Korea
k2woong92@postech.ac.kr

Seung-Hwan Baek
POSTECH
South Korea
shwbaek@postech.ac.kr

Sunghyun Cho
POSTECH
South Korea
Pebblous
South Korea
s.cho@postech.ac.kr

ACM Reference Format:
Wonjoon Jin, Nuri Ryu, Geonung Kim, Seung-Hwan Baek, and Sunghyun Cho. 2022. Dr.3D: Adapting 3D GANs to Artistic Drawings: Supplemental Document. In *SIGGRAPH Asia 2022 Conference Papers (SA '22 Conference Papers), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3550469.3555422

1 OVERVIEW

In this Supplemental Document, we provide implementation details and additional qualitative results. Specifically, we present:

- pretraining details of EG3D,
- our detailed network architecture,
- additional analysis of our deformation-aware network,
- additional training details,
- domain adaptation of the compared methods,
- details on GAN inversion,
- visual examples of limitation, and
- additional results.

2 PRETRAINING OF EG3D

For domain adaptation, we pretrain an EG3D model [Chan et al. 2021a] on 25M real portrait images in the FFHQ dataset [Karras et al. 2019]. Since our primal interest lies in 3D-aware drawing synthesis, we reduce the network to generate 256×256 images for faster training/testing.

3 NETWORK ARCHITECTURE

Pose-estimation network. We use a pretrained ResNet50 network [He et al. 2016] as a backbone of our pose-estimation network. We replace the last fully-connected layer for classification with another fully-connected layer for pose estimation.

Deformation network. We modulate the 2D convolutional feature maps of the generator's coarse layers using our deformation network. We sample a deformation code $z_d \in R^{256}$ from the normal distribution. The deformation code is then converted into a feature $w_d \in R^{256}$ through our deformation network consisting of four fully-connected layers with 256 hidden layers. An affine-transformation layer then transforms w_d into 1D modulation code with the length of $HW + D$, where H, W and D are the height, width and number of channels of original 2D convolutional feature maps, respectively. We separate the modulation code into two tensors with the respective lengths of HW and D. We obtain residual feature maps with the size of $H \times W \times D$ from the two tensors with the size of $H \times W \times 1$ and $1 \times 1 \times D$ via broadcasting. We modulate the feature maps of the coarse layers at the resolutions of 8×8, 16×16, and 32×32 with feature dimension $D = 512$.

Image synthesis. Our StyleGAN-based generator [Karras et al. 2020b] synthesizes 2D convolutional features, which are fed to a volumetric renderer. Then, we obtain an image and a depth map of resolution of 64×64. High-resolution images with size 256×256 are then synthesized via a 2D convolutional upsampler.

Discriminator with pose condition. We use an estimated pose θ as a prior for the conditioned discriminator D following StyleGAN2-ADA [Karras et al. 2020a]. The 25-dimensional pose condition is composed of rearranged intrinsic and extrinsic matrices. An intrinsic matrix is constructed using a focal length, and an extrinsic matrix is obtained from the estimated camera pose θ. Then, this pose condition is fed to eight fully-connected layers and modulates the discriminator's features.

4 ADDITIONAL ANALYSIS ON DEFORMATION-AWARE NETWORK

The ablation study in Figure 7 in the main paper shows that the deformation-aware network helps improve the quality of synthesized geometries. It is because the deformation-aware network helps retain the original weights of the generator so that the knowledge about 3D structures of the original network can be better preserved for challenging styles such as anime [Anonymous et al. 2019]. Figure 1 visualizes the amounts of weight parameter changes in the feature generator caused by the adaptation process with and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SA '22 Conference Papers, December 6–9, 2022, Daegu, Republic of Korea
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9470-3/22/12…$15.00
https://doi.org/10.1145/3550469.3555422

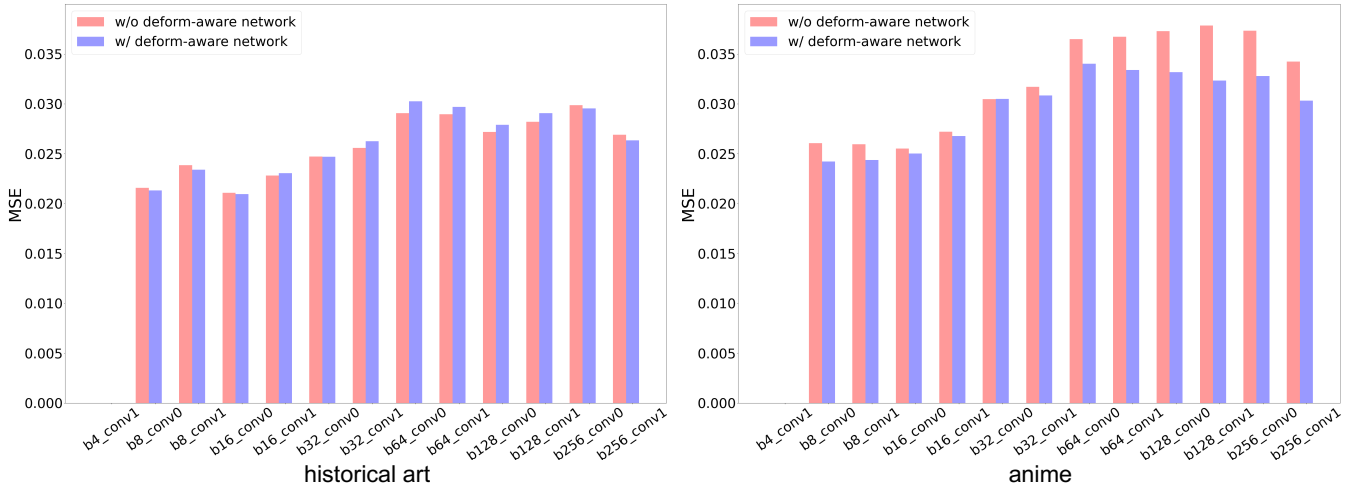

Figure 1: Amounts of weight parameter changes in the feature generator caused by the adaptation process with and without the deformation-aware network. The first convolutional layer (b4_conv0) is omitted because it is fixed for stable adaptation in both styles.

without the deformation-aware network. In the case of the historical art domain, which is closer to the original real human portrait domain than the anime domain, both adaptations with and without the deformation-aware network cause a similar amount of weight changes. On the other hand, in the case of anime domain, adaptation without the deformation-aware network causes large weight changes due to the large domain gap, which may lead to the lose of the knowledge learned in the original generator. In contrast, the deformation-aware network clearly helps suppress weight changes as shown in the figure.

5 TRAINING DETAILS

Adversarial Learning. We adapt the deformation-aware 3D synthesis network G from human faces to drawings via adversarial training. The pose-estimation network P estimates the pose θ of a real image x_{real}. A fake image x_{fake} is then synthesized by our generator G with a latent code z and a deformation code z_d at the camera pose θ. We train the deformation-aware 3D synthesis network by minimizing the discrepancy between fake and real image distributions with an adversarial loss $\mathcal{L}_a$ using the discriminator D:

$$\begin{aligned}\mathcal{L}_a(D,G) = \mathbb{E}_{z\sim p_z,\, z_d\sim p_{z_d}}[f(D(G(z,z_d,\theta)))] \\ + \mathbb{E}_{x_{real}\sim p_{data}}[\mathrm{f}(-D(x_{\text{real}},\theta)) + \lambda|\nabla D(x_{\text{real}},\theta))|^2],\end{aligned} \tag{1}$$

where p_{data} is the distribution of real images, $f(u) = -\log(1+\exp(-u))$, and $\lambda = 1.0$. The latent code z and deformation code z_d are sampled from normal distributions p_z and p_{z_d}. During the training of our synthesis network, we freeze the first convolutional layer of the feature generator for stable training.

Hyperparameters. In our experiments, we adapt 3D GAN from real portraits to the drawing datasets of historical art [Karras et al. 2019], ukiyo-e [Pinkney and Adler 2020], anime [Anonymous et al. 2019] and caricature [Huo et al. 2017]. For the weights (α, β, γ), we use $(3.0, 2.5, 10)$ for the historical-art dataset, $(2.0, 2.0, 10)$ for the ukiyo-e dataset, $(1.5, 1.0, 10)$ for the anime dataset, and $(3.0, 1.5, 10)$ for the caricature dataset. We run domain adaptation on 800K images for the historical-art and ukiyo-e datasets, and 400K images for the other datasets. We use learning rates of 0.00125 and 0.00075 for the deformation-aware 3D synthesis network and the discriminator. We set the learning rate as 0.0000125 for the mapping network and the deformation network. A learning rate of 0.0000075 is used for the pose-condition network of the discriminator. In our experiments, we use 8 NVIDIA RTX 3090 GPUs and the training takes 8.71 hours for metfaces and ukiyo-e. For the anime and caricature datasets, it took 4.36 hours for training.

6 DOMAIN ADAPTATION OF BASELINE METHODS

We compare Dr.3D with recent 3D GANs (π-GAN and StyleNeRF [Chan et al. 2021b; Gu et al. 2021]), a parametric fitting method (DECA [Feng et al. 2021]) and physics-based decomposition methods (Unsup3D, GAN2Shape [Pan et al. 2021; Wu et al. 2020]). We adapt all the models from real human portraits to drawings [Anonymous et al. 2019; Huo et al. 2017; Karras et al. 2020a; Pinkney 2020] except for DECA since a parametric model is not available for drawings.

π-GAN. We first retrain a pretrained model of π-GAN [Chan et al. 2021b] on FFHQ [Karras et al. 2019] with 7.2M images, improving its synthesis capability. We then perform domain adaptation of the retrained π-GAN model to drawings using 360K, 480K, 720K, and 600K iterations for historical-art, ukiyo-e, caricature, and anime, until the model converges.

StyleNeRF. For StyleNeRF [Gu et al. 2021], we also use the author-provided model pretrained on FFHQ. We adapt this model to drawings with 400K, 800K, 600K and 800K iterations for historical-art, ukiyo-e, caricature, and anime until the model converges.

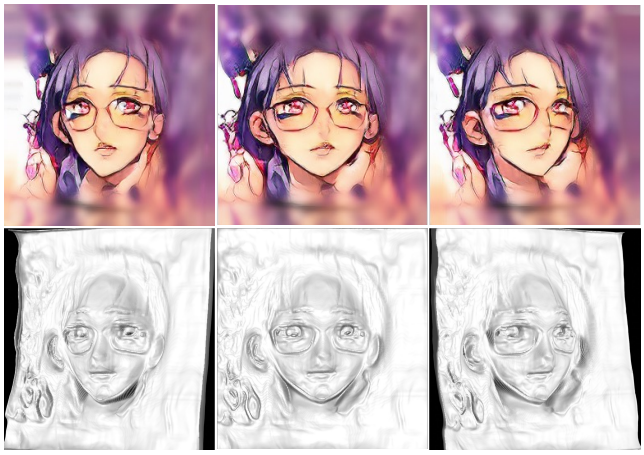

Figure 2: Limitations of Dr.3D. Dr.3D may sometimes produce flattened geometries for some latent codes of challenging target domains such as anime.

Unsup3D. We adapt an Unsup3D [Wu et al. 2020] model pretrained on CelebA [Liu et al. 2015] to drawings. We trained Unsup3D with 90K, 300K, 300K and 90K training iterations for historical-art, ukiyo-e, caricature, and anime until the model converges.

GAN2Shape. Since GAN2Shape [Pan et al. 2021] needs a pretrained StyleGAN2 model, we were only able to adapt its model to the historical-art dataset. The adaptation was done on a model pretrained with CelebA [Liu et al. 2015].

7 DETAILS ON GAN INVERSION

Dr.3D. We first use an off-the-shelf face detector [King 2009] to find the face region in a target drawing. Then, we align the drawing and crop it to the resolution of 256×256. We then use pivotal tuning inversion (PTI) [Roich et al. 2021] to find a latent code w, pose θ, and deformation code z_d that reconstruct the input image best using our generator. We first run the optimization for 500 iterations. For additional 350 iterations, we fix the latent code w and only optimize the deformation code z_d by finetuning our deformation-aware 3D synthesis network G.

π-GAN. π-GAN [Chan et al. 2021b] uses FiLM-SiREN structure which has the parameters of frequency and phase-shift. To reconstruct an input drawing using π-GAN, we optimize frequency, phase-shift, and rendering pose for 1000 iterations. As the rendering pose of the drawing is not available in the π-GAN setting, we use the estimated pose from our pose-estimation network for initialization. Note that the author-provided inversion code did not work in our experiments.

StyleNeRF. Since StyleNeRF [Gu et al. 2021] adopts similar architecture to StyleGAN2 [Karras et al. 2020b], we use the same inversion method of PTI [Roich et al. 2021]. We set the initial pose using our pose-estimation network. Latent code and rendering pose are optimized for 500 iterations, followed by generator tuning for additional 350 iterations.

8 LIMITATION EXAMPLES

Although Dr.3D outperforms previous methods, it may sometimes produce degenerate flattened geometries for some latent codes for challenging domains such as anime characters. Figure 2 shows such an example.

9 ADDITIONAL RESULTS

Here, we provide additional results comparing Dr.3D with other baseline methods. Figure 3 shows reconstructed images and shapes for the metfaces dataset [Karras et al. 2019] using DECA [Feng et al. 2021], Unsup3D [Wu et al. 2020], GAN2Shape [Pan et al. 2021], π-GAN [Chan et al. 2021b], styleNeRF [Gu et al. 2021] and Dr.3D. We extract the 3D shapes using the marching-cube algorithm. Figure 4 shows another comparison on the ukiyo-e, caricature, anime datasets [Anonymous et al. 2019; Huo et al. 2017; Pinkney 2020]. DECA fails to synthesize realistic shapes since it requires a parametric model which is not available for drawings. π-GAN and StyleNeRF suffer from distortion artifacts due to the lack of 3D knowledge of drawings. Our method outperforms the other methods in image fidelity and shape quality.

Figure 5 presents a qualitative comparison of randomly synthesized images of π-GAN [Chan et al. 2021b], StyleNeRF [Gu et al. 2021] and Dr.3D. For all drawing styles, Dr.3D shows the best synthesized results with clear and high-quality images.

Multi-view images with variable yaw angles can be synthesized by Dr.3D with fixed latent code z and deformation code z_d. Figures 7 and 6 show multi-view drawing synthesis results of 3D GAN baselines [Chan et al. 2021b; Gu et al. 2021] and Dr.3D. π-GAN results in low-quality images at steep angles. StyleNeRF suffers from flattened shapes. Dr.3D effectively covers the large shape deformation of drawings by expanding representation power using our deformation-aware 3D synthesis network.

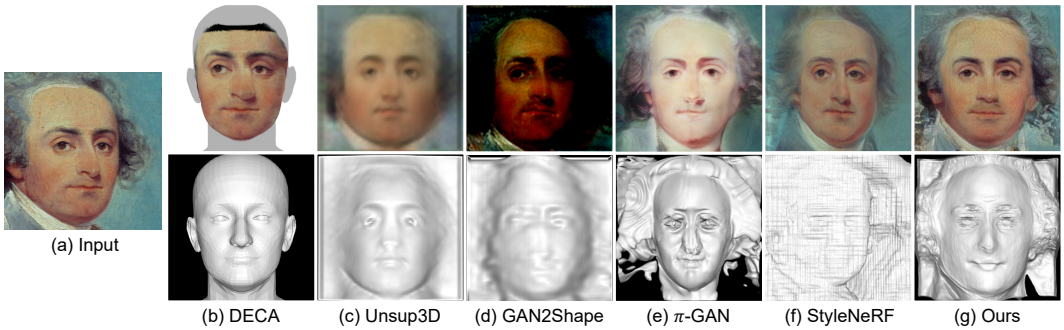

Figure 3: Qualitative comparison of reconstruction results. Input: Portrait of Giuseppe Ceracchi, 1792 by John Trumbull, Yale University Art Gallery [Public Domain] via (https://bit.ly/3KbMRjc).

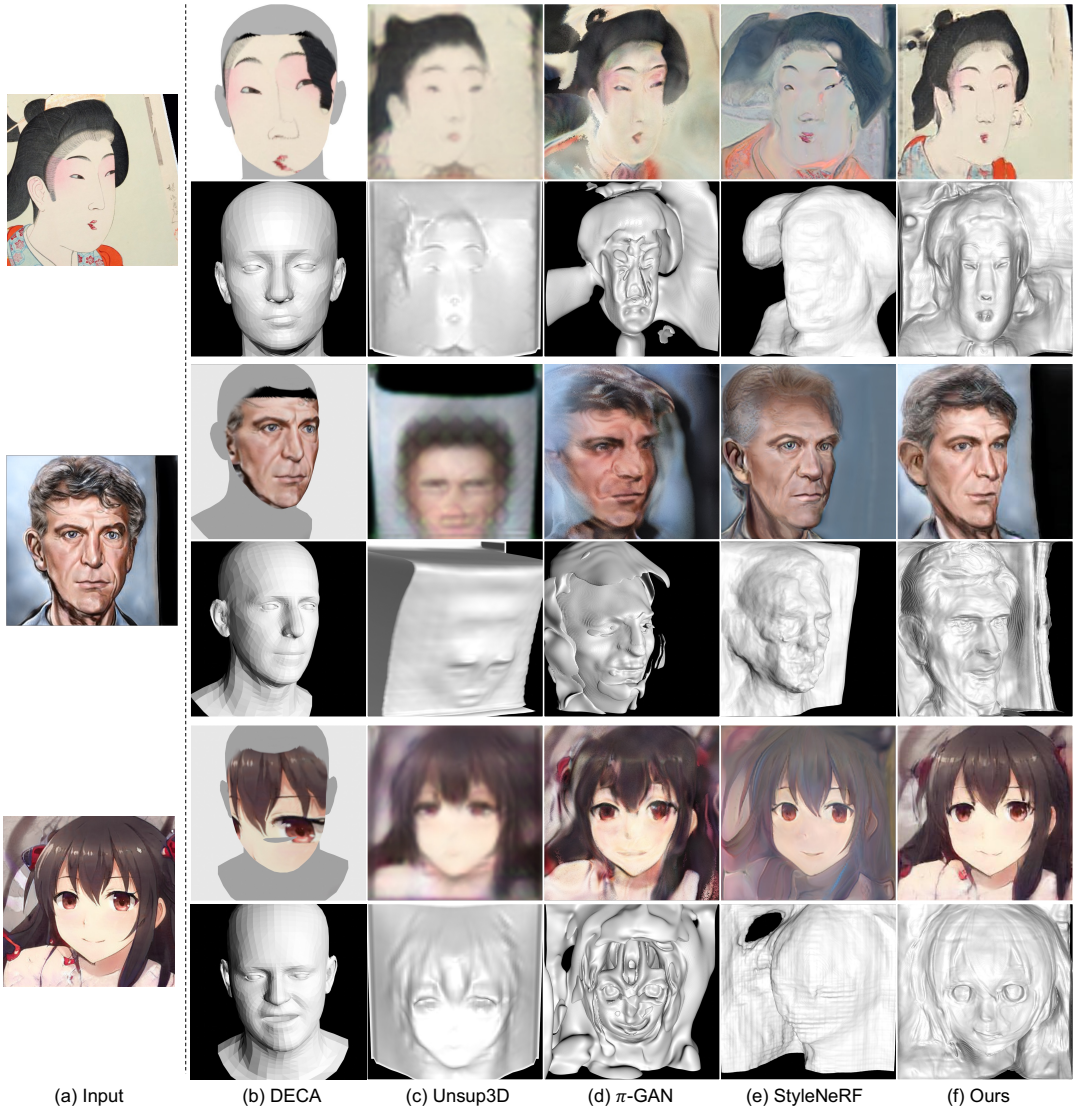

Figure 4: Qualitative comparison of reconstruction results. Input (top): A Beauty of The Kyoho Era, 1897 by Toyohara Chikanobu, Arthur M. Sackler Gallery [Fair Use] via (https://s.si.edu/3ewrtth).

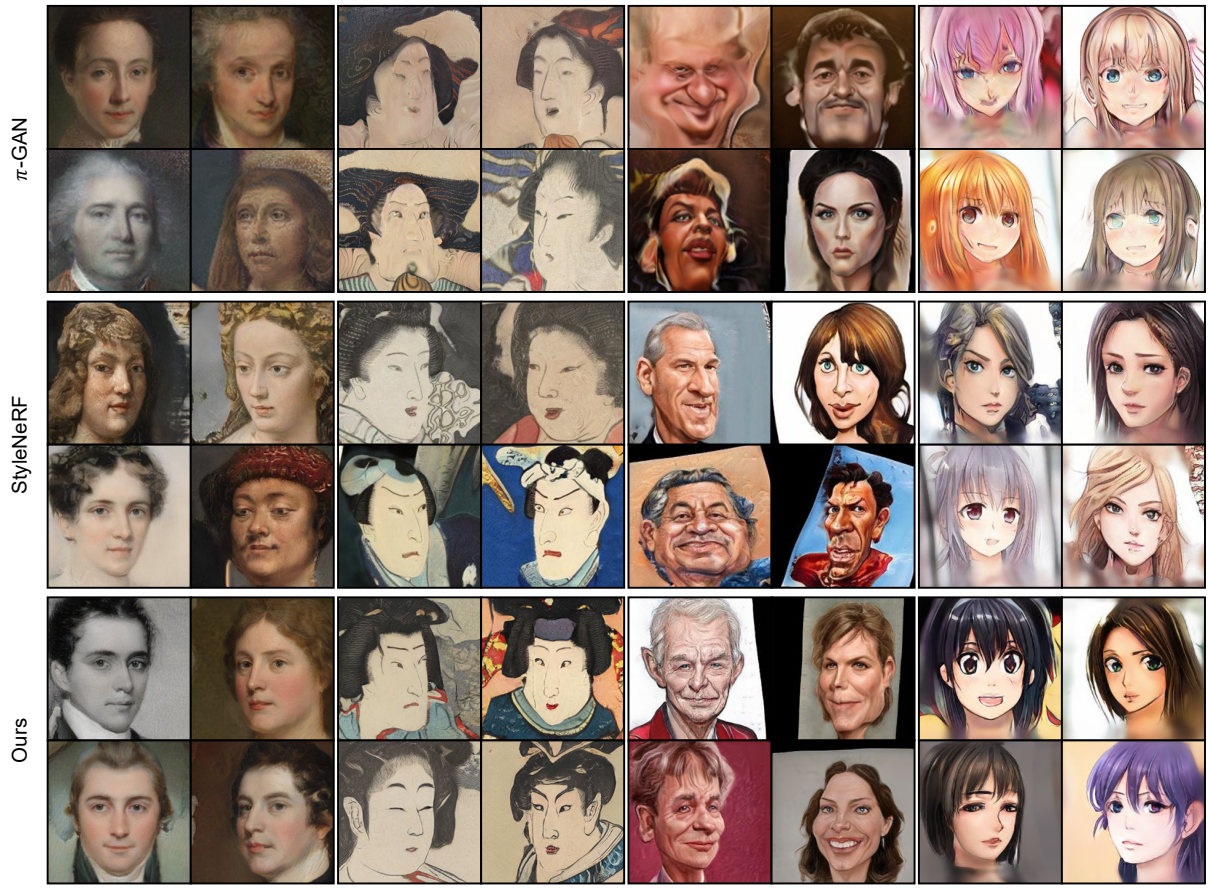

Figure 5: Qualitative comparison of randomly synthesized images.

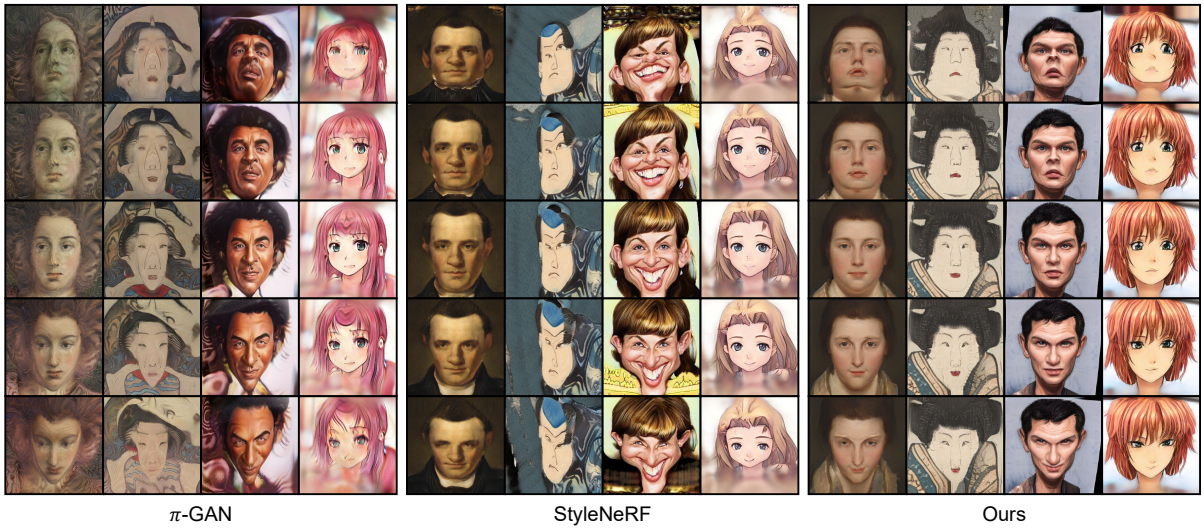

Figure 6: Drawing synthesis with different pitch angles. Dr.3D can synthesize higher-quality multi-view consistent drawings than the other 3D GAN baselines.

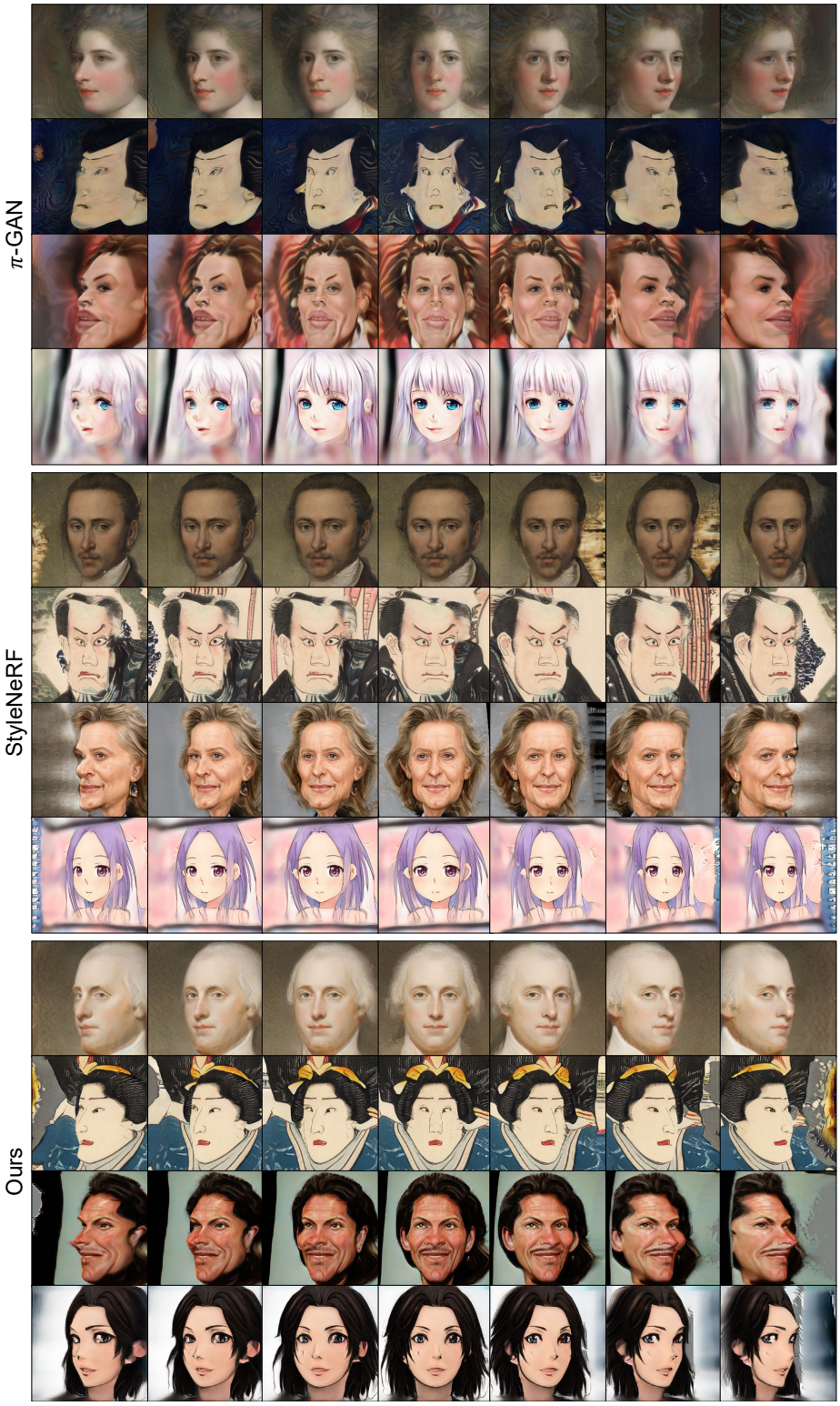

Figure 7: Drawing synthesis with different yaw angles. Dr.3D can synthesize higher-quality multi-view consistent drawings than the other 3D GAN baselines.

REFERENCES

Anonymous, the Danbooru community, Gwern Branwen, and Aaron Gokaslan. 2019. Danbooru2018: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset. https://www.gwern.net/Danbooru2018. https://www.gwern.net/Danbooru2018 Accessed: DATE.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2021a. Efficient geometry-aware 3D generative adversarial networks. *arXiv preprint arXiv:2112.07945* (2021).

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021b. PI-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *Proc. of IEEE/CVF CVPR*. 5799–5809.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. on Graphics (TOG)* 40, 4 (2021), 1–13.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985* (2021).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of IEEE/CVF CVPR*. 770–778.

Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. 2017. WebCaricature: a benchmark for caricature recognition. *arXiv preprint arXiv:1703.03230* (2017).

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020a. Training generative adversarial networks with limited data. In *Proc. of NeurIPS*. 12104–12114.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proc. of IEEE/CVF CVPR*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020b. Analyzing and improving the image quality of StyleGAN. In *Proc. of IEEE/CVF CVPR*. 8110–8119.

Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proc. of IEEE/CVF ICCV*.

Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. 2021. Do 2D GANs Know 3D Shape? Unsupervised 3D Shape Reconstruction from 2D Image GANs. In *Proc. of ICLR*.

Justin NM Pinkney and Doron Adler. 2020. Resolution dependent GAN interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334* (2020).

Justin N. M. Pinkney. 2020. Aligned Ukiyo-e faces dataset. https://www.justinpinkney.com/ukiyoe-dataset.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2021. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744* (2021).

Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. 2020. Unsupervised Learning of Probably Symmetric Deformable 3D Objects From Images in the Wild. In *Proc. of IEEE/CVF CVPR*.