PODIA-3D: Domain Adaptation of 3D Generative Model Across Large Domain Gap Using Pose-Preserved Text-to-Image Diffusion

Gwanghyun Kim[1] Ji Ha Jang[1] Se Young Chun[1,2,†]
[1]Dept. of Electrical and Computer Engineering, [2]INMC & IPAI
Seoul National University, Republic of Korea
{gwang.kim, jeeit17, sychun}@snu.ac.kr

Figure 1. Our PODIA-3D successfully adapts 3D generators across significant domain gaps, producing excellent text-image correspondence and 3D shapes, while the baselines fail. See the supplementary videos at gwang-kim.github.io/podia_3d.

Abstract

Recently, significant advancements have been made in 3D generative models, however training these models across diverse domains is challenging and requires an huge amount of training data and knowledge of pose distribution. Text-guided domain adaptation methods have allowed the generator to be adapted to the target domains using text prompts, thereby obviating the need for assembling numerous data. Recently, DATID-3D presents impressive quality of samples in text-guided domain, preserving diversity in text by leveraging text-to-image diffusion. However, adapting 3D generators to domains with significant domain gaps from

†Corresponding author.

the source domain still remains challenging due to issues in current text-to-image diffusion models as following: 1) shape-pose trade-off in diffusion-based translation, 2) pose bias, and 3) instance bias in the target domain, resulting in inferior 3D shapes, low text-image correspondence, and low intra-domain diversity in the generated samples. To address these issues, we propose a novel pipeline called PODIA-3D, which uses pose-preserved text-to-image diffusion-based domain adaptation for 3D generative models. We construct a pose-preserved text-to-image diffusion model that allows the use of extremely high-level noise for significant domain changes. We also propose specialized-to-general sampling strategies to improve the details of the generated samples. Moreover, to overcome the instance bias, we introduce a text-guided debiasing method that improves intra-domain diversity. Consequently, our method successfully adapts 3D generators across significant domain gaps. Our qualitative results and user study demonstrate that our approach outperforms existing 3D text-guided domain adaptation methods in terms of text-image correspondence, realism, diversity of rendered images, and sense of depth of 3D shapes in the generated samples.

1. Introduction

Recently, 3D generative models [20, 37, 8, 11, 23, 24, 40, 38, 5, 23, 25, 33, 10, 39, 4, 1] have been advanced to enable multi-view consistent and explicitly pose-controlled image synthesis. However, training state-of-the-art 3D generative models is challenging due to the requirement of a large number of images and knowledge about their camera pose distribution. This prerequisite has resulted in limited applications of these models to only a few domains.

Text-guided domain adaptation methods such as StyleGAN-NADA [9], HyperDomainNet [2], DATID-3D [18], and StyleGANFusion [35] have emerged as a promising solution to overcome the challenge of need for additional data of the target domain. These methods leverage CLIP [27] or text-to-image diffusion models [31, 28, 32] that are pretrained on a large number of image-text pairs.

Although non-adversarial fine-tuning methods like StyleGAN-NADA [9], HyperDomainNet [2], and StyleGANFusion [35] have demonstrated impressive results, they suffer from the inherent loss of diversity in a text prompt and suboptimal text-image correspondence, as illustrated in Fig. 1 (See results of StyleGAN-NADA*). Recently, a diversity-preserved domain adaptation method called DATID-3D [18] has been developed for 3D generators, which achieves compelling quality of multi-view consistent image synthesis in text-guided domains. This method generates pose-aware target dataset using text-to-image diffusion models and fine-tunes the 3D generator on the target dataset.

Despite the use of this method, the adaptation of 3D generators to domains that have significant domain gaps from the source domain remains challenging due to the problems encountered in the current text-to-image diffusion models. 1) shape-pose trade-off in diffusion-based translation: For text-guided pose-aware target generation, we first perturb the source image or latent $\boldsymbol{x}_0^{\text{src}}$ until $t_r \in [1, T]$ such that $\boldsymbol{x}_0^{\text{trg}}$ generated from $\boldsymbol{x}_{t_r}^{\text{src}}$ should represent the features corresponding the target domains without altering pose of $\boldsymbol{x}_0^{\text{src}}$. However, our investigations show that when the target domain requires selecting a high t_r to achieve a significant structural change, preserving the pose is not guaranteed, as depicted in Fig. 2(a). Consequently, shifting the generator to a target domain that requires significant shape changes can lead to poor 3D shapes or low text correspondence, as illustrated in Fig. 1 (See SpongeBob by DATID-3D [18]). 2) We found that a publicly available text-to-image diffusion model, has pose bias issues for certain target domain text prompts, as illustrated in Fig. 2(b). Pose bias represents that the position and orientation of certain objects in images from T2I diffusion models are biased (mainly toward the front or side). Accordingly, the shifted generators guided by these text prompts result in either poor 3D structure as represented in Fig. 1 (See Horse by DATID-3D [18]). 3) We also found that the text-to-image diffusion models often generate images with one or a few instances among many instances representing the text prompts as represented in Fig. 2. In consequence, the shifted generators guided by these text prompts result in low intra-domain diversity as represented in Fig. 1. (See Dog by DATID-3D [18]).

To address these issues, we propose a novel pipeline called PODIA-3D, a method of **PO**se-preserved text-to-image **DI**ffusion-based domain **A**daptation for **3D** generative model. We construct pose-preserved text-to-image diffusion models. We first collect target images that have the same pose but different shapes with source images through 3 strategies: identity mixing, text-guided image translation with pose-guaranteed prompts, utilizing the different domain generator. Then, we fine-tune the depth-guided diffusion model to make it ignore the shape information from the depth map and focus only on pose information. Furthermore, we propose a specialized-to-general sampling strategy to improve details of generated images and resolve the detail bias issue. Using pose-preserved diffusion models and specialized-to-general sampling, we are able to synthesize pose-consistent target images with excellent text-image correspondence by using extremely high-level noise for large shape change. We then fine-tune the state-of-the-art 3D generator adversarially on the generated target images. Moreover, to improve intra-domain diversity, we propose a text-guided debiasing method, which enables the fine-tuned generator to reach the diverse modes. As a result, our method effectively adapts 3D generators across significant domain gaps, generating excellent text-image correspondence and 3D shapes, as shown in Fig. 1. Our approach has been demonstrated to outperform

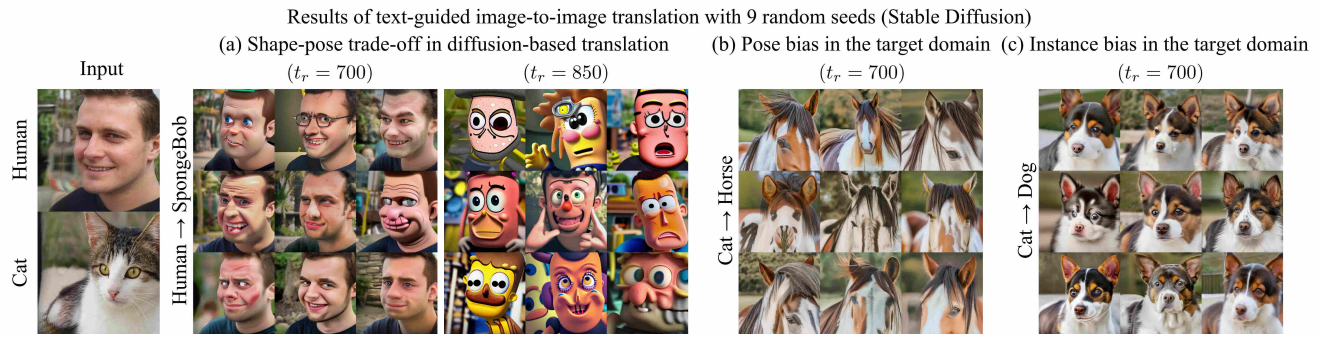

Figure 2. Issues in pose-aware target generation for domain adaption of 3D generative models using current text-to-image diffusion models: (a) a shape-pose trade-off in diffusion-based translation, (b) pose bias, and (c) instance bias in the target domain.

existing 3D text-guided domain adaptation methods in terms of text-image correspondence, realism, diversity of rendered images, and sense of depth of 3D shapes in the generated samples via the qualitative results and user study.

2. Related Works

2.1. 3D generative models

Recent advancements in 3D generative models [20, 37, 8, 11, 23, 24, 40, 38, 5, 23, 25, 33, 10, 39, 4, 1] have enabled multi-view consistent and explicitly pose-controlled image synthesis. Notably, EG3D [4], which uses StyleGAN2 [17] generator, in conjunction with neural rendering [22], succeeds in producing high resolution multi-view consistent images in real-time as well as highly detailed 3D shapes. However, training modern 3D generative models is more challenging than training 2D generative models, as it requires a significant number of images and detailed information on the camera parameter distribution for those images.

To expand the usability of state-of-the-art 3D generative models to a wider range of domains, including those with significant domain gaps, we introduce PODIA-3D, an approach that employs text-guided adaptation methods for 3D generators using pose-preserved diffusion models to enhance image-text correspondence, 3D shapes, and intra-domain diversity.

2.2. Text-to-image diffusion models

Diffusion models have demonstrated great success in the fields of image generation [12, 34, 36, 13, 7] and image-text multimodal applications [31, 28, 32, 19, 3]. In recent years, text-to-image diffusion models trained on large-scale image-text datasets [31, 28, 32] have exhibited remarkable performance in generating diverse 2D images from a single text prompt. One variant of text-to-image diffusion models referred to as depth-guided diffusion models [31], employs depth maps as a conditioning input throughout the generative process to synthesize images that correspond to the provided depth map.

In this work, we propose the pose-preserved text-to-image diffusion model to generate faithful pose-consistent target images to adapt the 3D generator to text-guided domains with large domain gaps.

2.3. Text-guided domain adaptation

Domain adaptation methods guided by textual prompts have been developed for 2D generative models, providing a promising solution to the challenge of acquiring additional data for the target domain. These methods utilize CLIP [27] or text-to-image diffusion models [31, 28, 32] pretrained on a large number of image-text pairs, allowing for text-driven domain adaptation. StyleGAN-NADA [9] and HyperDomainNet [2] fine-tune pretrained StyleGAN2 [17] models to shift the domain towards a target domain, utilizing a simple textual prompt guided by CLIP [27] loss. StyleGANFusion [35] adopts SDS loss [26] as guidance of text-guided adaptation of 2D and 3D generators using text-to-image diffusion models. Although these non-adversarial fine-tuning methods have demonstrated impressive results, they suffer from the inherent loss of diversity in a text prompt and suboptimal text-image correspondence. DATID-3D [18] achieves impressive quality in multi-view consistent image synthesis for text-guided domains by generating diverse pose-aware target dataset using text-to-image diffusion models and fine-tuning the 3D generator on the target dataset while preserving diversity in the text. However, adapting 3D generators to domains with significant domain gaps from the source domain using existing methods remains challenging. This is because these models suffer from several issues, such as a shape-pose trade-off in diffusion-based translation, pose bias, and instance bias in the target domain. As a result, the generated samples often exhibit inferior 3D shapes, low text-image correspondence, and low intra-domain diversity.

To mitigate these issues, we propose PODIA-3D, a novel method of pose-preserved text-to-image diffusion-based domain adaptation for the 3D generative models.

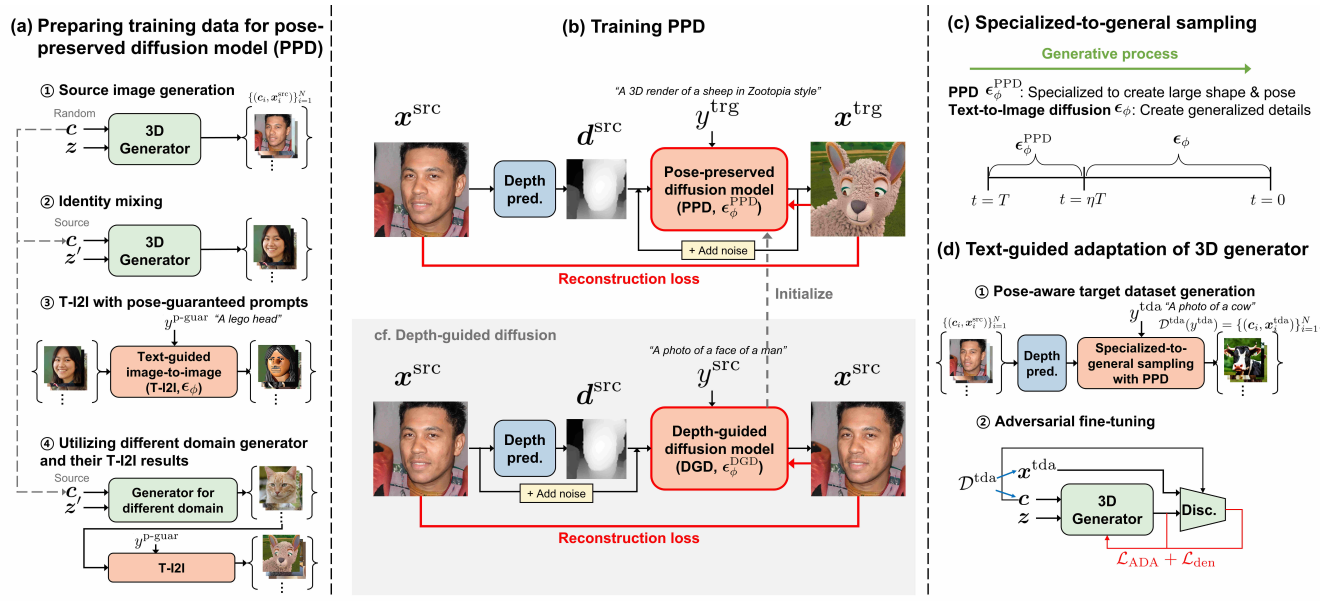

Figure 3. Overview of PODIA-3D. (a) We prepare data for training pose-preserved diffusion models (PPD) and (b) fine-tune the depth-guided diffusion models on the collected data. (c) We use a specialized-to-general sampling strategies to generate high quality pose-aware target images. (d) Finally, we fine-tune the state-of-the-art 3D generator on them adversarially.

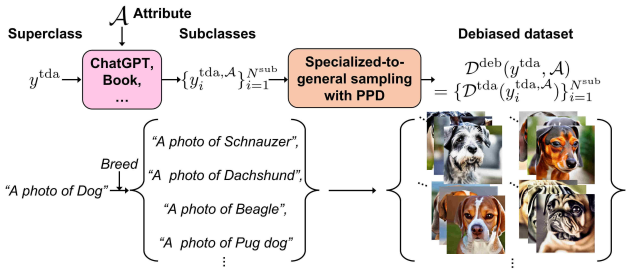

Figure 4. Our text-guided debiasing method includes obtaining a set of subclass texts, and then generating a pose-aware target dataset for each subclass text. We combine these datasets to construct a debiased target dataset.

3. PODIA-3D

We begin by constructing a text-to-image diffusion model that preserves the pose of the source images while generating target images, as shown in Fig. 3(a)(b). We then propose a specialized sampling strategy to enhance the pose-preserved diffusion and improve image details, as demonstrated in Fig. 3(c). Using this approach, we perform pose-preserved diffusion-driven text-guided domain adaptation in two steps: 1) generating a pose-aware target dataset and 2) fine-tuning the 3D generator using adversarial training, as depicted in Fig. 3(d). Furthermore, to address the instance bias observed in certain text prompts, we introduce a text-guided debiasing method, as illustrated in Fig. 4, to improve intra-domain diversity.

3.1. Pose-preserved text-to-image diffusion models

We aim to synthesize target images that have faithful pose-consistency, high diversity, and excellent text-image correspondence. To achieve this, we initially consider using the depth-guided diffusion model (DGD) [31], which generates images conditioned on both depth maps and text prompts, thus producing images consistent with the given depth maps. However, as shown in Fig. 8, we observed that the strong shape constraints imposed by the depth maps can result in low diversity and poor text-image correspondence, particularly for text prompts that require significant shape changes. To overcome this issue while retaining the benefits of DGD, we develop pose-preserved text-to-image diffusion models (PPD) by fine-tuning DGD to focus only on pose information and ignore shape information during image generation.

Preparation of training data for PPD. We prepare training data $\mathcal{D}^{\mathrm{PPD}} = \{(\boldsymbol{d}_i^{\mathrm{src}}, \boldsymbol{q}_i^{\mathrm{trg}}, y^{\mathrm{trg}})\}_{i=1}^{N^{\mathrm{PPD}}}$ for training PPD, which consists of a source depth map $\boldsymbol{d}_i^{\mathrm{src}}$, target diffusion latent $\boldsymbol{q}_i^{\mathrm{trg}} = E^V(\boldsymbol{x}_i^{\mathrm{trg}})$ from target image $\boldsymbol{x}_i^{\mathrm{trg}}$ encoded by VQGAN encoder E^V, and target text prompt y^{trg} following the process illustrated in Fig. 3(a). We start by generating N^{src} source images $\boldsymbol{x}^{\mathrm{src}} = G_\theta(\boldsymbol{z}, \boldsymbol{c})$ with random latent vectors $\boldsymbol{z}$ and camera parameters $\boldsymbol{c}$ given the pretrained source 3D generator G_θ, which in our case is the EG3D [4] model trained on 512^2 FFHQ [16] images. Next, we obtain the source depth maps $\boldsymbol{d}^{\mathrm{src}}$ using a pretrained depth estimation model. To collect the set of target images $\boldsymbol{x}^{\mathrm{trg}}$ with the

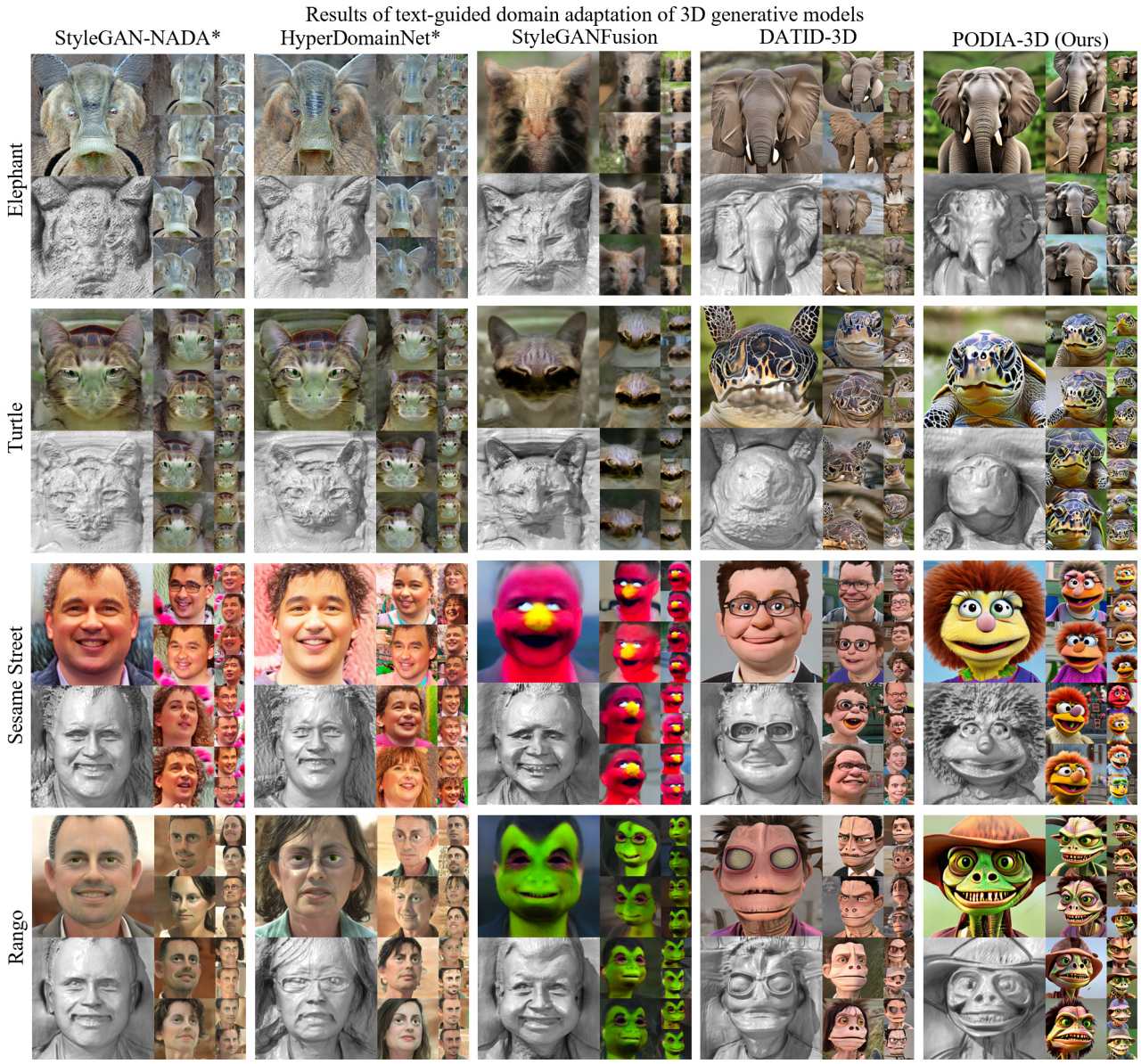

Figure 5. Qualitative comparison with existing text-guided domain adaptation methods with a star (*) indicating their 3D extensions. Our method allows to adapt the 3D generative models to domains with huge domain gap, presenting excellent text-image correspondence and 3D shape.

same pose as the source images but different shapes, we employ the following three strategies: 1) Identity mixing: We generate images by feeding the same camera parameters $\boldsymbol{c}$ with the source images into G but with different latent vectors $\boldsymbol{z}$. The prompts for y^{trg} are chosen to represent the source domain. 2) Text-guided image-to-image translation (T-I2I) [21] with pose-guaranteed prompts $y^{\text{p-guar}}$: We use the pretrained text-to-image model (Stable diffusion [31]) to perform T-I2I on the identity-mixed images for each prompt with guaranteed pose consistency and excellent text-image correspondence, based on our observation. We carefully select the text prompts to avoid overlapping visual features, mitigating bias issues. 3) Using a different domain generator: To achieve further large shape changes, we use the generator trained for a different domain. Specifically, we use the EG3D [4] model pretrained on AFHQ-cat [15, 6] dataset, which is transferred from the FFHQ EG3D model. We generate images to have the same pose as the source images with this model and also use the translated images using T-I2I.

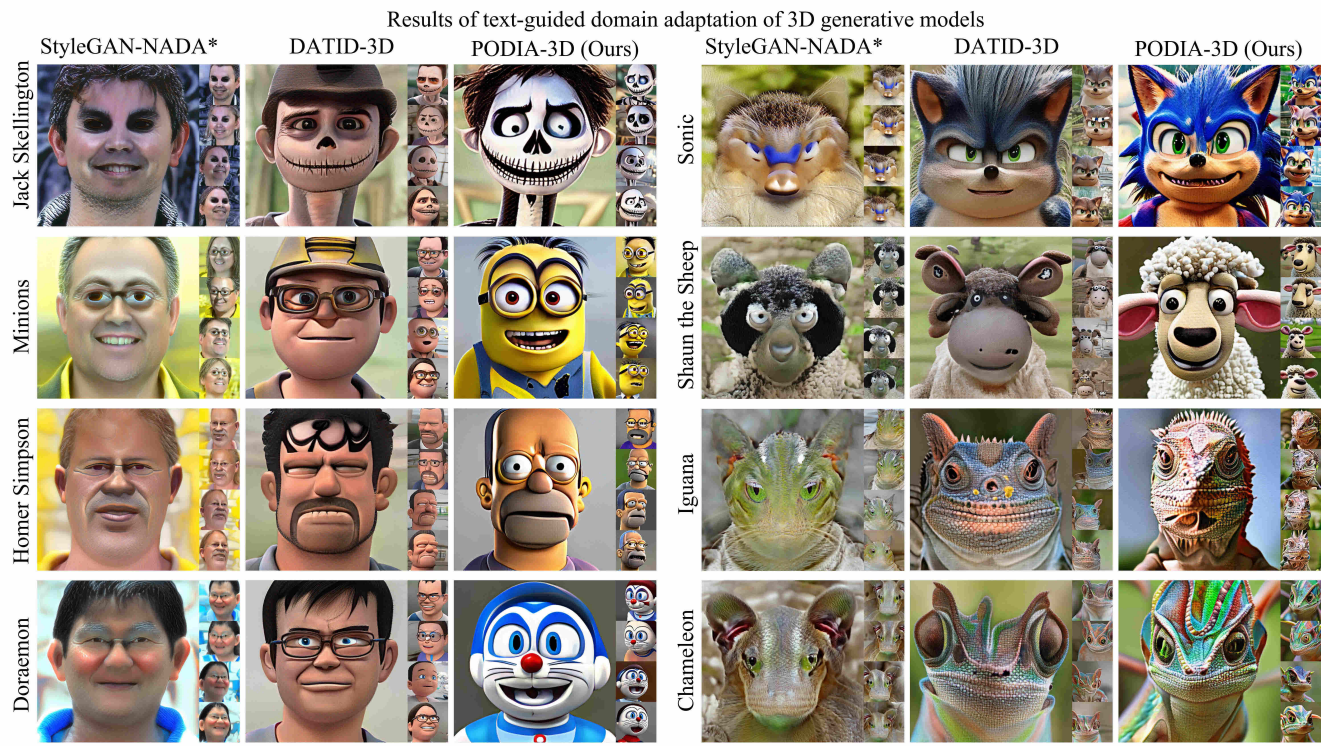

Figure 6. Our method succeeds in text-guided adaption to the wide range of domains while other baselines show the results with low text-image correspondence. For extended results, see the supplementary Fig. S2 and S3.

Fine-tuning objective for PPD. We fine-tune the copy of pretrained DGD $\epsilon_\phi^{\text{PPD}}$ on $\mathcal{D}^{\text{PPD}}$ using following objective:

$$\mathbb{E}_{(\boldsymbol{d}^{\text{src}},\boldsymbol{q}^{\text{trg}},y^{\text{trg}})\in\mathcal{D}^{\text{PPD}},\boldsymbol{\epsilon},t}[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\phi^{\text{PPD}}(\boldsymbol{q}_t^{\text{trg}},y^{\text{trg}},\boldsymbol{d}^{\text{src}},t)\|_2^2],$$

where $\boldsymbol{\epsilon}\sim\mathcal{N}(0,1), t\sim\mathcal{U}([1,T])$.

3.2. Specialized-to-general sampling

Although PPD trained on augmented data can generate images with corresponding pose and shape to the depth map and text prompt, we discovered the presence of style and detail biases that are inherent in the training data. To address this issue and enhance details, we propose specialized-to-general sampling strategies that leverage the pose-consistent generation capability of the PPD model and the generalization capability of text-to-image diffusion models as presented in Fig. 3(c). During the first ηT period, where $\eta\in[0,1]$ is a PPD ratio and T is the number of total diffusion steps, we use the PPD model to generate large structural components and pose information. For the remaining $(1-\eta)T$ period, we utilize Stable diffusion [31], the general text-to-image diffusion model, to generate small structures or details in the images.

3.3. Adapting 3D generator to broader domains

As illustrated in Fig. 3(d), we translate the source image $\boldsymbol{x}^{\text{src}}$ to yield the target image $\boldsymbol{x}^{\text{tda}}$ guided by a text prompt y^{tda} using PPD and specialized-to-general sampling, constructing the pose-aware target dataset $\mathcal{D}^{\text{tda}}(y^{\text{tda}})=\{(\boldsymbol{c}_i,\boldsymbol{x}_i^{\text{tda}})\}_{i=1}^N$. Then, we fine-tune 3D generator adversarially using the loss composed of ADA loss $\mathcal{L}_{\text{ADA}}$ [14] and density regularization loss $\mathcal{L}_{\text{den}}$, following EG3D [4] and DATID-3D [18].

PPD and specialized-to-general sampling not only enable us to produce pose-consistent target images, but also improve their text-image correspondence by leveraging the full expressiveness of the text-to-image diffusion model through the use of extremely high return steps. This approach enables us to adapt 3D generators to domains with significant domain gaps without the need for time-consuming CLIP- and pose reconstruction-based filtering processes in DATID-3D [18], making the pipeline more efficient and simplified.

3.4. Text-guided debiasing

We observe that text-to-image diffusion models often suffer from an instance bias issue where only a few instances representing the text prompts are generated in the images. However, when we specify subclasses (e.g. breeds of dog) of the objects represented by the text prompt, the images of the instance are synthesized well. Based on this observation, we propose a text-guided debiasing method to improve intra-domain diversity, as depicted in Fig. 4.

To debias the target domain $\mathcal{X}^{\text{tda}}$ represented by the text y^{tda} in terms of attribute $\mathcal{A}$, we first obtain a set $\{y_i^{\text{tda},\mathcal{A}}\}_{i=1}^{N^{\text{sub}}}$

Figure 7. Our text-guided debiasing method improves intra-domain diversity of the results of text-guided domain adaptation.

of N^{sub} subclass texts $y_i^{\text{tda},\mathcal{A}}$ from various sources such as books, web search, or AI-powered chatbots like ChatGPT. We can ask ChatGPT "Tell me the list of *Breeds* of *Dogs*." when $\mathcal{X}^{\text{tda}} = \textit{Dog}$ and $\mathcal{A} = \textit{Breed}$. Then, we generate pose-aware target dataset $\mathcal{D}(y_i^{\text{tda},\mathcal{A}})$ for each subclass text. Finally, combining these dataset, we can construct the debiased dataset $\mathcal{D}^{\text{deb}}(y^{\text{tda}}, \mathcal{A}) = \{\mathcal{D}(y_i^{\text{tda},\mathcal{A}})\}_{i=1}^{N^{\text{sub}}}$.

4. Experiments

We demonstrate the effectiveness of our approach by applying it to a range of diverse domains with significant domain gaps using state-of-the-art 3D generators, EG3D [4]. For the experiments, we employ a Stable diffusion and its variants, depth-guided diffusion [31]. We use MiDaS [29, 30] as our depth map estimation model. To fine-tune the 3D generators, 1,000 target images per text prompt are used. We set $\eta = 0.4$ as default. In case of text-guided debiased dataset, we use 300 images per subclass text. For more detailed information about the setup of experiments, see the supplementary Section C and D.

Table 1. User study results on text-image correspondence, realism and diversity of rendered images from adapted generators.

Rendered 2D images	Text-Corr.↑	Realism↑	Diversity↑
StyleGAN-NADA*	3.267	2.571	2.719
HyperDomainNet*	3.231	2.576	2.776
StyleGANFusion	3.502	2.812	2.871
DATID-3D	3.776	3.148	3.160
Ours	**4.071**	**3.455**	**3.426**

Table 2. User study results on text-image correspondence, sense of depth and details of 3D shape extracted from adapted generators.

3D shapes	Text-Corr.↑	Sense of depth & Details↑
StyleGAN-NADA*	2.707	2.779
HyperDomainNet*	2.688	2.802
StyleGANFusion	2.860	2.981
DATID-3D	3.214	3.260
Ours	**3.495**	**3.440**

4.1. Evaluation

Baselines. We compare our approach to several recent methods for domain adaptation in 3D generative models, including StyleGANFusion [35] and DATID-3D [18], both of which are based on text-to-image diffusion methods. We also compare our approach to CLIP-based methods for 2D generative models, StyleGAN-NADA [27] and HyperDomainNet [2], denoted by a star symbol (*) to indicate their extension to 3D models.

To evaluate our method, we use the EG3D [4] generator pretrained on 512^2 images from the FFHQ dataset [16] as our source generator, and adapt it to a range of diverse domains with significant domain gaps. In contrast, other methods used the EG3D generator pretrained on 512^2 FFHQ images for adaptation to movie or animation characters, and the EG3D generator pretrained on 512^2 AFHQ-cat images for adaptation to animal domains, following their original experimental settings.

Qualitative results. As shown in Fig. 1 and 5, Our method successfully adapts 3D generators to the domains with significant domain gaps, enabling the synthesis of diverse samples with excellent text-image correspondence and 3D shapes. In contrast, other methods fail to adapt to these domains. For instance, when the target domain is an elephant, the samples and 3D shapes generated by StyleGAN-NADA* [27], HyperDomainNet*[2], and StyleGANFusion[35] resemble cats more than elephants. Although DATID-3D succeeds in generating samples that resemble elephants, its pose is biased toward the front view, leading to poor quality of 3D shapes. In comparison, our method produces images that closely correspond to elephant images with detailed shapes.

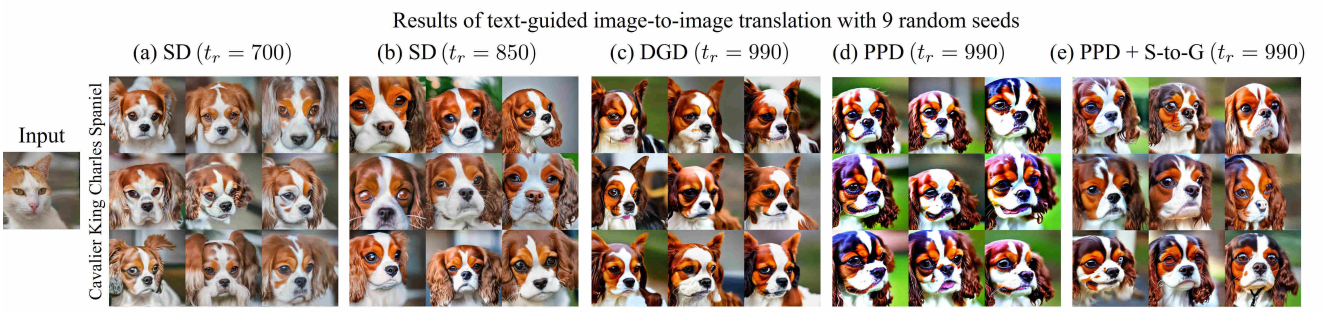

Figure 8. Results of text-guided image-to-image translation using Stable diffusion (SD) [31], depth-guided diffusion (DGD) [31], our pose-preserved diffusion (PPD), and specialized-to-general (S-to-G). Pose-preserved diffusion enables image translation with pose-consistency and domain adaptation with high-quality of 3D shapes. S-to-G allows to resolve the bias issue in details.

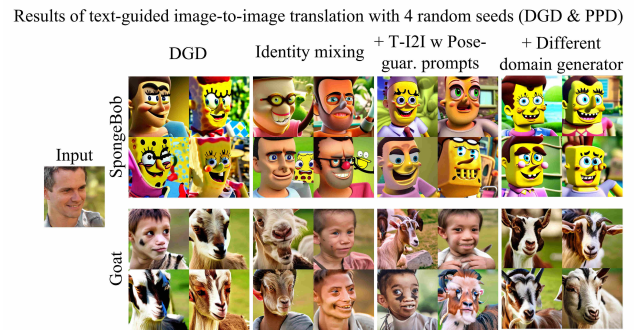

Figure 9. Results of text-guided image-to-image translation depending on the data collection stratgies for training the pose-preserved diffusion model.

User study. We conduct a user study to evaluate the quality of the generated samples and 3D shapes from the shifted generator through baselines and our methods and report the mean opinion score. The participants were requested to assess the visual quality of the generated images in terms of text-image consistency, realism, and diversity using a rating scale ranging from 1 to 5. Additionally, we asked users to rate the text-image correspondence and sense of depth & details for evaluating the 3D shapes. Our results, presented in Table 1 and Table 2, demonstrate the superior text-image correspondence, realism, diversity, and quality of 3D shapes compared to the baselines. See the supplementary Section **D** for further details on the comparison.

4.2. Text-guided debiasing

We apply text-guided debiasing to bear, 3D animation characters, and characters in Zootopia using the attribute species, 3D animation characters, and types of animals, respectively. We obtained the information on subclasses from ChatGPT. As represented in Fig. 7, our text-guided debiasing method enables to enhance intra-domain diversity (Bear, Characters in Zootopia) or further improve the text-image correspondence (3D animation characters).

4.3. Ablation studies

4.4. Pose-preserved text-to-image diffusion

In Fig. 8, we compare the results of text-guided image translation using Stable diffusion (SD), the depth-guided diffusion (DGD), our proposed pose-preserved diffusion (PPD), and our specialized-to-general sampling strategy (S-to-G). We observe that SD exhibits low image-text correspondence, such as unnatural ear shapes when using a low return step, and low pose consistency when using a high return step. DGD suffers from overly strong shape constraints from the depth map. In contrast, our PPD enables pose-consistent image generation but may exhibit biases in style or details. The S-to-G strategy resolves these biases by utilizing the general text-to-image diffusion for creating details. See the supplemetary Fig. S4 for extended results.

Data preparation for training PPD. The effectiveness of training PGD with only identity mixing is not optimal as shown in Fig. 9. However, when PGD is trained with the translated targets from identity mixing using pose-guaranteed prompts, PPD enables large shape changes, particularly when the target domain is similar to the human domain (e.g. SpongeBob). However, PPD fails to produce satisfactory results when the domain gap is significant (e.g. Goat). To overcome this limitation, we leverage another generator trained on a different domain, which enables PPD to translate the input image to target images even with a large domain gap.

5. Discussion and Conclusion

Limitation. Our methods may pose potential societal risks and therefore should be used with caution for appropriate purposes. Further information on limitations and potential negative social impacts can be found in the supplementary Section F.

Conclusion. We propose a novel pipeline called PODIA-3D, for domain adaptation of 3D generative models using pose-preserved text-to-image diffusion models. By utilizing PPD and specialized-to-general sampling models, our method is able to adapt 3D generators to the domains across large domain gaps, broadening applicability of 3D generative models. Our method achieves superior text-image correspondence and 3D shapes compared to existing methods. Additionally, we propose a text-guided debiasing method to address instance bias.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea(NRF) grants funded by the Korea government(MSIT) (NRF-2022R1A4A1030579, NRF-2022M3C1A309202211), a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0316) and Creative-Pioneering Researchers Program through Seoul National University. Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.

References

[1] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davatargan: Bridging domains for personalized editable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4552–4562, 2023.

[2] Aibek Alanov, Vadim Titov, and Dmitry Vetrov. Hyperdomainnet: Universal domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2210.08884*, 2022.

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[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.

[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.

[6] 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.

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

[8] 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.

[9] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

[10] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022.

[11] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9993, 2019.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

[13] Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. *arXiv preprint arXiv:2009.05475*, 2020.

[14] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020.

[15] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[18] Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model. *arXiv preprint arXiv:2211.16374*, 2022.

[19] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[20] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5871–5880, 2020.

[21] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[22] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020.

[23] 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.

[24] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems*, 33:6767–6778, 2020.

[25] 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.

[26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021.

[30] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[33] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[35] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022.

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[37] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019.

[38] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

[39] 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.

[40] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. *Advances in neural information processing systems*, 31, 2018.