

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

Mijeong Kim¹ Namgi Kim² Bohyung Han^{1,2}
Computer Vision Lab., ¹ECE & ²IPAI, Seoul National University, Korea
{mijeong.kim, rlaskar177, bhhan}@snu.ac.kr

Abstract

We propose a novel framework for training 3D-aware Generative Adversarial Networks (GANs) from a collection of 2D images, effectively learning both image distribution and 3D geometric configurations without relying on strong 3D priors such as camera poses, depth information, or target-specific 3D models. To achieve these objectives, we introduce hyper-pose embeddings alongside a novel pose disentanglement technique that effectively separates pose and scene information. This crucial disentanglement helps the generative model overcome the inherent conflict between learning photo-realism and accurate 3D geometry. Furthermore, we propose soft contrastive learning to robustly handle the continuous nature of camera poses, and a non-match loss to further enhance disentanglement and refine embedding training. Experiments on challenging datasets demonstrate the effectiveness of our method in 3D-aware image synthesis, particularly for scenes with complex or diverse objects.

1. Introduction

Recent advances in neural radiance fields [42, 47, 51] have expanded the scope of generative models from 2D to 3D domains. They are called as 3D-aware generative models, and synthesize multiple views of a single scene with explicit control over camera poses. We aim to train 3D-aware generative models using only a collection of 2D images without relying on 3D labels like camera pose and depths. While both diffusion-based and Generative Adversarial Network (GAN)-based approaches exist in generative model paradigms, diffusion models typically require 3d labels for training due to its inherent reconstruction loss. In contrast, GANs facilitate label-free training by learning the real data distribution through min-max optimization, thus the GAN-based method is more suitable to achieve our goal.

While most existing 3D-aware GANs are limited to homogeneous structures like human or cat faces, recent ef-

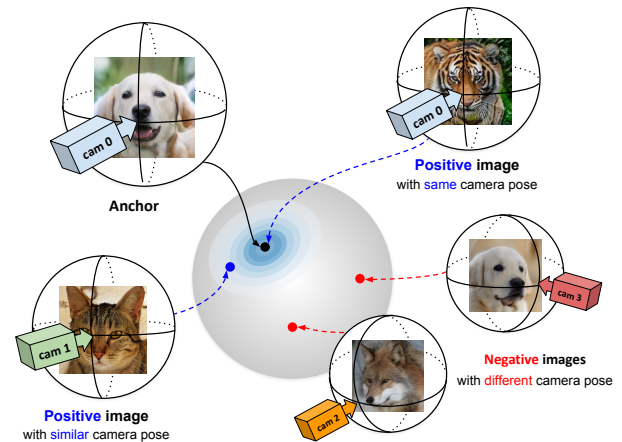


Figure 1. Illustration of the proposed soft contrastive learning on the hyper-pose embedding space. The *positive* and *negative* images denote images rendered in the same/similar or different directions with the *anchor* image, respectively. The distances between pose embeddings of positive pairs are learned to be closer than those of negative pairs, considering their physical similarity on continuous space.

orts [3, 14, 23, 35, 61, 77] have expanded target domains. However, these methods often necessitate additional geometric priors (e.g., depth maps, explicit camera poses, or 3D models), incurring substantial costs or limiting their applicability to specific domains. Learning 3D configurations without such priors is crucial for real-world applicability, yet it presents significant challenges.

Training 3D-aware GANs poses a significantly greater challenge than 2D GANs. This is because the discriminator’s decision boundary must simultaneously account for visual fidelity and the inherent 3D geometry. This dual complexity, particularly with intricate 3D configurations, often places a substantial burden on the discriminator, leading to degraded sample quality. Our primary focus, therefore, is to enhance the discriminator’s ability to comprehend 3D configurations without compromising its capacity to learn the real image distribution. To achieve this, we propose a re-

laxed technique that reduces the burden on the generative model for explicitly handling 3D information, while still accurately learning 3D configurations based on our hyper-pose embeddings.

In our approach, we first modify the discriminator design of a GAN, allowing it to additionally estimate high-dimensional embeddings from generated images, thus achieving a more expressive representation of the geometry. Furthermore, we carefully disentangle the implicit pose information and the scene information, from which we derive hyper-pose embeddings. Through this pose disentanglement process, our model reserves the capability to generate more diverse scenes by preventing unintended alignment of scene appearance, all while maintaining accurate pose matching with rendering direction.

To align the pose embeddings of generated images with their corresponding rendering directions, we employ self-supervised contrastive learning [48] on these embeddings. Unlike traditional contrastive methods that typically operate on discrete classification labels, our approach is uniquely designed for the continuous pose space. Specifically, we introduce a novel soft contrastive learning approach, as depicted in Figure 1. By deeply considering the physical relationships among multiple examples within this continuous domain, our soft contrastive learning method effectively captures rich 3D configurations. Complementing this, we introduce a non-match loss. This loss functions by providing additional negative examples, which are constructed by incorrectly pairing pose embeddings with scene information. This mechanism further enhances the disentanglement of pose and scene, leading to a more robust and accurate 3D representation.

Our experiments demonstrate that the proposed approach outperforms existing algorithms in both 3D generation and reconstruction metrics. Figure 2 illustrates generated examples by the proposed algorithm in challenging datasets. Our main contributions are summarized below:

- We propose a novel self-supervised framework for 3D-aware GANs that effectively learns 3D geometric configurations from 2D image collections without strong 3D priors, addressing challenging datasets with complex or diverse objects.
- We introduce hyper-pose embeddings alongside a novel pose disentanglement technique that effectively separates pose and scene information, crucial for maintaining appearance diversity during contrastive learning.
- We propose soft contrastive learning for continuous camera poses, enabling robust 3D structure learning by considering physical relationships among examples
- Our framework achieves state-of-the-art performance in 3D-aware image synthesis, thoroughly validated through extensive qualitative and quantitative experiments.

2. Related Work

2.1. Neural Scene Representations

Encoding a scene into a neural network becomes a promising direction in the research of representing a 3D scene. It has been renowned for its spatially-continuity and memory-efficiency compared to point-based, voxel-based, and mesh-based representation methods. More specifically, a neural network maps 3D coordinates to any representation, usually geometric representation based on the occupancy [11, 41] or sign distance function [36, 51, 59, 65], and sometimes included appearance [42, 47, 58].

Neural Radiance Field (NeRF) [42] is one of the most fundamental researches drawing a lot of attention to this field. It proposes a neural volume rendering technique, which synthesizes a 2D image corresponding to a given camera pose by repeatedly sampling the representation along a ray per each pixel and aggregating them based on traditional volume rendering. Its photorealistic image quality of rendered 2D images results in many following researches [2, 15, 25, 26, 32, 37, 40, 43, 66, 68, 71, 73, 79]. Some of them [5–7, 13, 45, 49, 50, 63], including this work, incorporate NeRF into the generative model.

2.2. 3D-aware Generative Models

Many 3D-aware generative models [5–7, 13, 18, 34, 45, 49, 50, 54, 60, 63] integrate volume rendering techniques [42, 47, 51] into Generative Adversarial Networks (GANs) [4, 12, 16, 29–31, 75]. This integration allows for the synthesis of multiple views of a single scene with explicit control over camera poses, even enabling training with unorganized 2D image datasets—the same datasets used for 2D GANs. EG3D [7] incorporates StyleGAN [29] and generates high-resolution images through a super-resolution module. Following the success of EG3D, many subsequent works [10, 45, 49, 56, 62, 70, 78] have been proposed based on this model. Some research, including CAM-PARI [45] and PoF3D [56], relaxes the strict dependence on pre-defined camera pose distributions by introducing an additional learnable module that refines the camera poses. Another line of research has focused on improving the rendering quality of EG3D; for instance, Mimic3D [10] enhances the fidelity of local image patch with an additional patch-based discriminator, while GRAM-HD [70] improves 3D consistency by performing super-resolution directly in 3D space. Note that these recent 3d-aware GANs build upon EG3D, which is fundamental architecture, but they still rely on camera pose label of each training image or are restricted to domains with simple geometry, such as human faces. Thus, we introduce an orthogonal and novel line of research: **handling challenging geometries without camera pose labels or target-specific 3D modelings.**

Recently, with the success of diffusion models [22] in 2D

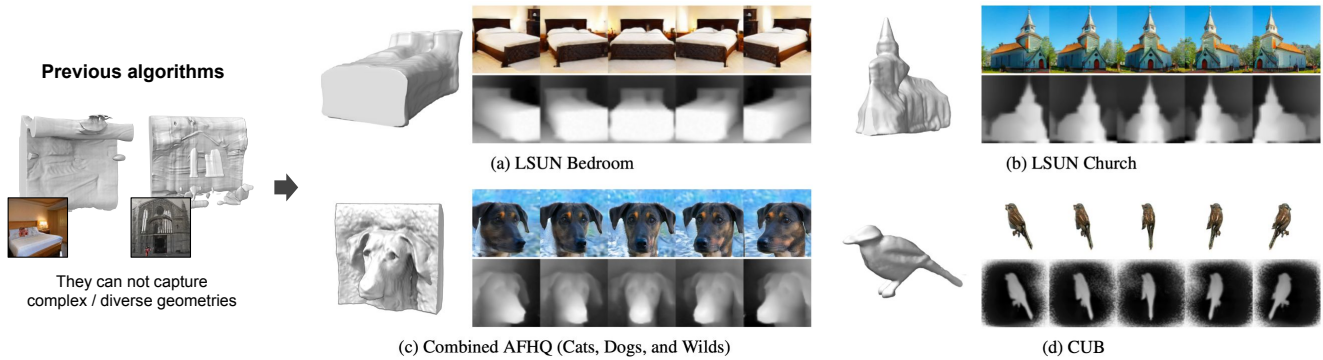


Figure 2. Illustration of generated examples by HyperPose. HyperPose enables the synthesis of scenes with complex geometry such as bedrooms, churches, animal faces, or birds, beyond simple geometries such as human faces. Our approach trains from a collection of 2D images without strong 3D priors such as the camera pose of each training data, depth information, and target-specific 3D modelings.

images, there have been attempts to use them for 3D scene generation. DreamFusion [53] and its variants [24, 38] introduce a text-to-3D framework that generates multiple views of a scene from a text prompt using pre-trained 2D diffusion models. However, these methods are limited by single-scene optimization, meaning each new scene requires optimization from scratch, resulting in significant computational costs. Another line of diffusion-based research [1, 8, 8, 44, 55, 57, 64] focuses on directly optimizing 3D-aware diffusion, but it inherently requires the precise camera pose or depth maps of each training images. In these views, our paper focuses suggesting a novel 3D-aware GAN approach capable for unsupervised scenarios.

2.3. 3D-aware GANs on Complex Scenes

Some 3D-aware GANs have tackled generation tasks on complex datasets composed of images with diverse geometric configurations. However, existing approaches mostly rely on the prior knowledge of scenes such as target-specific 3D modelings or depth information. For example, full-body human generation techniques [17, 23, 77] demonstrate impressive results with high-fidelity geometries and motions but cannot be generalized to other domains because they rely on the pre-trained human body modeling such as SMPL [39]. Some other algorithms [3, 14, 35, 61] utilize the estimated depth maps of training images to obtain direct 3D information. However, most of them are limited to 3D indoor scenes, where accurate depth maps can be obtained. On the other hand, our algorithm does not rely on explicit geometric information such as 3D modelings and depth or exact camera pose of training examples, so it can be generalized to various applications more easily.

2.4. Contrastive Learning

Contrastive learning is a widely used self-supervised representation learning scheme [9, 20, 48, 69]. The idea of

contrastive learning is to train a network to keep the representation of anchor data close to the representation of relevant positive data while pushing it away from those many mismatched negative data. Recently, some generative algorithms have adopted contrastive learning scenarios. CntrGAN [80] adds contrastive learning to train GANs together with image augmentations, where it serves as a regularizer to improve generation fidelity. Contrastive learning has also been used in image-to-image translation [19, 33, 52] and cross-modal translation [76] to enforce patch-wise correspondence and mutual information between image and text, respectively. Also, ContraGAN [28] proposes a class-conditional contrastive learning objective to increase the correlations between images of the same class.

SideGAN [27] introduces a contrastive learning-inspired framework for 3D-aware GAN training, specifically to address the limited number of side-view images in training datasets. Their approach can be seen as label-conditioned contrastive learning, as it depends on ground truth pose labels of the training images. Thus, their method can not be applied to complex scenes where pose labels are unavailable. In contrast, our paper pioneers a self-supervised contrastive learning framework integrated directly into a 3D-aware generative model. Our novel approach introduces several key techniques, including pose disentanglement, a non-match loss, and the use of soft positive labels, marking a significant advancement in the field.

3. Method

We aim to train 3D-aware GANs using unorganized 2D image collections that feature complex and diverse geometric configurations. To achieve this goal, we focus on enabling the discriminator to effectively learn 3D geometry in a self-supervised manner by reducing its learning burden, yet maintaining accurate 3D configuration acquisition. This section details the methodology.

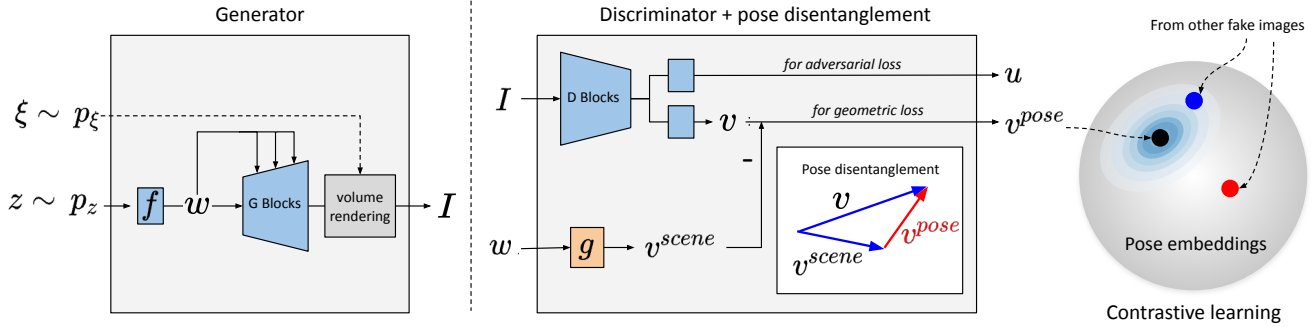


Figure 3. Overview of the proposed 3D-aware GAN framework. Given a camera pose $\xi \sim p_\xi$ and a random vector $z \sim p_z$, the generator G produces a fake image I , where w is an intermediate output representing scene-specific vector. When provided with the fake image, the discriminator D estimates both an embedding v and an adversarial logit u . To disentangle an implicit pose embedding v^{pose} from the embedding v , we employ w from the generator, and train the pose embeddings via the proposed soft contrastive learning.

3.1. Overview

We present an overview of our generative model framework, designed to learn 3D scene configurations from only 2D image sets in a self-supervised manner.

Generator Our generator borrows the architecture from EG3D [7], primarily based on StyleGAN [31], as illustrated in Figure 3. Let p_z and p_ξ represent the distributions of the latent variable z and camera pose ξ , respectively. Given $z \sim p_z$, the mapping network $f(\cdot)$ in the generator produces a 3D scene specific descriptor w as follows:

$$f : z \rightarrow w. \quad (1)$$

Given the scene descriptor w , the remaining modules of the generator construct a 3D neural radiance field corresponding to w . Subsequently, for a given camera pose ξ , we render a 2D image via a neural volume rendering technique [42]. The overall procedure of generator is summarized as follows:

$$G : (z, \xi) \rightarrow I. \quad (2)$$

Note that we can produce multi-view images of the same scene by varying ξ with a fixed z . For clarity and generalization of explanation, we omit details about EG3D-specific components, such as the super-resolution module. Please refer to the original paper [7] for more details.

Discriminator Our model is designed to learn both photorealistic image generation and the underlying 3D scene configuration. To achieve this, our discriminator incorporates two distinct branches, where they outputs an adversarial logit u and a latent scene embedding $v \in \mathbb{R}^m$, respectively, as follows:

$$D : I \rightarrow (u, v), \quad (3)$$

where $u \in \mathbb{R}$ represents the adversarial logit for discerning between real and generated images, and v is specifically designed for evaluating the 3D configuration.

Motivation Although ground-truth poses for training images are unavailable, we can use the camera pose ξ used for rendering fake images as a substitute. Existing approaches [5, 6, 13, 45, 49, 50, 63] employ a discriminator to estimate an explicit camera pose $v \in \mathbb{R}^2$, typically representing yaw and pitch. This representation is aligned with the rendering direction ξ via regression loss as follows:

$$\mathcal{L}_{\text{regression}} = \mathbb{E}_{z \sim p_z, \xi \sim p_\xi} \|v - \xi\|, \quad (4)$$

which ensures that images generated with the same ξ share a consistent object pose. However, this regression loss has significant limitations. First, it struggles to adequately capture complex scene geometries due to its limited capacity. Second, the embeddings often encode irrelevant information unrelated to pose, leading to pose-scene geometry entanglement, where pose and scene structure become intertwined. This entanglement is further amplified by the naïve regression loss, ultimately restricting the diversity of generated scenes.

To address these challenges, we propose a higher-dimensional embedding vector $v \in \mathbb{R}^{m \gg 2}$, which provides a richer and more expressive representation. Additionally, we introduce a pose disentanglement technique and soft contrastive learning, which together ensure that pose and scene geometry are disentangled effectively, allowing for more accurate modeling of 3D configurations while preserving the diversity of generated scenes.

3.2. Hyper-pose Embedding

The discriminator outputs the latent scene embedding v , as detailed in Equation (3), which possesses both scene geometry and rendering direction. However, such an entangled

embedding may incur unintended visual alignments of the scenes generated from different z values, instead of solely matching the camera pose information. To tackle this limitation, we introduce a disentangling strategy to separate the camera pose component from v .

Scene embedding The scene-specific embedding w in Equation (1) is one-to-one mapping to a generated 3D neural radiance field and involves no information about the camera direction ξ for volume rendering. Thus, we extract the scene embedding $v^{\text{scene}} \in \mathbb{R}^m$ from w as follows:

$$g : w \rightarrow v^{\text{scene}}, \quad (5)$$

where $g(\cdot)$ is an MLP network.

Pose disentanglement Given a scene embedding v^{scene} in Equation (5), we separate the pose embedding $v^{\text{pose}} \in \mathbb{R}^m$ from v as follows:

$$v^{\text{pose}} = v - v^{\text{scene}}, \quad (6)$$

where a pose embedding captures only the camera pose information by eliminating scene-specific information v^{scene} from v , as also shown in Figure 3. This pose disentanglement allows our model to generate more diverse scenes by preventing unintended visual alignments across different z .

3.3. Training Objective

This section details our training methodology for learning robust pose embeddings v^{pose} that encode pose information. We employ a contrastive learning framework that maximizes mutual information between synthesized images sharing the same camera pose. Unlike standard contrastive approaches designed for discrete labels, we develop soft contrastive learning to handle the continuous pose space. Furthermore, we propose an enhanced contrastive scheme with strategic negative sampling to improve disentanglement between pose and content representations.

Pose similarity and positiveness During training, we have the rendering pose information ξ of every fake example and can compute the pose similarity between two arbitrary fake images, which is given by

$$S(\xi_1, \xi_2) = \exp\left(-\frac{d(\xi_1, \xi_2)^2}{2\sigma^2}\right), \quad (7)$$

where ξ_1 and ξ_2 are two camera pose parameters in the (*yaw*, *pitch*) space, $d(\cdot, \cdot)$ denotes the cosine distance, and σ corresponds to the standard deviation. Based on the similarity between two poses computed in Equation (7), we define a smooth positive mask, which is given by

$$M(v_1^{\text{pose}}, v_2^{\text{pose}}) = \begin{cases} S(\xi_1, \xi_2) & \text{if } S(\xi_1, \xi_2) > \epsilon \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

where v_1^{pose} and v_2^{pose} are the estimated pose embeddings from the images rendered in the directions of ξ_1 and ξ_2 , respectively. This soft positive pairing, illustrated in Figure 1, naturally handles the continuous pose space by assigning graded similarity rather than binary positive/negative labels.

Soft contrastive loss Let us formally define the proposed soft contrastive loss, which considers the similarity of camera poses in a self-supervised way. Denote by $y_i = G(z_i, \xi_i)$ ($i = 1, \dots, m$) a fake example generated with (z_i, ξ_i) . Then, our soft contrastive loss is given by

$$\ell(y_i) = -\log\left(\frac{\sum_{j \neq i} M(v_i^{\text{pose}}, v_j^{\text{pose}}) \exp(\langle v_i^{\text{pose}}, v_j^{\text{pose}} \rangle / \tau)}{\sum_{j \neq i} \exp(\langle v_i^{\text{pose}}, v_j^{\text{pose}} \rangle / \tau)}\right) \quad (9)$$

where v_i^{pose} is the estimated implicit pose embedding of a fake example y_i and τ is a temperature. This loss term is applied to every example in the mini-batch, which also includes the training data stored in a memory because the proposed contrastive learning approach is based on MoCo [20]. To sum up, the proposed soft contrastive loss is given by

$$\mathcal{L}_{\text{SCL}} = \mathbb{E}_{z \sim p_z, \xi \sim p_\xi} \ell(y_i). \quad (10)$$

Non-matching loss Besides the straightforward contrastive loss in Equation (10), we incorporate an additional soft contrastive loss to provide a more concrete constraint for implicit pose disentanglement, named non-matching loss. The additional soft contrastive loss is defined with additional negative examples given by

$$\bar{\mathcal{V}}_i = \{\bar{v}^{\text{pose}} | \bar{v}^{\text{pose}} = v_i - v_j^{\text{scene}}, j \neq i\}, \quad (11)$$

where v_i and v_j are associated with fake examples y_i and y_j , respectively, and $\bar{\mathcal{V}}_i$ is a set of the implicit camera poses that are incorrectly disentangled. Using the set $\bar{\mathcal{V}}_i$ consisting of negative examples, another soft contrastive loss is defined as

$$\ell'(y_i) = -\log\left(\frac{\sum_{j \neq i} M(v_i^{\text{pose}}, v_j^{\text{pose}}) \exp(\langle v_i^{\text{pose}}, v_j^{\text{pose}} \rangle / \tau)}{\sum_j \exp(\langle v_i^{\text{pose}}, \bar{v}_j^{\text{pose}} \rangle / \tau)}\right) \quad (12)$$

where \bar{v}^{pose} is an element of $\bar{\mathcal{V}}_i$. The only difference between (9) and (12) is the denominator, which indicates the augmentation of negative examples. Thus, the proposed non-matching loss, referred to as $\mathcal{L}_{\text{non-match}}$, is given by

$$\mathcal{L}_{\text{non-match}} = \mathbb{E}_{z \sim p_z, \xi \sim p_\xi} \ell'(y_i). \quad (13)$$

Overall objective On top of the standard GAN loss, the \mathcal{L}_{SCL} and $\mathcal{L}_{\text{non-match}}$ are employed to provide 3D awareness

and the total loss of our algorithm is given by

$$\begin{aligned} \mathcal{L}(D, G) = & \mathbb{E}_{\mathbf{I} \sim p_{\text{data}}} [f(-D(\mathbf{I}) + \lambda \|\nabla D(\mathbf{I})\|^2)] \\ & + \mathbb{E}_{\mathbf{z} \sim p_z, \mathbf{c} \sim p_c} [f(D(G(\mathbf{z}, \mathbf{c})))] \\ & + \lambda_1 \cdot \mathcal{L}_{\text{SCL}} + \lambda_2 \cdot \mathcal{L}_{\text{non-match}}, \end{aligned} \quad (14)$$

where the first term is active for real images, updating only the discriminator, while the remaining loss terms optimize both the generator and the discriminator with fake images.

4. Experiments

This section reports the performance of our method, HyperPose, on benchmarks featuring complex geometric structures. Furthermore, we discuss how our high-dimensional pose embeddings effectively enhance the 3D configuration capabilities within 3D-aware generative models.

4.1. Experiments Settings

Datasets We report results on four different image datasets: LSUN Bedroom [74], LSUN Church [74], AFHQ (Animal Faces-HQ)[12], and CUB[67]. These datasets are challenging for 3D-aware generative models due to difficulties in defining a canonical pose for the LSUN datasets, while AFHQ and CUB include complex and diverse geometric structures. For AFHQ, we unified the three animal face categories—*cats*, *dogs*, and *wildlife*—into a single dataset to increase both diversity and complexity.

Evaluation protocols For the evaluation of image quality, we adopt Fréchet inception distance (FID) [21] and Precision & Recall. For the evaluation of 3D configuration, we utilize Depth FID and the Non-Flatness Score (NFS) [61]. Depth FID measures the FID between the ground-truth depth maps of training images and the rendered depth maps, where it is important to note that ground-truth depth maps are only available for the LSUN Bedroom indoor scene. NFS, computed as the average entropy of normalized depth map histograms, serves as a proxy measure of the flatness of generated scene surfaces.

4.2. Main Results

4.2.1. Quantitative performance

Tables 1, 2, 3, and 4 demonstrate that our HyperPose consistently outperforms existing algorithms¹ across all datasets with remarkably lower FID and higher NFS scores, confirming that its synthesized 3D scenes effectively reflect true geometries. Additionally, our algorithm generates high-quality images with a high recall performance, thereby demonstrating the diversity of the generated scenes. This

¹Our comparisons are primarily with fundamental architectures. For a discussion on compatibility with recent orthogonal lines of extension, please refer to our supplementary material.

Table 1. Quantitative comparison on the LSUN Bedroom dataset with 128² resolution. Our models outperform existing methods by significant margins in all image quality metrics, where HyperPose successfully learns 3D geometry information.

Method	Pose	Depth FID ↓	FID ↓	Recall ↑	Precision ↑	NFS ↑
GRAF [54]	Parametric	97.4	70.7	0.00	0.42	19.4
π -GAN [6]		124.1	56.3	0.11	0.44	9.7
GIRAFFE [46]		145.6	42.8	0.02	0.55	16.9
GIRAFFE-HD [72]		—	27.7	0.13	0.44	—
HyperPose (ours)	Hyper	49.5	12.5	0.23	0.56	28.2

Table 2. Quantitative comparison on the LSUN Church dataset with 128² resolution. Our models outperform baselines by significant margins in all image quality metrics, where HyperPose achieves the best performance. The astro (*) denotes that the scores are taken from GIRAFFE-HD [72].

Method	Pose	FID ↓	Recall ↑	Precision ↑	NFS ↑
GRAF [54]	Parametric	91.1	0.00	0.53	9.3
π -GAN [6]		56.8	0.18	0.49	24.4
GIRAFFE [46]		38.4	0.02	0.51	13.5
GIRAFFE-HD [72]*		10.3	—	—	—
HyperPose (ours)	Hyper	5.8	0.37	0.60	29.9

Table 3. Quantitative comparison on the CUB dataset, having large object pose variations. HyperPose significantly outperforms existing methods in all image quality metrics.

Method	Pose	FID ↓	Recall ↑	Precision ↑	NFS ↑
GRAF [54]	Parametric	46.3	0.09	0.67	21.3
π -GAN [6]		48.8	0.10	0.64	22.1
GIRAFFE [46]		49.3	0.04	0.68	30.6
GIRAFFE-HD [72]		24.3	0.17	0.67	—
HyperPose (ours)	Hyper	10.8	0.39	0.62	44.5

Table 4. Quantitative comparison on the unified AFHQ dataset with 256² resolution. HyperPose achieves the best performance, generating high-fidelity images with accurate 3D geometry. Scores with an astro (*) are from StyleNeRF [18].

Method	Pose	FID ↓	Recall ↑	Precision ↑	NFS ↑
GRAF [54]	Parametric	107.0	0.00	0.35	8.5
π -GAN [6]		48.4	0.12	0.41	21.4
GIRAFFE [46]		31.3	0.04	0.51	14.2
GIRAFFE-HD [72]		14.2	0.10	0.55	—
StyleNeRF [18]*		14	—	—	—
HyperPose (ours)	Hyper	7.5	0.30	0.53	19.2

performance stems from the replacement of existing regression losses [6, 13, 45, 50, 63] for \mathbf{z} and pose mappings with a contrastive learning approach, allowing for a more flexible representation. Despite this relaxation, our method still effectively learns rich 3D information, thanks to our pose disentanglement and a sophisticated training scheme, ultimately leading to both high-quality image generation and accurate 3D configurations.

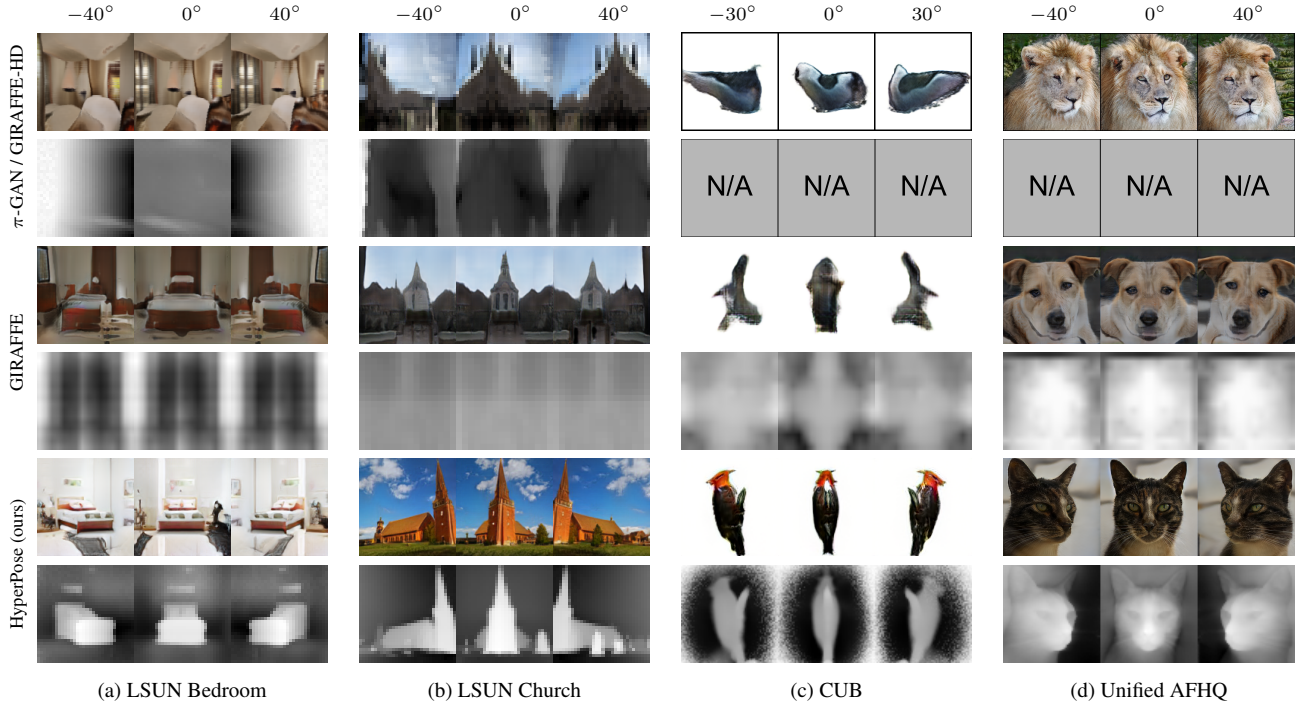


Figure 4. Comparison between HyperPose and existing 3D-aware GANs, which are not relying on strong 3D priors. We show RGB images and depth maps by rotating the rendering pose horizontally. The first two results in the first row are from π -GAN, and the rest from GIRAFFE-HD. HyperPose generates high-fidelity images with accurate depth maps across all domains thanks our pose disentanglement technique with a novel training objective; the existing methods usually produce unrealistic 3D structure.

4.2.2. Qualitative performance

Figure 4 illustrates the generated images and their depth maps with different rendering views. The quality of the depth maps from HyperPose looks impressive, and we notice that all the compared methods mostly fail to reconstruct 3D structures accurately. The experimental results demonstrate that the propose algorithm can learn 3D configurations more easily compared to other algorithms, and it even achieves success on challenging datasets.

4.3. Analysis

Comparison with pose-regression technique To fairly evaluate the effectiveness of our proposed hyper-pose embedding and contrastive learning, we constructed an ablation model built upon an identical architecture to ours, employing a common pose-regression loss $\mathcal{L}_{\text{regression}}$ in Equation (4). In this ablation model, the second branch of its discriminator estimates a 2-dimensional pose vector, on which a standard regression loss is applied. For detailed implementation, please refer to our supplementary material.

Table 5 shows that HyperPose achieves significantly higher recall and lower FID scores. This verifies that our contrastive learning on hyper-pose space effectively enables the model to learn not only the 3D configuration but also the 2D image distribution. This strong performance primarily

Table 5. Comparison with the regression loss, $\mathcal{L}_{\text{regression}}$. Compared to the regression loss, our algorithm (based on contrastive learning on hyper-pose space) can synthesize high-fidelity images with accurate 3D geometry showing the remarkable performance.

Dataset	Method	Metric				
		FID ↓	Recall ↑	Precision ↑	NFS ↑	Depth-FID ↓
Bedroom	w/ $\mathcal{L}_{\text{regression}}$	12.8	0.21	0.46	16.5	138.4
	HyperPose (ours)	10.8	0.23	0.56	28.2	49.5
CUB	w/ $\mathcal{L}_{\text{regression}}$	18.7	0.31	0.61	37.8	-
	HyperPose (ours)	12.5	0.39	0.62	44.5	-
Church	w/ $\mathcal{L}_{\text{regression}}$	7.3	0.35	0.55	22.7	-
	HyperPose (ours)	5.8	0.37	0.60	29.9	-
AFHQ (unified)	w/ $\mathcal{L}_{\text{regression}}$	7.8	0.23	0.58	22.1	-
	HyperPose (ours)	7.3	0.32	0.54	22.1	-

stems from our ability to relax the rigid correlation between z and explicit pose parameters, a constraint typically enforced by regression losses. Despite this relaxation, ours still effectively learns rich 3D information, thanks to our pose disentanglement and a sophisticated training scheme.

Benefit of soft contrastive learning Our soft contrastive learning is specifically designed to handle the continuous nature of camera poses. Real-world datasets exhibit bi-

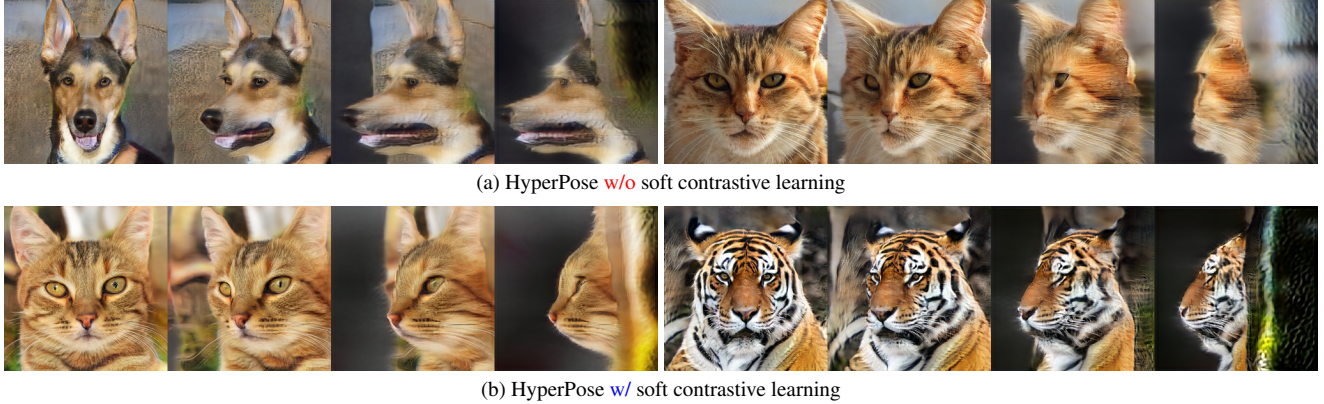


Figure 5. Qualitative comparisons between HyperPose with and without a smooth positive definition, conducted on the unified AFHQ dataset at 256^2 resolution, are presented. By enforcing soft positive pairs based on their similarity in the physical space, our model learns more complex 3D representations. In other words, this soft definition is more appropriate for this continuous pose domain.

Table 6. Ablation study of pose disentanglement and mismatch loss $\mathcal{L}_{\text{non-match}}$ on the LSUN bedroom dataset with 128^2 resolution. Adversarial training and contrastive training benefit from pose disentanglement and mismatch loss, respectively.

Pose disentangle.	$\mathcal{L}_{\text{non-match}}$	FID↓	Recall↑	Precision↑	NFS↑
		13.4	0.22	0.51	28.4
✓		12.6	0.23	0.54	28.0
✓	✓	12.5	0.23	0.56	28.2

ased pose distributions (e.g., frontal faces, side-view birds), creating many similar-pose pairs. Naïve hard contrastive learning treats these negligibly different poses as negatives, causing unstable training, while our approach treats similar poses as smooth positive examples. Thus, our methodology provides more stable gradients to enable learning feature information in a continuous label manifold. Figure 5 demonstrates this effect, where the generative model struggles to capture 3D configurations effectively without our soft contrastive formulation.

Benefit of pose disentanglement & $\mathcal{L}_{\text{non-match}}$ Table 6 presents ablation results on the LSUN bedroom dataset, highlighting the individual contributions of our pose disentanglement method and the non-match loss. When contrastive learning is applied directly to entangled embeddings (the embedding v in Equation (3)), it severely hampers the generative model’s capacity for effective 3D configuration acquisition and high-fidelity 2D image synthesis. This degradation arises because the contrastive objective, when imposed on pose-content entangled representations, conflicts with the adversarial goal of producing perceptually realistic outputs—making discriminator optimization unstable. In contrast, applying contrastive learning within the disentangled hyper-pose space mitigates this conflict. It allows the discriminator to focus on pose-related differences

Table 7. Experiments on the AFHQ dataset with 512^2 resolution. HyperPose achieves the similar performance with 256^2 resolution, and outperforms the existing method in high-resolution settings.

Resolution	Method	FID↓	Recall↑	Precision↑
256	HyperPose (ours)	7.5	0.30	0.53
512	GIRAFFE-HD [72]	13.4	0.23	0.61
	HyperPose (ours)	7.7	0.27	0.63

while preserving its ability to judge image realism. Additionally, the proposed non-match loss strengthens supervision by encouraging more discriminative pose embeddings, leading to improved 3D configuration understanding.

Evaluation with high resolution To verify that our algorithm performs well on higher-resolution images, we test our algorithms on the AFHQ dataset with the resolution of 512^2 . Table 7 presents that our methods still significantly outperform the previous one on the 512^2 resolution, similar to the 256^2 resolution setting in Table 4. For the qualitative results, please check our supplementary document.

5. Conclusion

We present HyperPose, a novel algorithm for 3D-aware GANs designed to handle more challenging in-the-wild datasets without relying on pose label or strong 3D priors. To achieve this, we introduce hyper-pose embeddings, trained via a contrastive learning scheme that incorporates a pose disentanglement technique to prevent spurious alignment across generated scenes. For robust disentanglement and to specifically address the continuous nature of camera poses, we further propose a soft contrastive loss and non-match loss. Despite its architectural simplicity, the HyperPose proves highly effective in learning both 3D configurations and 2D image distributions on the various datasets.

Acknowledgements This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants [RS-2025-25442338, AI star Fellowship Support Program (Seoul National University); RS-2022-II220959 (No.2022-0-00959), (Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making; No.RS-2021-II212068, AI Innovation Hub (AI Institute, Seoul National University); No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] funded by the Korea government (MSIT).

References

- [1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *CVPR*, 2023. 3
- [2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2
- [3] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Josh Susskind. Gaudi: A neural architect for immersive 3d scene generation. In *NeurIPS*, 2022. 1, 3
- [4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2
- [5] 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, 4
- [6] 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. 4, 6
- [7] 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. 2, 4
- [8] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *ICCV*, 2023. 3
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3
- [10] Xingyu Chen, Yu Deng, and Baoyuan Wang. Mimic3d: Thriving 3d-aware gans via 3d-to-2d imitation. In *ICCV*, 2023. 2
- [11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 2
- [12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2, 6
- [13] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022. 2, 4, 6
- [14] Terrance Devries, Miguel Ángel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *ICCV*, 2021. 1, 3
- [15] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2
- [17] Artur Grigorev, Karim Isakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *CVPR*, 2021. 3
- [18] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *ICLR*, 2022. 2, 6
- [19] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. Dual contrastive learning for unsupervised image-to-image translation. In *CVPR*, 2021. 3
- [20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 5
- [21] 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 *NIPS*, 2017. 6
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2
- [23] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections, 2022. 1, 3
- [24] Tianyu Huang, Yihan Zeng, Zhilu Zhang, Wan Xu, Hang Xu, Songcen Xu, Rynson WH Lau, and Wangmeng Zuo. Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. In *CVPR*, 2024. 3
- [25] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022. 2
- [26] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *ICCV*, 2021. 2
- [27] Kyungmin Jo, Wonjoon Jin, Jaegul Choo, Hyunjoon Lee, and Sunghyun Cho. Sidegan: 3d-aware generative model for improved side-view image synthesis. In *ICCV*, 2023. 3
- [28] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. In *NeurIPS*, 2020. 3
- [29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

- [30] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020.
- [31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 4
- [32] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*, 2022. 2
- [33] 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 *CVPR*, 2022. 3
- [34] 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. 2
- [35] Yidi Li, Yiqun Wang, Zhengda Lu, and Jun Xiao. Depthgan: Gan-based depth generation of indoor scenes from semantic layouts. In *ICCV*, 2022. 1, 3
- [36] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdfsrn: Learning signed distance 3d object reconstruction from static images. In *NeurIPS*, 2020. 2
- [37] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 2
- [38] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 3
- [39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *TOG*, 2015. 3
- [40] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 2
- [41] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2
- [42] 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. In *ECCV*, 2020. 1, 2, 4
- [43] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *CVPR*, 2022. 2
- [44] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Peter Kotschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *CVPR*, 2023. 3
- [45] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *International Conference on 3D Vision (3DV)*, 2021. 2, 4, 6
- [46] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 6
- [47] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 1, 2
- [48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 3
- [49] 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. 2, 4
- [50] 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. 2, 4, 6
- [51] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 1, 2
- [52] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020. 3
- [53] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023. 3
- [54] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. 2, 6
- [55] Katja Schwarz, Seung Wook Kim, Jun Gao, Sanja Fidler, Andreas Geiger, and Karsten Kreis. WildFusion: Learning 3d-aware latent diffusion models in view space. In *ICLR*, 2024. 3
- [56] 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. 2
- [57] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *CVPR*, 2023. 3
- [58] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 2
- [59] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. MetaSDF: Meta-learning signed distance functions. In *NeurIPS*, 2020. 2
- [60] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. In *NeurIPS*, 2022. 2
- [61] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *ICLR*, 2023. 1, 3, 6
- [62] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3D: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. In *SIGGRAPH*, 2022. 2

- [63] 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. 2, 4, 6
- [64] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *ICCV*, 2023. 3
- [65] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *CVPR*, 2021. 2
- [66] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *CVPR*, 2022. 2
- [67] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. 6
- [68] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2
- [69] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3
- [70] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *ICCV*, 2023. 2
- [71] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *CVPR*, 2022. 2
- [72] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe-hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. 6, 8
- [73] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 2
- [74] 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. 6
- [75] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 2
- [76] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, 2021. 3
- [77] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: A 3d generative model for animatable human avatars. *arXiv preprint arXiv:2208.00561*, 2022. 1, 3
- [78] 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 *CVPR*, 2022. 2
- [79] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *CVPR*, 2022. 2
- [80] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020. 3