HyperPose: Hyper-pose Embeddings for 3D-Aware Generative Models with Self-Supervised Disentangling of Pose and Scene

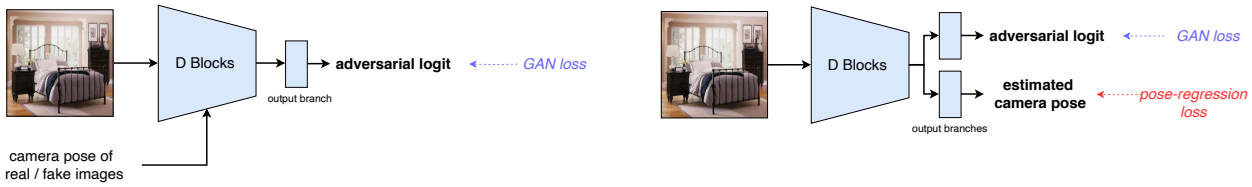
–Supplementary Document–

A. Detailed Comparison with Pose Regression Loss

To supplement the Table 5 in the main paper, this section describes the details of EG3D [3] and our ablative model employing pose regression loss.

A.1. Discriminator Design

Figure Aa shows our base architecture, EG3D [3], which employs a pose-conditional discriminator and thus requires ground-truth camera poses of training images. This limits its applicability to our target domains, where such pose annotations are unavailable. To address this, we modify EG3D by removing the pose-conditioning input and introducing an additional branch to handle learned pose embeddings. Figure Ab illustrates our modified EG3D variant used in the ablation study with pose regression loss [1, 2, 5, 7–10, 15], denoted as EG3D† for clarity. This corresponds to the "w/ pose regression model" entry in Table 5 of the main paper. Unlike the HyperPose, the second branch in the discriminator explicitly predicts pose parameters (pitch and yaw), $\hat{\xi} \in \mathbb{R}^2$, from the input image.



(a) EG3D (Pose-conditioned discriminator)

(b) EG3D† (Discriminator with the pose regression loss)

Figure A. Discriminator designs of EG3D and EG3D†. The pose-conditioned discriminator in (a) utilizes camera pose information as an input, so the ground-truth pose should be given for each training image. (b) employs the pose regression loss on fake images using the estimated pose $\hat{\xi}$ and the rendering pose ξ.

A.2. Pose Regression Loss

The pose regression loss encourages the generator to synthesize the images congruent to the rendering direction $\xi \in \mathbb{R}^2$, where the estimated pose $\hat{\xi}$ from the discriminator is optimized as follows:

$$\mathcal{L}_{\text{regression}} = \mathbb{E}_{z\sim p_z, \xi\sim p_\xi} \left\| \hat{\xi} - \xi \right\|, \tag{A}$$

where $\|\cdot\|$ denotes ℓ_1 or ℓ_2 norm. As similar to our algorithm, the pose regression loss is operated only on fake images, and the overall objective is as follows:

$$\mathcal{L}(D, G) = \mathbb{E}_{I\sim p_{\text{data}}} \left[f(-D(I) + \lambda \|\nabla D(I)\|^2) \right] + \mathbb{E}_{z\sim p_z, \xi\sim p_\xi} \left[f(D(G(z, \xi))) \right] + \lambda_{\text{regression}} \cdot \mathcal{L}_{\text{regression}}, \tag{B}$$

where $f(u) = -\log(1 + \exp(-u))$ and p_{data} denotes the data distribution.

A.3. Additional Comparison with HyperPose

We already reported the comparison with EG3D† and HyperPose on the challenging datasets in Table 5 in the main paper. To enhance more comparison and performance on the simpler geometry dataset, we further test them on the FFHQ human face dataset.

Results on the FFHQ dataset As mentioned in our main paper, our target aims to tackle challenging domains without relying on 3D priors such as camera pose. It is important to note that the FFHQ dataset is not our target domain. However, we observed that HyperPose still performs well on the FFHQ dataset, as shown in Table A. Unlike other challenging datasets in the main paper, the performance gain on the FFHQ dataset is not very significant. This is likely because pose regression is sufficiently straightforward in this dataset due to its simple and homogeneous geometry. Nevertheless, this observation suggests that our algorithm remains competitive not only in complex, in-the-wild scenarios but also in simple geometry.

Table A. Comparison betwen HyperPose and EG3D†. Compared to the pose regression loss used in EG3D†, our algorithm synthesizes high-fidelity images with accurate 3D geometry, showing remarkable performance improvement.

Method	Pose-related	FID	Recall	Precision
EG3D†	Pose regression	5.340	0.415	0.548
HyperPose (ours)	Contrastive learning	5.332	0.419	0.551

Figure B compares EG3D† and HyperPose on the FFHQ dataset. In EG3D†, smiling faces in the front view often turn into non-smiling ones when rendered to the side, due to misalignment across the scenes. In contrast, since our algorithm employs pose matching on the disentangled pose embedding, it preserves facial expressions across viewpoints in the generated images. Additionally, our algorithm produces accurate 3D geometry even for profile faces, capturing fine details of ears and ensuring smooth facial surfaces, as shown in the last column of Figure B. This is thanks to our soft contrastive learning, which allows the generator to capture proper geometry for these out-of-distribution views by learning abundant soft-positive and negative pair relationships.

(a) EG3D† (w/ regression loss on explicit pose representation)

(b) HyperPose (w/ contrastive learning on hyper-pose embeddings)

Figure B. Comparison between EG3D† and HyperPose on the FFHQ dataset (subset with 20,000 images). Compared to EG3D† trained with the pose regression loss, HyperPose achieves improved view consistency and captures more realistic shapes, particularly in out-of-distribution viewpoints.

B. Comparisons with Recent 3D-aware GANs

B.1. Different Target Scopes

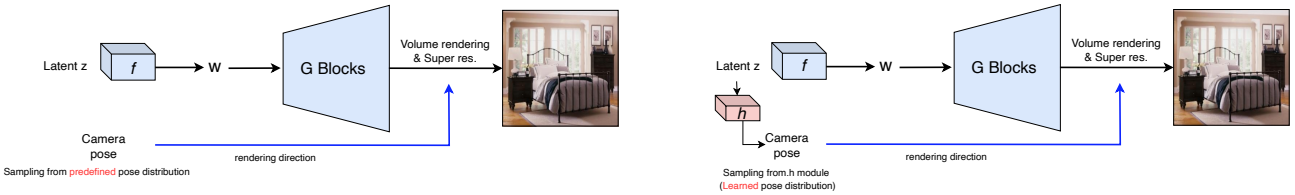
Recent research has generally focused on two main scopes: **(1) learning camera pose distribution priors** and **(2) improving the rendering quality of baseline models** (still using camera pose labels of training images, limited domains such as human face). Most of these works build upon the EG3D architecture, and are definitely orthogonal to our target scope.

Note that our approach presents a novel direction, **(3) handling challenging datasets without any camera pose labels or 3D priors.** Therefore, comparing our method to the fundamental EG3D architecture is still quite reasonable. Nevertheless, we further explore the compatibility of our altogirhtm on this orthogonal algorithm in next subsection.

B.2. Compatibility with PoF3D [13]

PoF3D [13] and HyperPose (ours) are actually orthogonal methods with different target aims: PoF3D tackles the problem of learning camera pose distribution priors (scope 1), while we address challenging datasets (scope 3). To verify their orthogonality and compatibility, we integrated PoF3D into our algorithm to estimate the camera distribution.

We set the camera pose distribution as a predefined prior, with the distributions detailed in Table C in this supplementary document. However, by incorporating PoF3D into HyperPose, we can estimate the pose distribution. PoF3D includes an additional learnable module that performs the role of sampling camera poses as shown in Figure Cb. Table B shows the qualitative results of them on the LSUN Bedroom dataset, where we achieved additional performance gains by integrating PoF3D. Additionally, only our contrastive learning-based training scheme can help to learn complex geometry. Figure D shows that the HyperPose with PoF3D can more closely mimic the real camera pose distribution of the training images (denoted as pseudo-GT in the FFHQ dataset in (b)).



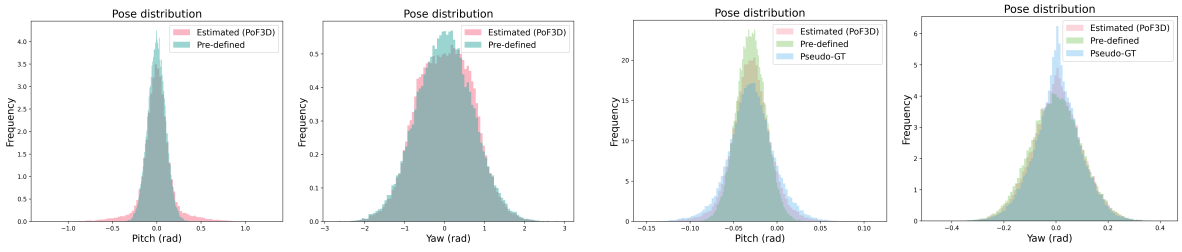
(a) Generator in the main paper

(b) Generator with PoF3D

Figure C. Comparison of the generator between the original and the version incorporated with PoF3D

Table B. Quantitative results on the LSUN bedroom dataset, including those integrating the PoF algorithm, indicate that while PoF enhances generation quality, HyperPose uniquely outperforms regression-based training schemes.

Method	camera distribution	pose-related	FID	Recall	Precision	NFS	Depth-FID
EG3D†	Predefined prior	Pose regression	12.8	0.21	0.46	16.5	138.4
HyperPose (ours)	Predefined prior	Contrastive learning	10.8	0.23	0.56	28.2	49.5
EG3D† + PoF3D	Learnable	Pose regression	12.4	0.21	0.45	16.8	142.4
HyperPose + PoF3D (ours)	Learnable	Contrastive learning	10.3	0.25	0.58	27.5	44.9



(a) LSUN Bedroom (pitch / yaw)

(b) FFHQ (pitch / yaw)

Figure D. Learned camera pose distribution with PoF3D. Through the incorporation of PoF3D, we are able to obtain a more accurate camera pose distribution.

C. Evaluation with High Resolution

To verify that our algorithm, HyperPose, performs well on higher-resolution images, we test our algorithms on the unified AFHQ (Cats, Dogs, and Wildlifes) dataset with the resolution of 512^2. It achieves FID of 7.65, recall of 0.30, precision of 0.56, and NFS of 21.7, similar to resolution 256^2. Figure E exhibits the results of HyperPose, where it produces high-resolution and high-fidelity images with accurate depth maps and surfaces.

Figure E. Curated examples from HyperPose trained on the unified AFHQ (Cats, Dogs, and Wildlifes) with 512^2 resolution. The first column exhibits images rendered with the mean yaw and pitch value, and the second column shows surface renderings with random poses. The last column shows RGB images, and their depth maps when varying camera yaw angles at $-30°$, $-15°$, $15°$, and $30°$.

D. Implementation Details

Architecture of generator Figure F illustrates the detailed architecture of the generator, adapted from EG3D. Let p_z and p_ξ be the distributions of latent variable and camera pose, respectively. Given $z \sim p_z$, the mapping network $f(\cdot)$ in the generator takes z as inputs and produces a scene-specific descriptor w. Subsequently, StyleGAN-v2 modules create a 3D neural radiance field from a scene-specific descriptor w, and the neural radiance field is employed for volume rendering in the direction ξ, producing a low-resolution of 2D feature map and a low-resolution image. Lastly, an image super-resolution module generates a high-resolution image using the provided 2D feature map and the low-resolution image.

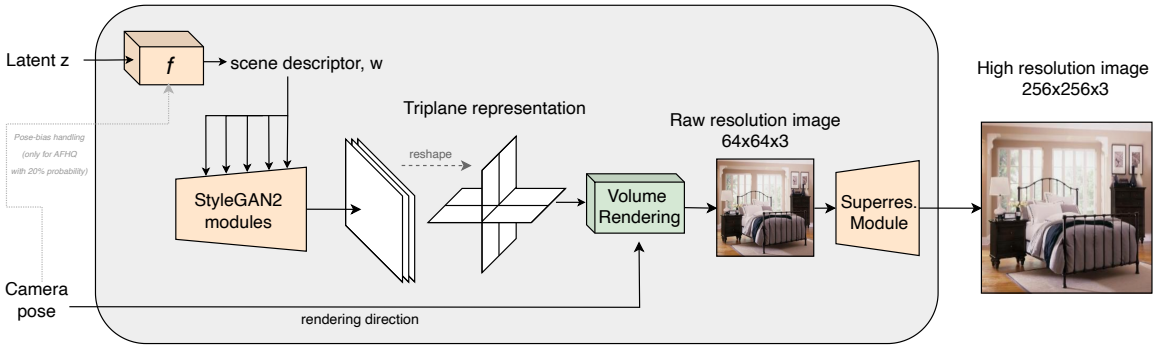


Figure F. Architecture of 3D-aware generator. We adopt the base architecture from EG3D for generator. For the AFHQ dataset, we adopt the pose-bias handling technique proposed in EG3D with only a probability of 0.2, where the scene descriptor w from the mapping network f still represents the scene information regardless of this.

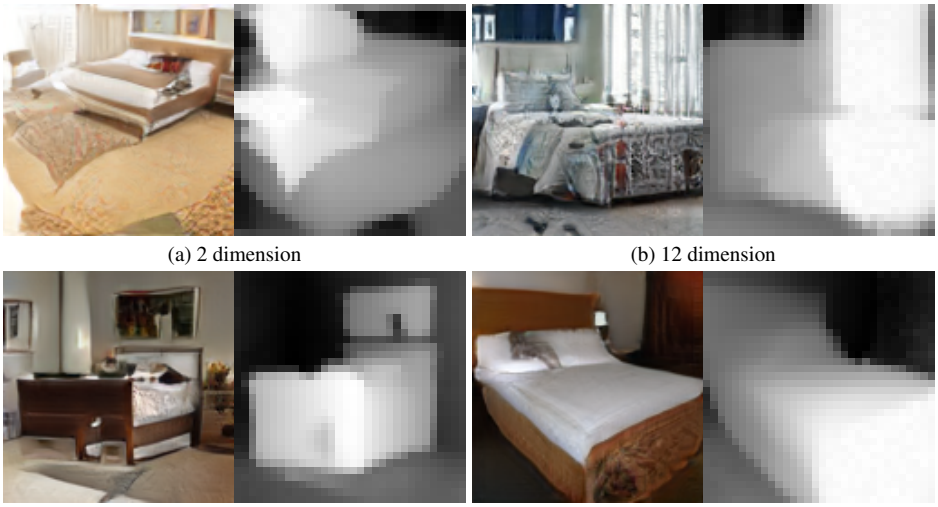
Training Our implementation is based on PyTorch [11] and all experiments are conducted on 4 NVIDIA A6000 or 4 A100 GPUs. During training, we use a batch size of 48 and apply horizontal flips for data augmentation. The pose embedding dimension is set to 24 for the LSUN Bedroom and AFHQ datasets, and 96 for the LSUN Church dataset. For CUB, we also set the pose embedding dimension to 24 and utilize available instance masks to place the birds on a white background. The λ_1 and λ_2 are set to 0.7, 0.9, 0.5, and 0.5 for the LSUN Bedroom, LSUN Church, AFHQ, and CUB datasets, respectively. Table C provides additional details, including the number of images and prior pose distribution, p_ξ, for each dataset.

Table C. The number of images and prior camera pose distribution p_ξ of the LSUN Bedroom, LSUN Church and AFHQ datasets.

Dataset	Number of Images	Pitch		Yaw	
		Distribution	Detail	Distribution	Detail
LSUN Bedroom [18]	3,033,042	Gaussian	$\mu = \pi/2, \sigma = 0.10$	Gaussian	$\mu = \pi/2, \sigma = 0.70$
LSUN Church [18]	126,227	–	$\pi/2$	Uniform	$[\pi/2 - 5\pi/18, \pi/2 + 5\pi/18]$
AFHQ [4]	14,630	Gaussian	$\mu = \pi/2, \sigma = 0.13$	Gaussian	$\mu = \pi/2, \sigma = 0.19$
CUB [16]	11,788	Gaussian	$\mu = \pi/2, \sigma = 0.13$	Uniform	$[\pi/2 - 3\pi/4, \pi/2 + 3\pi/4]$

E. Dimensionality of pose embedding

We analyze the impacts of the dimensionality of pose embedding on 3D reconstruction quality on the LSUN Bedroom dataset and visualize its ablative results in Figure G. Our framework successfully captures 3D structures if it has a sufficient size of embedding dimension $m \geq 24$. Even with low-dimensional embedding vectors, we still obtain decent quality of reconstructed images and depth maps with minor blurs.



(a) 2 dimension (b) 12 dimension

(c) 24 dimension (d) 48 dimension

Figure G. Effects of the pose embedding dimension on the quality of rendered image on the LSUN Bedroom dataset. Our framework successfully captures underlying 3D structures with a sufficient number of embedding dimensions.

F. Limitation with failure cases

HyperPose mostly produces high-fidelity volumetric scenes, but we rarely observe failure cases on the LSUN bedroom dataset. Figure H shows some cases having planar scenes. We presume outlier training samples, such as watermarked images or images captured from out-of-distribution camera poses, may result in these outputs.

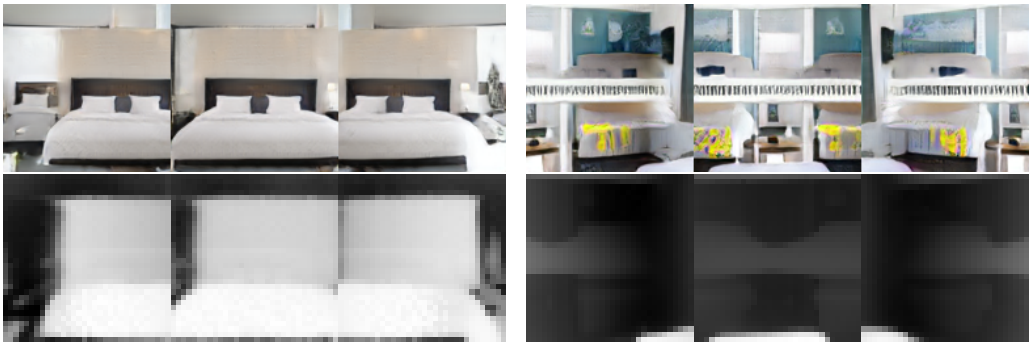
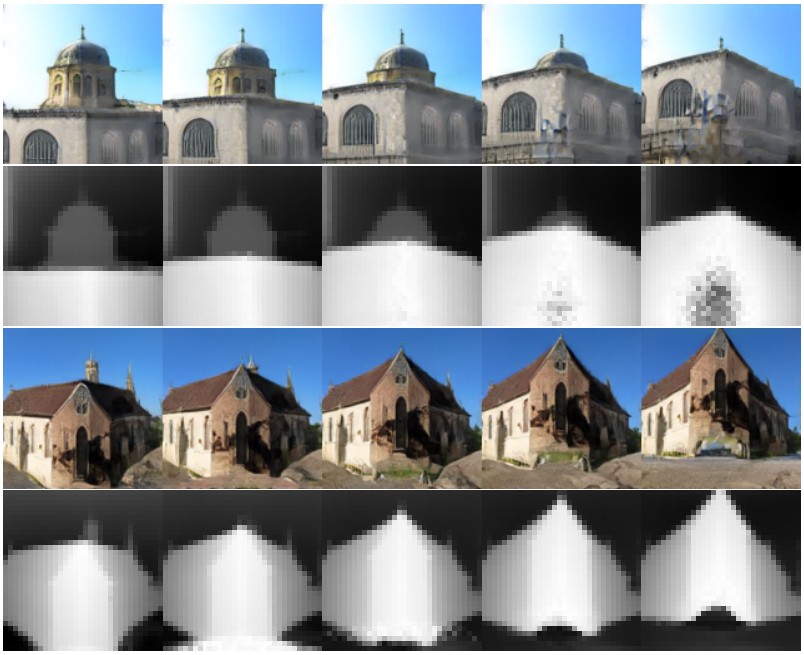


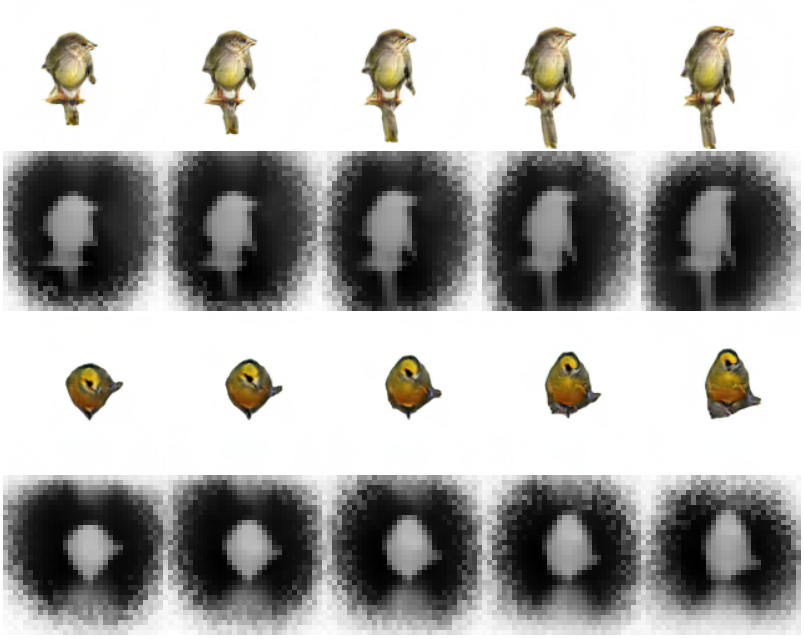
Figure H. Failure cases. HyperPose rarely produces images with unrealistic geometries, such as planar scenes. The first case (left) is the example that HyperPose generates translated images when rotating camera poses, and the second one (right) illustrates the results with almost uniform depth.

G. Qualitative Evaluation with Vertical Rotation

In the main paper, we presented our results using a camera featuring horizontal rotations. To further assess the effectiveness of our algorithms, particularly HyperPose, we display supplementary qualitative results incorporating vertical rotations, as depicted in Figure I and J. The scenes are observed from 5 distinct views, varying camera pitch angles at $20°$, $10°$, $0°$, $-10°$, and $-20°$. For the church dataset, the results show that ours even can generate images with a camera pose from out-of-distribution; during training, the pitch value is fixed for the church dataset.

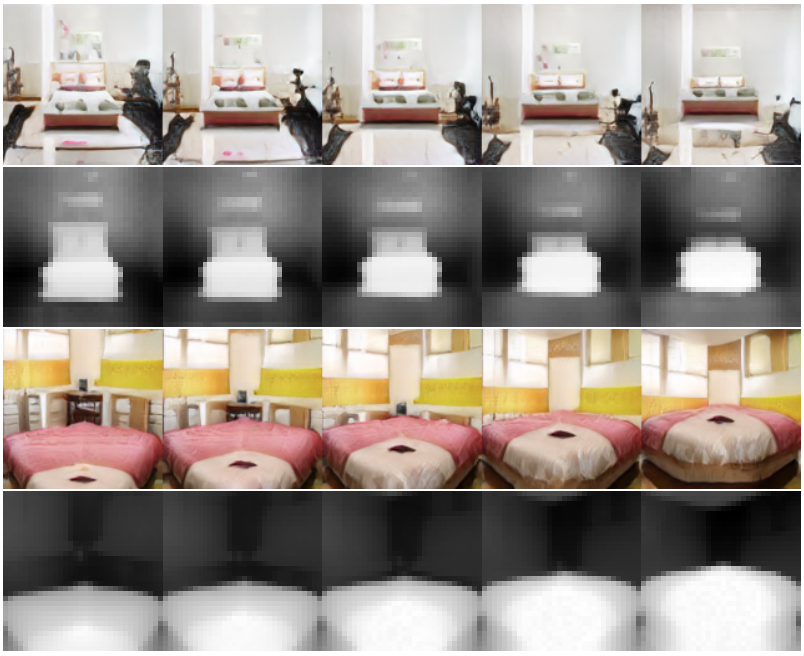


(a) LSUN Church



(b) CUB

Figure I. Qualitative results of HyperPose on the LSUN Church, and CUB dataset. The view angles of each scene are rotated vertically at regular intervals: $20°$, $10°$, $0°$, $-10°$, and $-20°$.



(a) LSUN Bedroom

(b) AFHQ dataset

Figure J. Qualitative results of HyperPose on the LSUN Bedroom and AFHQ dataset. We visualized the rendering results with RGB image and its depth map. The view angles of each scene are rotated vertically at regular intervals: 20°, 10°, 0°, -10°, and -20°.

H. Additional Qualitative Results

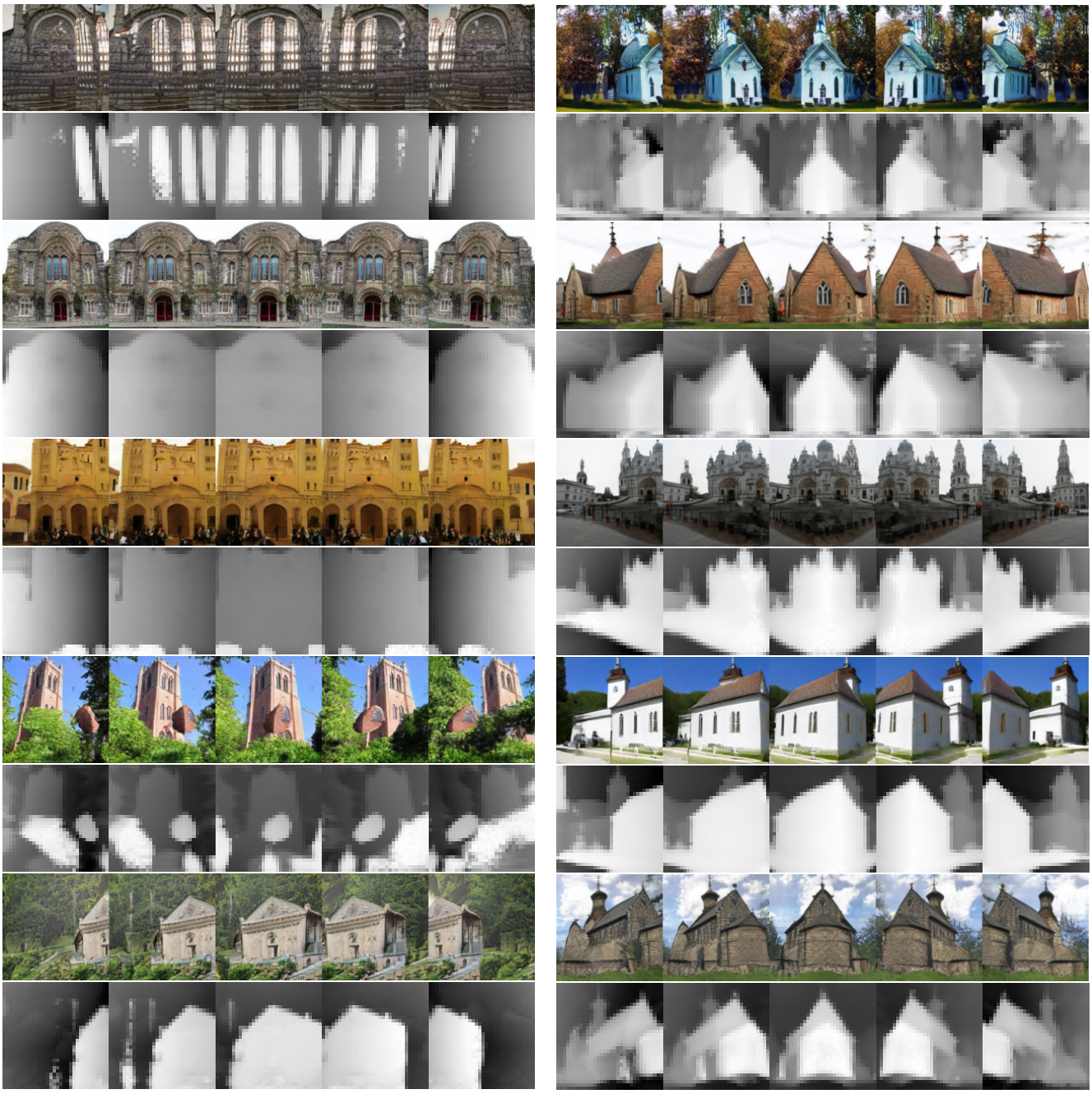
This section demonstrate the qualitative results of individual examples corresponding to the following four datasets: LSUN Bedroom, LSUN Church, AFHQ, and CUB.

H.1. LSUN Bedroom

(a) EG3D† (b) HyperPose

Figure K. Qualitative results of EG3D† and HyperPose on the LSUN Bedroom dataset. We visualized the rendering results with RGB image and its depth map. The view angles of each scene are rotated at regular intervals: $-40°$, $-20°$, $0°$, $20°$, and $40°$.

H.2. LSUN Church



(a) EG3D† (b) HyperPose

Figure L. Qualitative results of EG3D† and HyperPose on the LSUN Church dataset. We visualized the rendering results with RGB image and its depth map. The view angles of each scene are rotated at regular intervals: $-40°$, $-20°$, $0°$, $20°$, and $40°$.

H.3. CUB

(a) EG3D† (b) HyperPose

Figure M. Qualitative results of EG3D† and HyperPose on the CUB dataset. We visualized the rendering results with RGB image and its depth map. The view angles of each scene are rotated at regular intervals: $-40°$, $-20°$, $0°$, $20°$, and $40°$.

H.4. AFHQ

(a) EG3D† (b) HyperPose

Figure N. Qualitative results of EG3D† and HyperPose on the AFHQ dataset. We visualized the rendering results with RGB image and its depth map. The view angles of each scene are rotated at regular intervals: -30°, -15°, 0°, 15°, and 30°.

I. Future Works

Extension to the multi-category datasets While our algorithm effectively tackles challenging indoor and outdoor datasets, scaling to more in-the-wild datasets such as ImageNet presents significant room for improvement, which we believe is crucial for real-world applications. Existing methods that tackle ImageNet, most notably the pioneering work by 3DGP [14], heavily rely on additional 3D priors, such as estimated depth maps. Consequently, the performance of these methods can be severely limited or disproportionately impacted by the accuracy of these priors, especially when such priors are unreliable, which is often the case in complex scenes. For instance, due to inaccurate depth maps, 3DGP often generates flat scenes, as illustrated by failure cases in Figure Oa. Since our approach does not utilize these priors, directly applying it to ImageNet is not straightforward. However, given that the contrastive learning in HyperPose is inherently generic, we believe its effectiveness on multi-category datasets like ImageNet could be further enhanced by incorporating recent techniques, such as hierarchical contrastive learning [6]. We acknowledge this as an important direction for future work and intend to explore it further.

Extension to diffusion-based algorithms While our methodology primarily focuses on GANs, we believe our novel contrastive learning scheme provides general concepts. Score Distillation Sampling (SDS)-based diffusion methods [12, 17] typically provide guidance with the difference between conditional and unconditional scores: $\nabla_\theta \mathcal{L} = w(t)(\epsilon_\phi(z_t; c, t) - \epsilon_\phi(z_t; \emptyset, t))$, where c is camera pose. Our contrastive framework could motivate **HyperPose-SDS** with negative pairs: $\nabla_\theta \mathcal{L} = w(t)(\epsilon_\phi(z_t; c^+, t) - \alpha\epsilon_\phi(z_t; c^-, t) - \beta\epsilon_\phi(z_t; \emptyset, t))$ where c^+ and c^- represent positive and negative conditions and α would be related to our pose similarity, providing richer guidance. We also acknowledge this as important and valuable future work.

(a) 3DGP [14] (b) HyperPose (ours)

Figure O. Comparison with 3DGP [14], illustrating its failure cases often caused by a reliance on inaccurate depth maps.

References

[1] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *CVPR*, 2022.

[2] 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 *CVPR*, 2021.

[3] 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 *CVPR*, 2022.

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

[5] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022.

[6] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *CVPR*, 2023.

[7] Jeong-gi Kwak, Yuanming Li, Dongsik Yoon, Donghyeon Kim, David Han, and Hanseok Ko. Injecting 3d perception of controllable nerf-gan into stylegan for editable portrait image synthesis. In *ECCV*, 2022.

[8] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *International Conference on 3D Vision (3DV)*, 2021.

[9] 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 *CVPR*, 2022.

[10] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *NeurIPS*, 2021.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[12] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

[13] Zifan Shi, Yujun Shen, Yinghao Xu, Sida Peng, Yiyi Liao, Sheng Guo, Qifeng Chen, and Dit-Yan Yeung. Learning 3d-aware image synthesis with unknown pose distribution. In *CVPR*, 2023.

[14] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *ICLR*, 2023.

[15] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *CVPR*, 2022.

[16] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011.

[17] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*.

[18] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.