

ContraNeRF: 3D-Aware Generative Model via Contrastive Learning with Unsupervised Implicit Pose Embedding

Mijeong Kim^{1*}¹ECE & ²IPAI, Seoul National University

{mijeong.kim, bhhan}@snu.ac.kr malfo.lee@kakaobrain.com

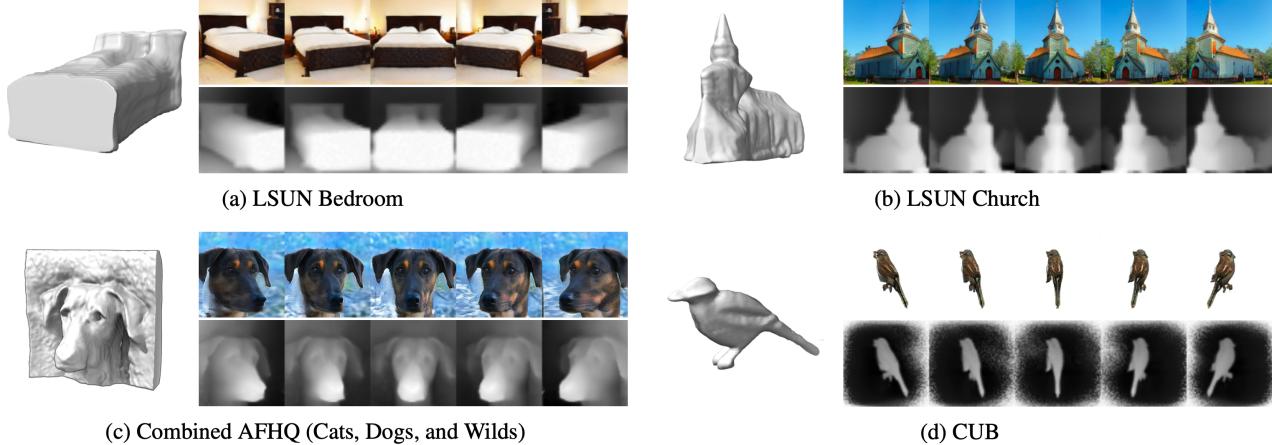
Hyunjoon Lee³Bohyung Han^{1,2}³Kakao Brain

Figure 1: Illustration of generated examples. Our 3D GAN enables synthesis of complex geometric scenes such as bedroom, church, animal faces, or birds, beyond simple geometries such as human face. Our approach trains from a collection of 2D images without ground-truth camera poses, depth information, target-specific shape priors, or multi-view supervision.

Abstract

Although 3D-aware GANs based on neural radiance fields have achieved competitive performance, their applicability is still limited to objects or scenes with the ground-truths or prediction models for clearly defined canonical camera poses. To extend the scope of applicable datasets, we propose a novel 3D-aware GAN optimization technique through contrastive learning with implicit pose embeddings. To this end, we first revise the discriminator design and remove dependency on ground-truth camera poses. Then, to capture complex and challenging 3D scene structures more effectively, we make the discriminator estimate a high-dimensional implicit pose embedding from a given image and perform contrastive learning on the pose embedding. The proposed approach can be employed for the dataset, where the canonical camera pose is ill-defined because it does not look up or estimate camera poses. Experi-

mental results show that our algorithm outperforms existing methods by large margins on the datasets with multiple object categories and inconsistent canonical camera poses.

1. Introduction

3D-aware Generative Adversarial Networks (GANs) aim to synthesize multiple views of a single scene with explicitly control of camera poses. Recent methods [4, 3, 28, 5, 33, 32, 39, 8, 13, 37, 38] incorporate the advances of neural radiance fields [34, 30, 27] into generative models [11, 20, 19] and reconstruct 3D scenes using a collection of 2D images without 3D shape priors. This technique allows us to predict not only 2D projected images but also their underlying 3D structures. However, they still suffer from the limited scope of target domains; most algorithms deal with only a few object categories, e.g., human or cat faces, where ground-truth or estimated camera poses are available and the canonical camera pose is well-defined. Al-

*This work was partly done during an internship at Kakao Brain.

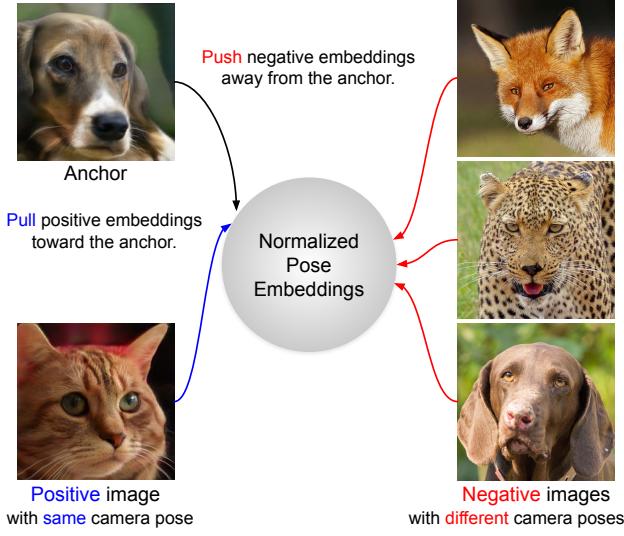


Figure 2: Illustration of the proposed contrastive learning on the pose embedding space. The ‘positive’ and ‘negative’ images denote images rendered in the same or different directions with the ‘anchor’ image, respectively. The distance between pose embeddings of positive pairs are learned to be closer than those of negative pairs.

though there exist some methods [24, 9, 1] that extend the scope of target domains to realistic ones with less geometric priors, they rely on additional geometric cues such as depth maps of training examples.

To alleviate the drawbacks of existing 3D-aware GAN approaches, we design a novel discriminator for representing the 3D structure of complex scenes in diverse domains without extra information. As an intermediate goal of our algorithm, we start from removing the dependency on ground-truth camera poses in the discriminator employed in previous work [4] and make the discriminator learn a camera pose regressor in a self-supervised way using generated images and their rendering poses. Such a simple approach turns out to be effective in synthesizing novel views without ground-truth camera poses of training images, but it sometimes fails to reconstruct 3D structures properly. We argue that the limitation is mainly due to explicit camera pose regression, which is not desirable to handle scenes or objects with complex and heterogeneous geometries.

To further improve the quality of complex geometric and photometric structures, we propose implicit camera pose embedding in a high-dimensional space for robust and comprehensive 3D reconstruction. For training, we employ a self-supervised contrastive learning [31], which captures rich geometric information of scenes from diverse pairwise relations of camera pose embeddings, improving camera pose regression and consequently enhancing 3D reconstruction quality without any ground-truths of camera poses.

Our experiments demonstrate that the proposed approach achieves state-of-the-art performance in both standard GAN evaluation and 3D reconstruction metrics without extra information. Our main contributions are summarized below:

- We present a simple yet effective camera pose representation method, implicit pose embedding, for training discriminators of 3D-aware GANs without ground-truth camera poses.
- We train the discriminator of 3D-aware GANs by contrastive learning, which allows our model to learn 3D structures of scenes with ill-defined canonical poses due to heterogeneous geometric configurations.
- Our framework achieves state-of-the-art performance on challenging benchmarks without any 3D related information and is validated via extensive experiments.

2. Related Work

We review existing approaches of GANs in 3D domain and discuss contrastive learning algorithms used in other generative tasks.

2.1. 3D-aware GANs

After the success of Generative Adversarial Network (GAN) [11, 20, 19, 21, 2, 45, 7] on generating high quality 2D images, several 3D-aware GANs [4, 3, 28, 5, 33, 32, 39, 8, 13, 37, 38] have been proposed to synthesize images based on 3D understanding instead of just image level understanding. By plugging the ideas of volume rendering and neural implicit representation techniques [34, 30, 27] into networks, 3D-aware GANs gain the capability of synthesizing multiple views of a single 3D scene. In addition, the networks are even trained to generate images with specific viewpoints using unorganized 2D image datasets—the same datasets used for 2D GANs. However, the domains of such datasets are limited; most 3D-aware GANs have shown successful examples only in a few object classes, including human and animal faces, cars, and a few synthetic object categories. Unlike existing 3D-aware GANs, we extend the domain range of the 3D-aware GAN framework to the complex scenes.

2.2. 3D-aware GANs on complex scenes

Some 3D-aware GANs have tackled generation task on the complex datasets which are composed of the images with diverse geometric configurations. However, existing approaches mostly rely on the prior knowledge of scenes such as object classes, ground-truths camera poses, or depth supervision. For example, full human body generation techniques [17, 12, 47] 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 [25]. On another direction, several algorithms learn to generate 3D indoor environments [9, 10, 1] and their generation processes are conditioned on camera pose information. These methods can synthesize new view synthesis of complex scenes in reasonable quality but require additional information such as ground truth camera poses or depth maps must be provided while training those networks. DepthGAN [24] also generates 3D indoor environments, but it utilizes the estimated depth maps given by the pre-trained depth estimation model [43] to obtain direct 3D information. However, our algorithm does not rely on explicit geometric information such as depth maps or ground-truths camera poses, and it can be generalized to various domains.

2.3. Contrastive learning

Contrastive learning is a widely used self-supervised representation learning schemes [15, 41, 6, 31]. Cntr-GAN [48] adds contrastive learning to train GANs together with image augmentations, where it serves as a regularizer to improve the fidelity of generation. Contrastive learning has also been used in image-to-image translation [35, 14, 23] and cross-modal translation [46] to enforce patch-wise correspondence and mutual information between image and text, respectively. Also, ContraGAN [18] proposes a class-conditional contrastive learning objective to increase the correlations between images of the same class. Unlike prior works, we are the first to adopt contrastive learning to 3D-aware GAN, employing it on the proposed implicit pose embeddings.

3. Preliminaries: EG3D

Our goal is to learn 3D-aware GANs on complex objects and scenes without prior knowledge or prediction models for camera poses of training examples. Since our algorithm relies on a state-of-the-art model, EG3D [4], we summarize the design of the generator $G(\cdot)$ and discriminator $D(\cdot)$ of EG3D.

3.1. Generator

Let p_z and p_ξ be the distributions of latent variable and camera pose, respectively. Given $z \sim p_z$ and $c \sim p_\xi$, the generator produces a 3D feature based on a tri-plane structure. The 3D feature is employed for rendering in the direction c , producing a low-resolution of 2D feature map and image. Then, an image super-resolution module synthesizes a high-resolution image from a given 2D feature map and a low-resolution image. In summary, the 3D-aware generator synthesizes a high-resolution image as

$$G : z, c \rightarrow I. \quad (1)$$

The generator $G(\cdot)$ can produce an image with different viewpoints of the same object, *i.e.*, using the identical z but

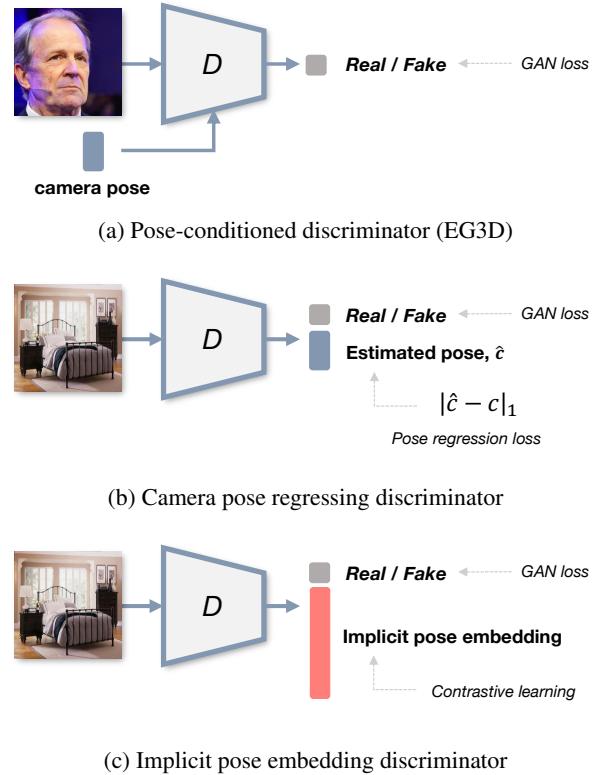


Figure 3: Comparison of different discriminator architectures. The pose-conditioned discriminator in (a) utilizes camera pose information as input, where ground truth pose should be given for each training image. On the other hand, (b) and (c) does not use such extra information, but instead, they additionally learn camera pose estimator explicitly or implicitly on rendered images. To this end, (b) uses direct pose regression loss with rendering camera pose c , while (c) employs contrastive learning to learn implicit pose embeddings. Note that our PRNeRF and ContraNeRF employ (b) and (c) as their discriminators, respectively.

different c 's.

3.2. Discriminator

The discriminator in 3D-aware GANs encourages the paired generator to draw realistic images given camera poses. To this end, EG3D [4] utilizes pose conditional discriminator as illustrated in Figure 3a. The discriminator takes both an image and a camera pose, and returns a logit as follows:

$$D : I, c \rightarrow l, \quad (2)$$

where $l \in \mathbb{R}$ is a logit for the standard GAN loss. Note that, following the design of EG3D, the discriminator takes both low-resolution and high-resolution image as its inputs. However, for the simplicity of the notations, we disregard

low-resolution image inputs from the equations for EG3D and our algorithm, which include (2), (3), and (6). Please refer to [4] for more detailed information.

3.3. Discussion

While EG3D [4] achieves competitive performance, it requires the ground-truth poses of training examples as inputs to the discriminator. This limitation significantly reduces the applicability of EG3D because camera poses are defined relatively with respect to a certain viewpoint and consequently ill-defined except for a few object categories with common-sense central poses such as faces. Other methods [28, 5, 33, 39, 32] trained without ground-truth camera poses typically yield incompetent performance and are evaluated only on less challenging datasets, *e.g.*, human faces. We propose a novel 3D-aware GAN algorithm that does not require camera pose labels but works well on images of natural scenes with heterogeneous geometric configurations.

4. Camera Pose Regression in Discriminator

This section describes our intermediate solution for training 3D-aware GAN models on top of EG3D without using camera pose labels of training data.

4.1. Discriminator design

To make the discriminator trainable without ground-truth camera poses, we first revise the original discriminator in EG3D [4] as illustrated in Figure 3b. Specifically, the new discriminator takes camera poses as its inputs no more while having additional output branch to predict the pose. The formal definition of the discriminator operation is given by

$$D : I \rightarrow l, \hat{c}, \quad (3)$$

where $l \in \mathbb{R}$ is a logit for the standard GAN loss and $\hat{c} \in \mathbb{R}^2$ is an estimated pitch and yaw for camera pose of an input image I . For implementation, we remove the pose conditional module in the discriminator of EG3D [4] and modify the dimension of the last fully connected layer. Note that the generator has an identical architecture with EG3D [4].

4.2. Pose regression loss

The estimated \hat{c} is optimized by the pose regression loss, which encourages the generator to synthesize the images congruent to a given camera direction c , which is given by

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

where $\|\cdot\|$ denotes ℓ_1 or ℓ_2 norm. This loss is operated only on fake images because their true camera poses, denoted by c 's, are always available as rendering direction while we do not have ground-truths of real images.

4.3. Overall objective

To enable the generator to learn the real data distribution, we use unsaturated adversarial losses with R1 regularization [26]. On top of the standard GAN loss, the pose regression loss is employed to provide 3D awareness with the model as follows:

$$\begin{aligned} \mathcal{L}(D, G) = & \mathbb{E}_{I \sim p_{\text{data}}} [f(-D(I) + \lambda \|\nabla D(I)\|^2)] \\ & + \mathbb{E}_{z \sim p_z, c \sim p_{\xi}} [f(D(G(z, c)))] \\ & + \lambda_{\text{pose}} \cdot \mathcal{L}_{\text{pose}}, \end{aligned} \quad (5)$$

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

5. Contrastive Learning in Discriminator

We now discuss the formulation and optimization of our final model, ContraNeRF, for 3D-aware generative model via contrastive learning.

5.1. Motivation

Although the pose regression loss discussed in the previous section is effective in terms of generated image quality, it often suffers from a lack of fidelity in reconstructing the underlying 3D structures. We build a novel discriminator based on an implicit pose embedding in a high-dimensional space and train the network using a new loss based on pairwise relations of implicit camera poses in a mini-batch. To be specific, we maximize the similarity between the implicit pose embeddings of the images with the same camera pose while minimizing the rest of the embedding pairs. This strategy is helpful for encoding camera poses by estimating the underlying scene structures via rich and flexible relations between many implicit pose embedding pairs. Note that the pose regression loss is now inaccessible due to the use of implicit camera pose embedding but our approach still free from the ground-truth camera poses.

5.2. Implicit pose embedding discriminator

The proposed discriminator has a similar architecture as the network given by (3). The only difference is that, instead of extracting a two-dimensional explicit camera pose from the input image I , the new model estimates an high-dimensional implicit camera pose embedding as follows:

$$D : I \rightarrow l, v, \quad (6)$$

where $l \in \mathbb{R}$ is a logit for the standard GAN loss and $v \in \mathbb{R}^m$ is an implicit pose embedding of the input image after ℓ_2 normalization. We set the dimensionality of v sufficiently high, *i.e.* $m \gg 2$, to make the implicit pose embedding vector more expressive than the typical camera pose representation based on yaw and pitch. Figure 3 illustrates our discriminator in comparison to other options. The other parts of the discriminator are identical to EG3D [4].

5.3. Mutual information maximization

The idea of contrastive learning is to train a network to keep the representation of anchor images close to the representation of relevant positive images while pushing it away from those of many mismatched negative images. Our goal is to make the synthesized images rendered on the same camera pose strongly associated with each other rather than the images generated by different poses as shown in Figure 2. In this respect, we employ contrastive learning on the pose embedding in the discriminator, which aims to maximize the mutual information between synthesized images with the same camera pose.

Positive and negative examples Given an anchor image $I^a = G(z^a, c^a)$, a positive image I^+ and a negative image I^- are defined as follows:

$$I^+ \in \mathcal{I}^+ = \{I = G(z, c) | z \sim p_z, c = c^a\} \quad (7)$$

$$I^- \in \mathcal{I}^- = \{I = G(z, c) | z \sim p_z, c \sim p_\xi, c \neq c^a\}. \quad (8)$$

A positive image is an example that is rendered in the same direction but may be generated using a different latent vector from the anchor image. In contrast, a negative image is the one rendered with a different camera pose from the anchor image. Then, the implicit pose embedding of an anchor v^a , its positive embedding v^+ and negative embedding v^- are given by

$$\begin{aligned} l^a, v^a &= D(I^a) \\ l^+, v^+ &= D(I^+) \\ l^-, v^- &= D(I^-), \end{aligned} \quad (9)$$

where l^a, l^+ and l^- are logits for the standard GAN loss.

Contrastive loss We adopt the InfoNCE loss [31] for contrastive learning of the implicit pose embedding. Let v_i^a , v_i^+ , and $v_{i,j}^a$ be camera pose embeddings in a mini-batch, where v_i^+ is a positive pair with the same camera pose of an anchor image as v_i^a , $i \in \{1, \dots, N\}$, while $v_{i,j}^-$, $j \in \{1, \dots, S\}$, is a negative pair with a different pose from v_i^a . We denote a collection of the negative examples for each anchor by v_i^- , i.e., $v_{i,j}^- \in v^-$. Given v_i^a , v_i^+ , and v_i^- , we obtain the following contrastive loss term:

$$\begin{aligned} \ell(v_i^a, v_i^+, v_i^-) &= \\ -\log \left(\frac{\exp(d(v_i^a, v_i^+)/\tau)}{\exp(d(v_i^a, v_i^+)/\tau) + \sum_{j=1}^S \exp(d(v_i^a, v_{i,j}^-)/\tau)} \right), \end{aligned} \quad (10)$$

where $d(u, v) = u^\top v / \|u\| \|v\|$ denotes the cosine similarity between u and v . This loss enforces the synthesized image to be similar as the rendered images in the same camera

viewpoint but dissimilar to those in other camera directions. To sum up, the overall contrastive loss is given by

$$\mathcal{L}_{\text{InfoNCE}} = \mathbb{E}_{z \sim p_z, c \sim p_\xi} \ell(v^a, v^+, v^-), \quad (11)$$

where v^+ and v^- are defined for each anchor, v^a .

5.4. Overall Objective

The final objective function of our algorithm is given by replacing the pose regression term in (5) by the InfoNCE loss term proposed for contrastive learning, as follows:

$$\begin{aligned} \mathcal{L}(D, G) &= \mathbb{E}_{I \sim p_{\text{data}}} [f(-D(I) + \lambda \|\nabla D(I)\|^2)] \\ &\quad + \mathbb{E}_{z \sim p_z, c \sim p_\xi} [f(D(G(z, c))] \\ &\quad + \lambda_{\text{pose}} \cdot \mathcal{L}_{\text{InfoNCE}}, \end{aligned} \quad (12)$$

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

6. Experiments

This section describes our benchmarks with complex geometric structures and reports the performance of our methods compared to previous ones quantitatively and qualitatively. We referred to our two models, one with camera pose regression and the other with contrastive learning, as PRNeRF and ContraNeRF, respectively.

6.1. Datasets and Settings

We report results on four different image datasets, LSUN Bedroom [44], LSUN Church [44], AFHQ (Animal Faces-HQ) [7], CUB [40]. These datasets are challenging for 3D-aware GANs, where the canonical pose is hard to define on the LSUN datasets, and AFHQ and CUB datasets contain complex and diverse geometric shapes. For AFHQ, we compute low-resolution features and images at a resolution of 32^2 with a total of 96 depth samples per ray. The final images are generated at a resolution of 256^2 . The resolution of feature maps and final images for the other datasets is set to 32^2 and 128^2 , respectively.

6.2. Results

Several ablations and analyses are performed to justify our contributions and proposed modules. For image synthesis evaluation, we report Fréchet inception distance (FID) [16] and Precision & Recall, which measure the fidelity and diversity of generated samples. For the evaluation of 3D reconstruction quality, we visualize rendered depth images along with their Depth FID, which measures FID between the estimated depth maps of training images given by a depth estimation model [43] and the rendered depth maps from the generated images. We also provide the quality of depth in rendered images using three subjective levels: **Bad**, **Fair**, and **Good**.

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

Method	FID ↓	Precision ↑	Recall ↑	Depth FID ↓	3D
GRAF [37]	70.71	0.42	0.00	97.41	Bad
π -GAN [5]	56.32	0.44	0.11	124.10	Bad
GIRAFFE [29]	42.78	0.55	0.02	145.63	Bad
PRNeRF (ours)	14.97	0.55	0.19	110.42	Bad
ContraNeRF (ours)	15.31	0.54	0.15	49.30	Good

6.2.1 LSUN Bedroom

We compare PRNeRF and ContraNeRF with GRAF [37], GIRAFFE [29], π -GAN [5], and DepthGAN [24] on the LSUN bedroom dataset. Figure 4a illustrates the generated images and their depth maps from three different viewpoints. Most algorithms including π -GAN [29], GIRAFFE [37], and PRNeRF produce unrealistic depth maps, where their generated images are almost identical and have planer depth maps from all viewpoints. In contrast, ContraNeRF generates RGB images and depth maps that reflect true 3D scene structures faithfully. Table 1 presents overall quantitative results, where ContraNeRF outperforms other algorithms in terms of Depth FID with considerable margins. It indicates that the synthesized 3D scenes given by ContraNeRF reflect true geometries effectively. Within our methods, although PRNeRF outperforms ContraNeRF in terms of 2D image synthesis metrics, it struggles with learning 3D structures accurately.

6.2.2 LSUN Church

Our models are compared with GRAF [37], GIRAFFE [29], π -GAN [5], and GIRAFFE-HD [42] on the LSUN church dataset. Again, our models outperform previous models in all metrics, as shown in Table 2. Figure 4b illustrates output examples from our models, where only ContraNeRF produces reasonable 3D scene structures and images. Similar to LSUN Bedroom, although PRNeRF shows a better 2D image synthesis quality within our methods, it fails to capture realistic 3D information in the scene.

6.2.3 AFHQ: Cats, Dogs, Wildlifes

AFHQ includes three categories for the animal face of *cats*, *dogs*, and *wildlife*, and we merge the three subsets and construct a single dataset having more diverse object shapes. Figure 4c presents the images of animal faces generated from different viewpoints, and Table 3 summarizes the experimental results on AFHQ. Both PRNeRF and ContraNeRF exhibit reasonable 3D-aware image synthesis performance, but ContraNeRF works slightly better in all evaluation metrics. PRNeRF does not produce planar depth maps,

Table 2: Quantitative comparison on the LSUN Church dataset. Our models outperform all other existing methods by significant margins in all image quality metrics, where ContraNeRF achieves the best performance. The dagger (\dagger) denotes that the scores are taken from GIRAFFE-HD [42].

Method	FID ↓	Precision ↑	Recall ↑	3D
GRAF [37]	91.11	0.53	0.00	Bad
π -GAN [5]	56.80	0.49	0.18	Bad
GIRAFFE [29]	38.36	0.51	0.02	Bad
GIRAFFE-HD [42] \dagger	10.28	—	—	—
PRNeRF (ours)	5.48	0.58	0.40	Bad
ContraNeRF (ours)	5.94	0.61	0.36	Good

Table 3: Quantitative comparison on the unified AFHQ dataset, including the Cats, Dogs, and Wilds categories with 256^2 resolution. PRNeRF and ContraNeRF synthesize high-fidelity images capturing accurate 3D geometry, where ContraNeRF shows the best performance. The dagger (\dagger) denotes that the scores are taken from StyleNeRF [13].

Method	FID ↓	Precision ↑	Recall ↑	3D
GRAF [37]	107.00	0.35	0.00	Bad
π -GAN [5]	48.43	0.41	0.12	Bad
GIRAFFE [29]	31.31	0.51	0.04	Fair
GIRAFFE-HD [42]	14.21	0.55	0.10	—
StyleNeRF [13] \dagger	14	—	—	Good
PRNeRF (ours)	9.14	0.54	0.19	Good
ContraNeRF (ours)	9.03	0.55	0.21	Good

Table 4: Quantitative comparison on the CUB dataset, having large object pose variations. ContraNeRF significantly outperforms existing methods in all image quality metrics.

Method	FID ↓	Precision ↑	Recall ↑	3D
GRAF [37]	46.31	0.67	0.09	Bad
GIRAFFE [29]	49.34	0.68	0.04	Bad
π -GAN [5]	48.82	0.64	0.10	Bad
GIRAFFE-HD [42]	24.31	0.67	0.17	—
PRNeRF (ours)	48.34	0.65	0.08	Bad
ContraNeRF (ours)	17.65	0.71	0.24	Good

unlike the LSUN datasets, because the geometric structures of the AFHQ dataset are more straightforward.

6.2.4 CUB

We evaluate our framework in a more challenging dataset, CUB, which consists of images with large object pose variations. Table 4 presents that ContraNeRF still outperforms others by large margins, offering the lowest FID values. This is because our pose embedding captures rich geometric information of scenes from diverse pairwise relations of camera pose embeddings. As illustrated in Figure 4d, the quality of the depth maps estimated by ContraNeRF looks impressive, and we notice that all the compared algorithms mostly fail to reconstruct 3D structures accurately.

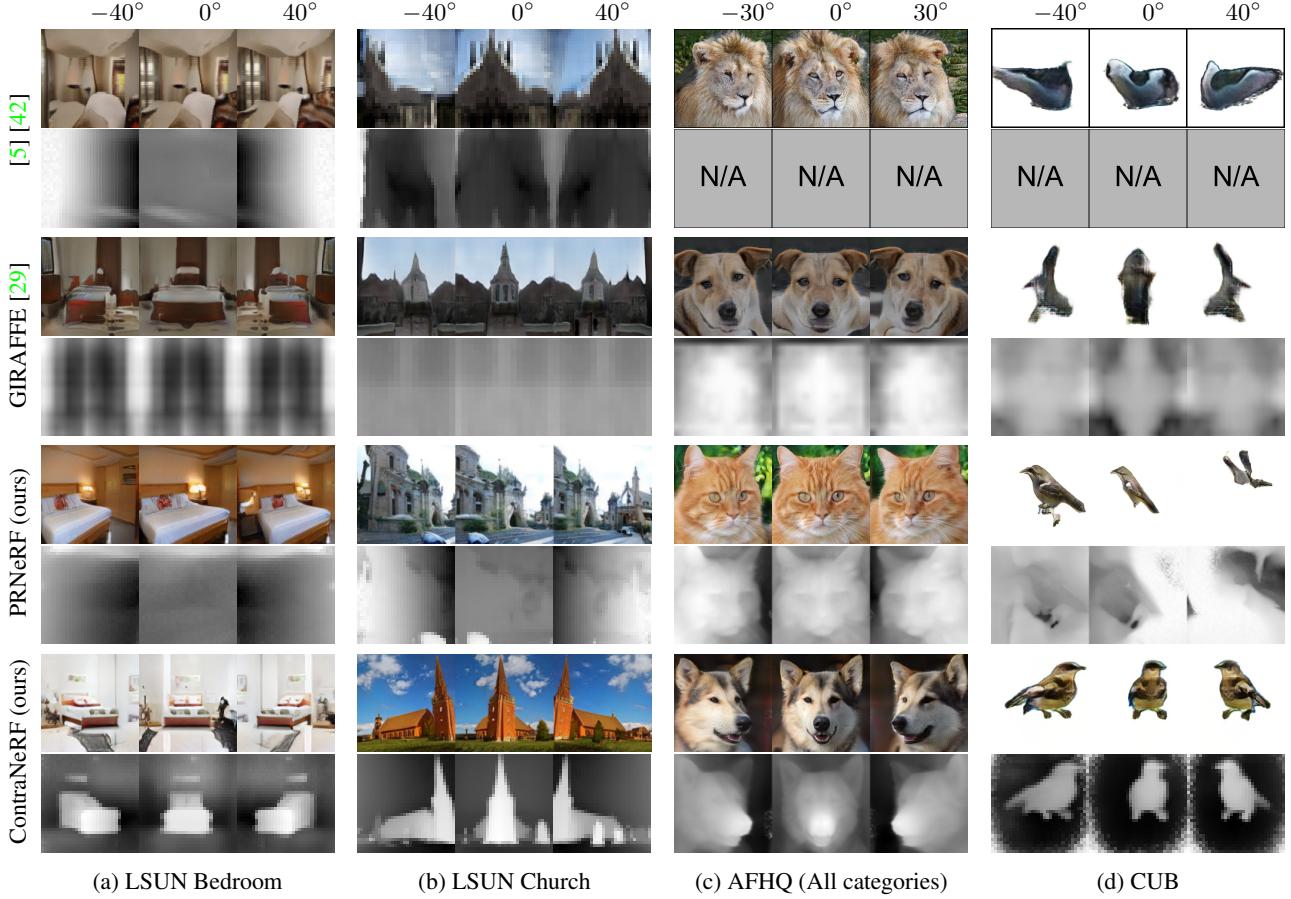


Figure 4: Comparing samples of ContraNeRF, PRNeRF, and modern 3D-aware GANs. We visualized the RGB image and depth map by rotating the rendering pose horizontally. For the first row, the first two results are from π -GAN [5], and the rest is from GIRAFFE-HD [42]. ContraNeRF produces high-fidelity images with accurate depth maps in all domains since its implicit pose embedding can capture complex geometries. However, others methods, including PRNeRF, usually produce planar depth maps with unrealistic 3D structure. For additional results, please refer to our supplementary materials.

6.3. Analysis

Combination of $\mathcal{L}_{\text{Pose}}$ and $\mathcal{L}_{\text{InfoNCE}}$ We evaluate the ensemble trained with the pose regression loss, $\mathcal{L}_{\text{Pose}}$, and the contrastive loss, $\mathcal{L}_{\text{InfoNCE}}$, on the FFHQ dataset [20]. Table 5 presents that the combination of the two losses achieves the best performance on FFHQ, except EG3D¹. Unlike other datasets, PRNeRF outperforms ContraNeRF in FFHQ, probably because the pose regression of PRNeRF is more straightforward in this dataset, having homogeneous geometry. However, ContraNeRF or its ensemble version always shows the best performance including the FFHQ dataset.

High resolution image synthesis To verify that our algorithm performs well on higher-resolution images, we test

¹EG3D exploits the ground-truth camera poses of the training set, which makes the direct comparison with EG3D unfair. For reference, EG3D can not be evaluated by other datasets used in this paper.

Table 5: Experiments on the FFHQ dataset with 256 resolution. The dagger (\dagger) denotes that the scores are taken from StyleNeRF [13] and EpiGRAF [38].

Method	GT pose	$\mathcal{L}_{\text{Pose}}$	$\mathcal{L}_{\text{InfoNCE}}$	FID \downarrow	Precision \uparrow	Recall \uparrow
EG3D [4]	✓	-	-	4.92	0.554	0.435
StyleNeRF [13] \dagger	-	-	-	8	-	-
EpiGRAF [38] \dagger	-	-	-	9.71	-	-
PRNeRF	-	✓	-	5.94	0.548	0.415
ContraNeRF	-	-	✓	6.85	0.552	0.405
PR-ContraNeRF	✓	✓	-	5.73	0.551	0.421

our algorithms on the AFHQ dataset with the resolution of 512^2 . Table 6 presents that our methods still significantly outperform the previous one on the 512^2 resolution, similar to the 256^2 resolution setting in Table 3.

Dimensionality of pose embedding We analyze the impacts of the dimensionality of pose embedding on 3D recon-

Table 6: Experiments on the AFHQ dataset with 512^2 resolution. ContraNeRF and PRNeRF still outperform the existing method in high resolution setting, where ContraNeRF achieves the best performance.

Method	FID↓.	Precision↑	Recall↑
GIRAFFE-HD [42]	13.42	0.61	0.23
PRNeRF	8.21	0.61	0.31
ContraNeRF	8.02	0.63	0.32

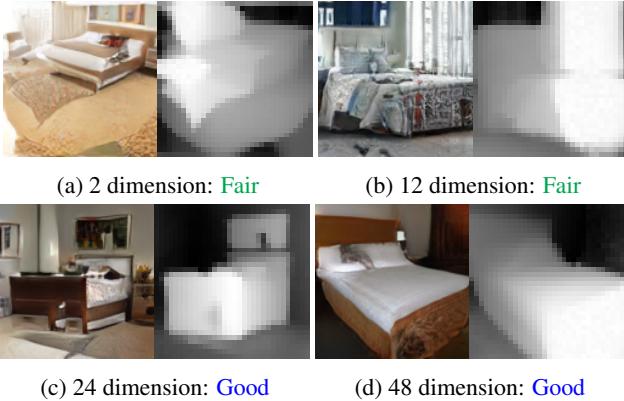


Figure 5: Effects of the pose embedding dimension on the quality of rendered image on the LSUN Bedroom dataset. Our framework successfully captures underlying 3D structures with a sufficient number of embedding dimensions.

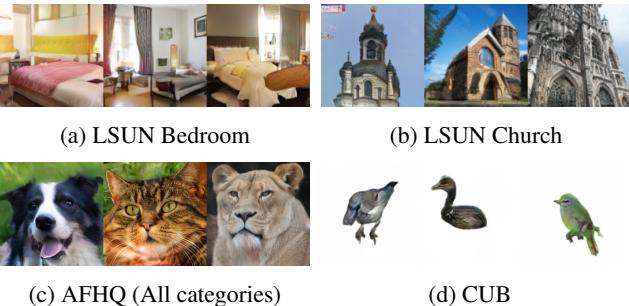


Figure 6: Qualitative results of rendering for the same pose with different latent vector z on ContraNeRF. ContraNeRF synthesizes diverse scenes with different geometry in the LSUN Bedroom and LSUN Church datasets while it produces images with identical geometry in the AFHQ dataset.

struction quality on the LSUN Bedroom dataset, and visualize its ablative results in Figure 5. Our framework successfully captures 3D structures if it has a sufficient number 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.

Handling dataset with diverse camera poses Figure 6 illustrates images rendered by ContraNeRF, from the same



Figure 7: Example of failure cases. ContraNeRF sometimes produces images with unrealistic geometries such as planar scenes. The first case (left) is the example that our algorithm generate translated images by varying camera poses, and the second one (right) illustrates the results with almost uniform depth maps.

camera pose but with different content latent vector z . The generated images on the AFHQ dataset have almost identical viewpoints, indicating that our contrastive learning works as previous ones for these simple cases. On the other hand, since the LSUN Bedroom, LSUN Church, and CUB have various scene geometries and object shapes, there is no canonical center pose applicable to all images, and the generated images do not have the same viewpoints perceptually. However, ContraNeRF successfully reconstructs the underlying 3D structure of the scenes as presented earlier, which shows the strength and potential of ContraNeRF for naturally handling datasets with images captured from heterogeneous viewpoints.

Failure cases Although ContraNeRF produces high-fidelity volumetric scenes in most cases, we observe some failure cases on the LSUN bedroom dataset. Figure 7 illustrates failure cases in which ContraNeRF produces planar scenes. We presume outlier training samples, such as watermarked images or images captured from out-of-distribution camera poses, may result in degenerate outputs.

7. Conclusion

By extending 3D-aware GANs to handle more diverse domains of objects and scenes, the proposed models improve their usability and expands the possible applications from face synthesis to 3D world modeling. To this end, we first show that a pose regression-based framework can be used to effectively remove the camera pose dependency in 3D-aware GAN training. We then propose a contrastive learning-based framework that uses implicit pose embeddings at higher dimensions for rich descriptions of pose information in natural scenes with diverse and complex geometries. The effectiveness of implicit pose embedding and contrast learning frameworks has been experimentally demonstrated through evaluation on multiple benchmark datasets.

References

- [1] 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, 2022. [2](#), [3](#)
- [2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018. [2](#)
- [3] 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. [1](#), [2](#)
- [4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. [1](#), [2](#), [3](#), [4](#), [7](#)
- [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 *CVPR*, 2021. [1](#), [2](#), [4](#), [6](#), [7](#)
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. [3](#)
- [7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. [2](#), [5](#), [11](#)
- [8] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022. [1](#), [2](#)
- [9] 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. [2](#), [3](#)
- [10] Marco Fraccaro, Danilo Jimenez Rezende, Yori Zwols, Alexander Pritzel, S. M. Ali Eslami, and Fabio Viola. Generative temporal models with spatial memory for partially observed environments, 2018. [3](#)
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. [1](#), [2](#)
- [12] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *CVPR*, 2021. [2](#)
- [13] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylererf: A style-based 3d-aware generator for high-resolution image synthesis. In *ICLR*, 2022. [1](#), [2](#), [6](#), [7](#)
- [14] Junlin Han, Mehrdad Shoeibi, Lars Petersson, and Mohammad Ali Armin. Dual contrastive learning for unsupervised image-to-image translation. In *JCVPR*, 2021. [3](#)
- [15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. [3](#)
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. [5](#)
- [17] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections, 2022. [2](#)
- [18] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. In *NeurIPS*, 2020. [3](#)
- [19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020. [1](#), [2](#)
- [20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. [1](#), [2](#), [7](#)
- [21] 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](#)
- [22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. [11](#)
- [23] 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](#)
- [24] Yidi Li, Yiqun Wang, Zhengda Lu, and Jun Xiao. Depthgan: Gan-based depth generation of indoor scenes from semantic layouts. In *ICCV*, 2022. [2](#), [3](#), [6](#)
- [25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *TOG*, 2015. [3](#)
- [26] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. [4](#)
- [27] 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](#)
- [28] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *International Conference on 3D Vision (3DV)*, 2021. [1](#), [2](#), [4](#)
- [29] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. [6](#), [7](#)
- [30] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. [1](#), [2](#)
- [31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. [2](#), [3](#), [5](#)
- [32] 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. [1](#), [2](#), [4](#)
- [33] 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. [1](#), [2](#), [4](#)

- [34] 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](#)
- [35] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020. [3](#)
- [36] 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. [11](#)
- [37] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. [1](#), [2](#), [6](#)
- [38] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. In *NeurIPS*, 2022. [1](#), [2](#), [7](#)
- [39] 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. [1](#), [2](#), [4](#)
- [40] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. [5](#), [11](#)
- [41] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. [3](#)
- [42] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe-hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. [6](#), [7](#), [8](#)
- [43] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, 2021. [3](#), [5](#)
- [44] 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. [5](#), [11](#)
- [45] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. [2](#)
- [46] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfai Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, 2021. [3](#)
- [47] 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, 2022. [2](#)
- [48] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020. [3](#)

A. Ensemble of $\mathcal{L}_{\text{Pose}}$ and $\mathcal{L}_{\text{InfoNCE}}$

We further evaluate the combination of pose regression loss, $\mathcal{L}_{\text{Pose}}$, and contrastive loss, $\mathcal{L}_{\text{InfoNCE}}$, on the AFHQ and LSUN Bedroom datasets, with comparative results presented in Table A. For the AFHQ dataset, the combined approach enhances 2D image synthesis metrics without compromising 3D reconstruction quality, though the differences are minor. Conversely, on the LSUN Bedroom dataset, the ensemble leads to a slight decline in 3D reconstruction quality, partly due to the pose regression loss being ill-suited for representing complex scene structures.

Table A: Ablative results of pose regression loss and contrastive loss on the AFHQ and LSUN Bedroom dataset.

Algorithms	$\mathcal{L}_{\text{Pose}}$	$\mathcal{L}_{\text{InfoNCE}}$	AFHQ				LSUN Bedroom			
			FID ↓	Precision ↑	Recall ↑	3D	FID ↓	Precision ↑	Recall ↑	DepthFID ↓
PRNeRF	✓		9.14	0.54	0.19	Good	14.97	0.55	0.19	110.42
ContraNeRF		✓	9.03	0.55	0.21	Good	15.31	0.54	0.15	49.30
PR-ContraNeRF	✓	✓	9.00	0.55	0.21	Good	15.29	0.55	0.16	55.77



Figure A: Qualitative results of the PR-ContraNeRF on the AFHQ dataset (left) and LSUN bedroom (right) dataset. The view angles of each scene are rotated at regular intervals: -30° , -15° , 0° , 15° , and 30° .

B. Implementation Details

Our implementation is based on PyTorch [36] and employs the Adam optimizer [22]. 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. Table B provides additional details, including the number of images and prior camera pose distribution, p_ξ , for each dataset.

Table B: The number of images and prior camera pose distribution p_ξ of the datasets.

Dataset	Number of Images	Pitch		Yaw	
		Distribution	Detail	Distribution	Detail
LSUN Bedroom [44]	3,033,042	Gaussian	$\mu = \pi/2, \sigma = 0.10$	Gaussian	$\mu = \pi/2, \sigma = 0.70$
LSUN Church [44]	126,227	–	$\pi/2$	Uniform	$[\pi/2 - 5\pi/18, \pi/2 + 5\pi/18]$
AFHQ [7]	14,630	Gaussian	$\mu = \pi/2, \sigma = 0.13$	Gaussian	$\mu = \pi/2, \sigma = 0.19$
CUB [40]	11,788	Gaussian	$\mu = \pi/2, \sigma = 0.13$	Uniform	$[\pi/2 - 3\pi/4, \pi/2 + 3\pi/4]$

C. Qualitative Evaluation with High Resolution

Figure B exhibits the results of ContraNeRF on the AFHQ dataset, including Cats, Dogs, and Wildlifes categories with 512^2 resolution. ContraNeRF produces high-resolution and high-fidelity images with accurate depth maps and surface.

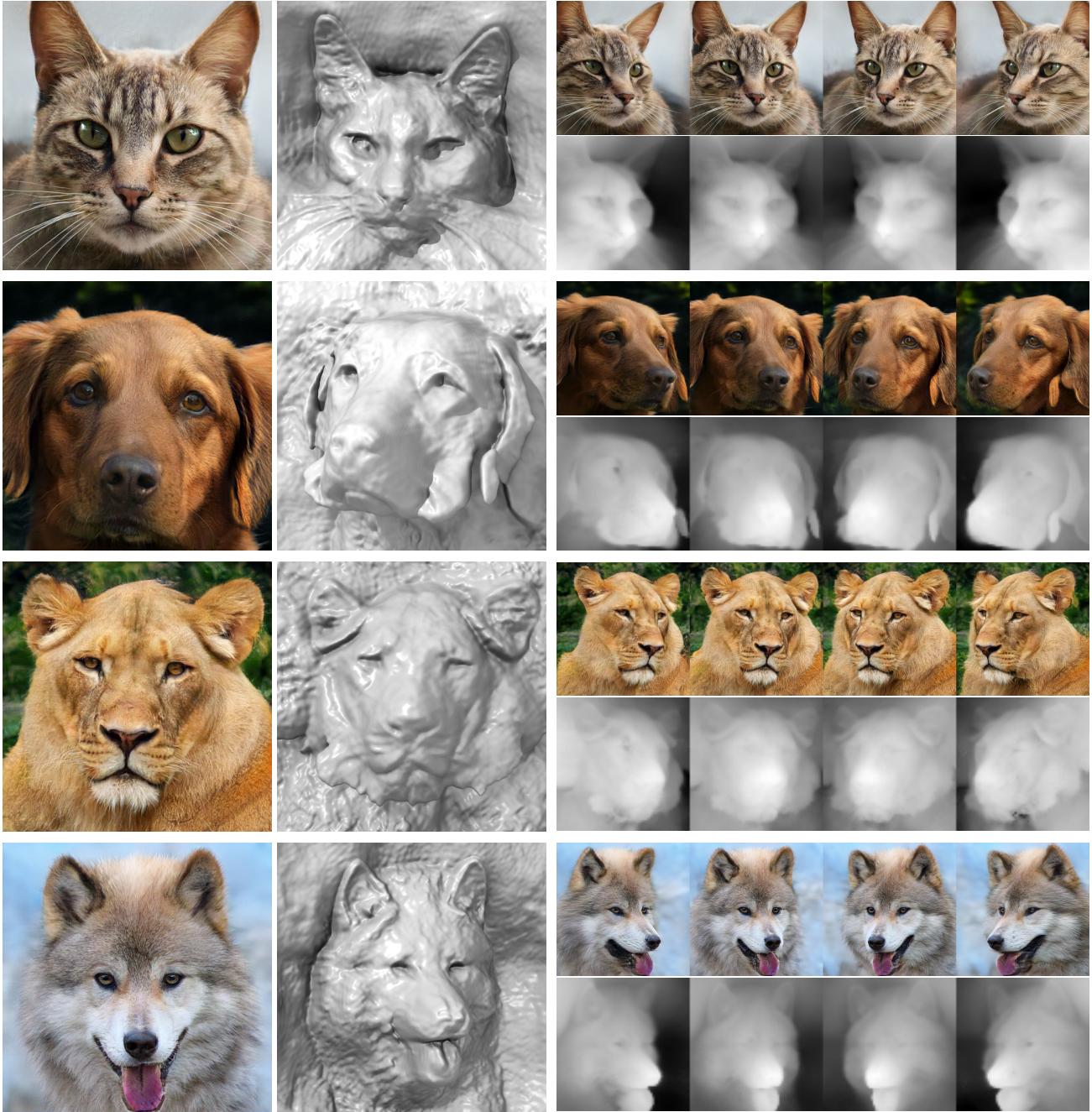
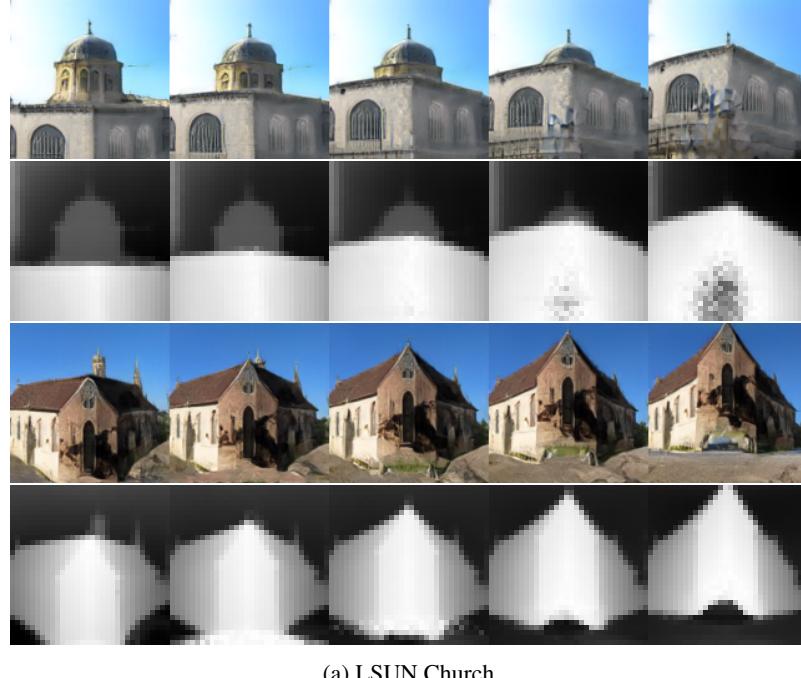


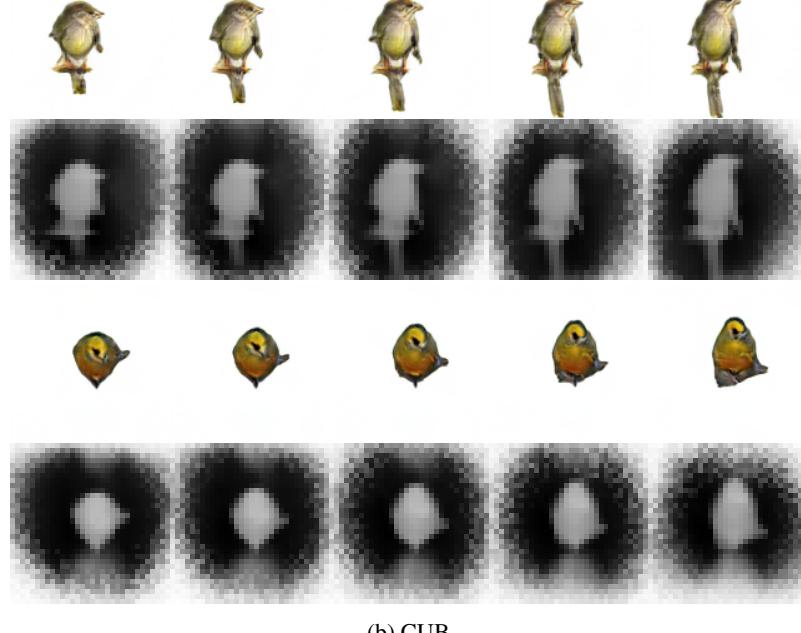
Figure B: Curated examples from ContraNeRF trained on the combined AFHQ (Cats, Dogs, and Wildlifes) with 512^2 resolution. The first column exhibits images rendered with the mean yaw and pitch value, and the second column shows surface renderings with random poses. The last column shows RGB images, and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

D. 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 ContraNeRF, we display supplementary qualitative results incorporating vertical rotations, as depicted in Figure C and D. 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.

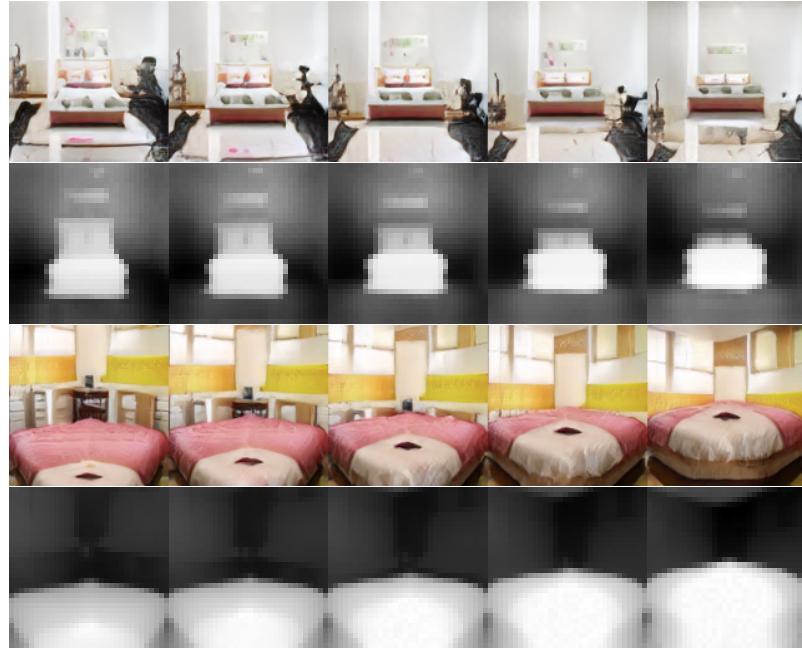


(a) LSUN Church

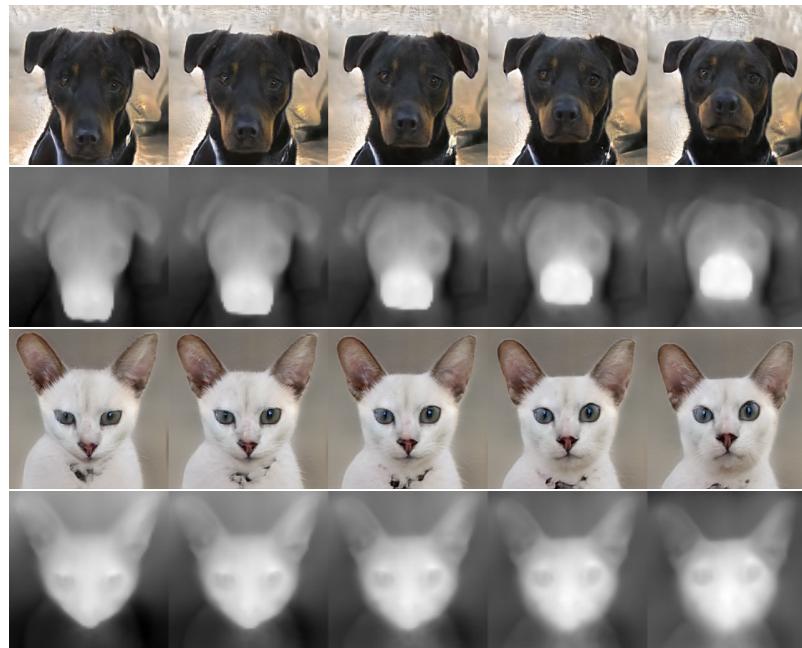


(b) CUB

Figure C: Qualitative results of ContraNeRF 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 D: Qualitative results of ContraNeRF 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° .

E. Qualitative Evaluation of 3D Reconstruction Results

We categorize the reconstruction quality of each scene into three levels: Bad, Fair, and Good. Each level generally exhibits the following characteristics:

- **Bad:** A model is unable to discern reasonable 3D structures, resulting in planar depth maps, as demonstrated in Figure E. This quality is frequently observed in PRNeRF’s outcomes on the LSUN Bedroom dataset.
- **Fair:** While a model generates volumetric scenes, the estimated depth maps contain coarse or partially inaccurate information about 3D structures, as illustrated in Figure F. ContraNeRF with low-dimensional camera pose embedding spaces (up to 12 dimensions) tends to produce this level of depth map quality.
- **Good:** A model produces high-quality depth maps that accurately represent the 3D scene structures, as shown in Figure G. ContraNeRF typically attains high-quality depth maps when sufficiently high-dimensional camera pose embeddings are used.

Note that the classification of 3D scene structure quality is evident for each algorithm when applied to a specific dataset, as depicted in Figure E, F, and G.

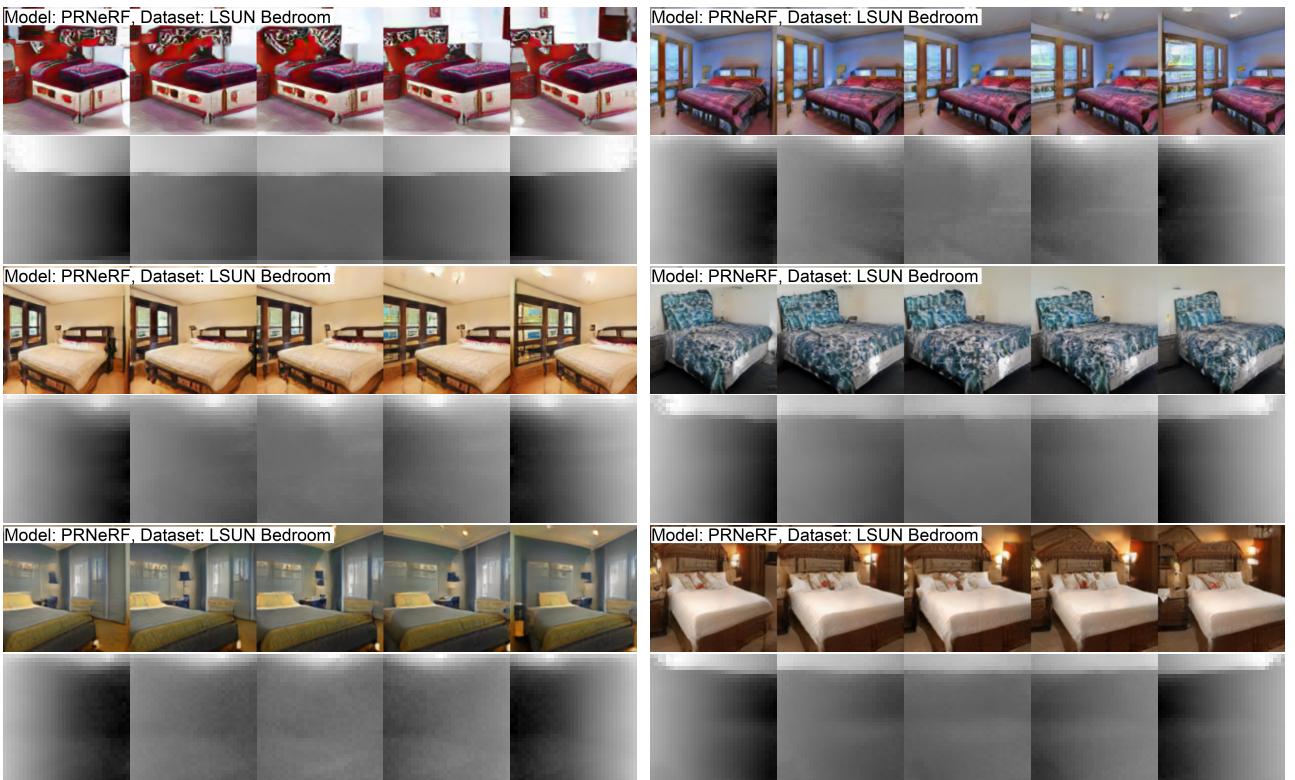


Figure E: Rendered images and their **Bad** depth maps. PRNeRF, when applied to the LSUN Bedroom dataset, produces images of this quality.

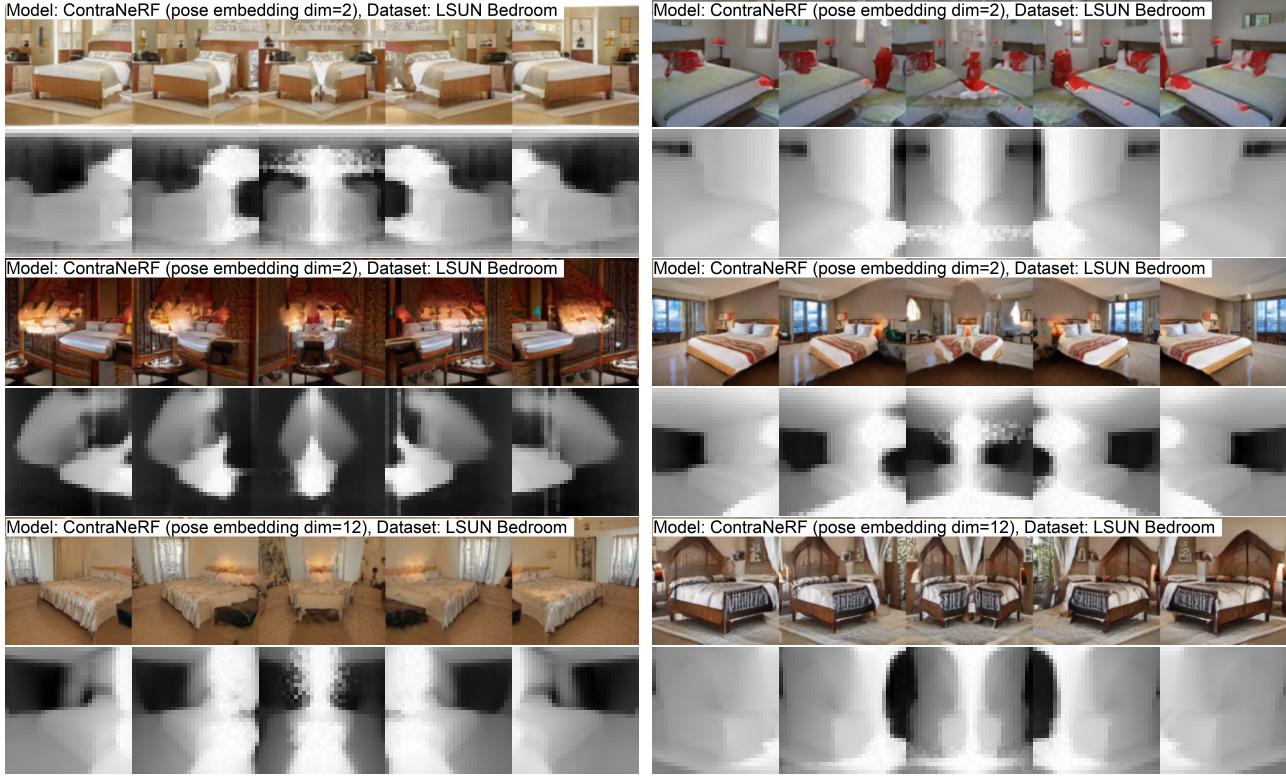


Figure F: Rendered images and their **Fair** depth maps. ContraNeRF with a low-dimensional camera pose embedding tends to generate images in the **Fair** quality of depth maps on the LSUN Bedroom dataset.

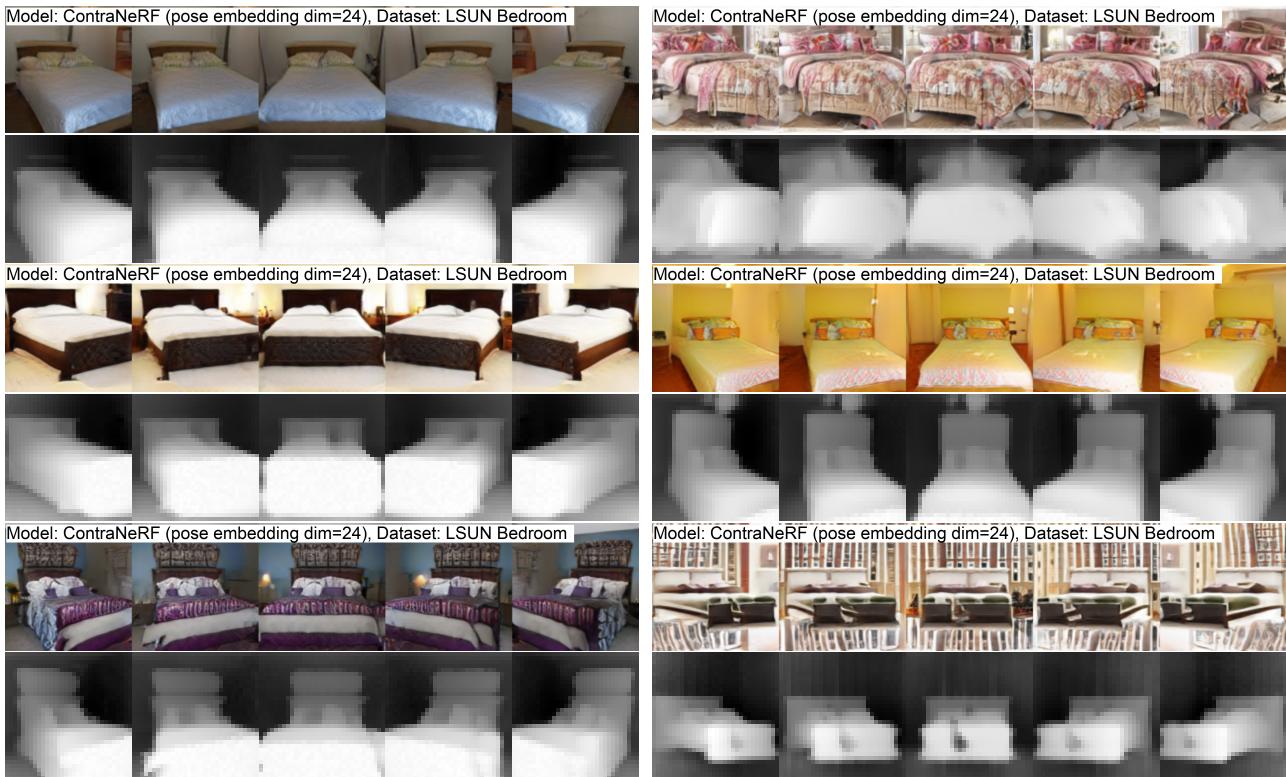


Figure G: Rendered images and their **Good** depth maps. ContraNeRF, when a high-dimensional camera pose embedding is utilized, typically produces images with **Good** quality depth maps.

F. 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.

F.1. LSUN Bedroom

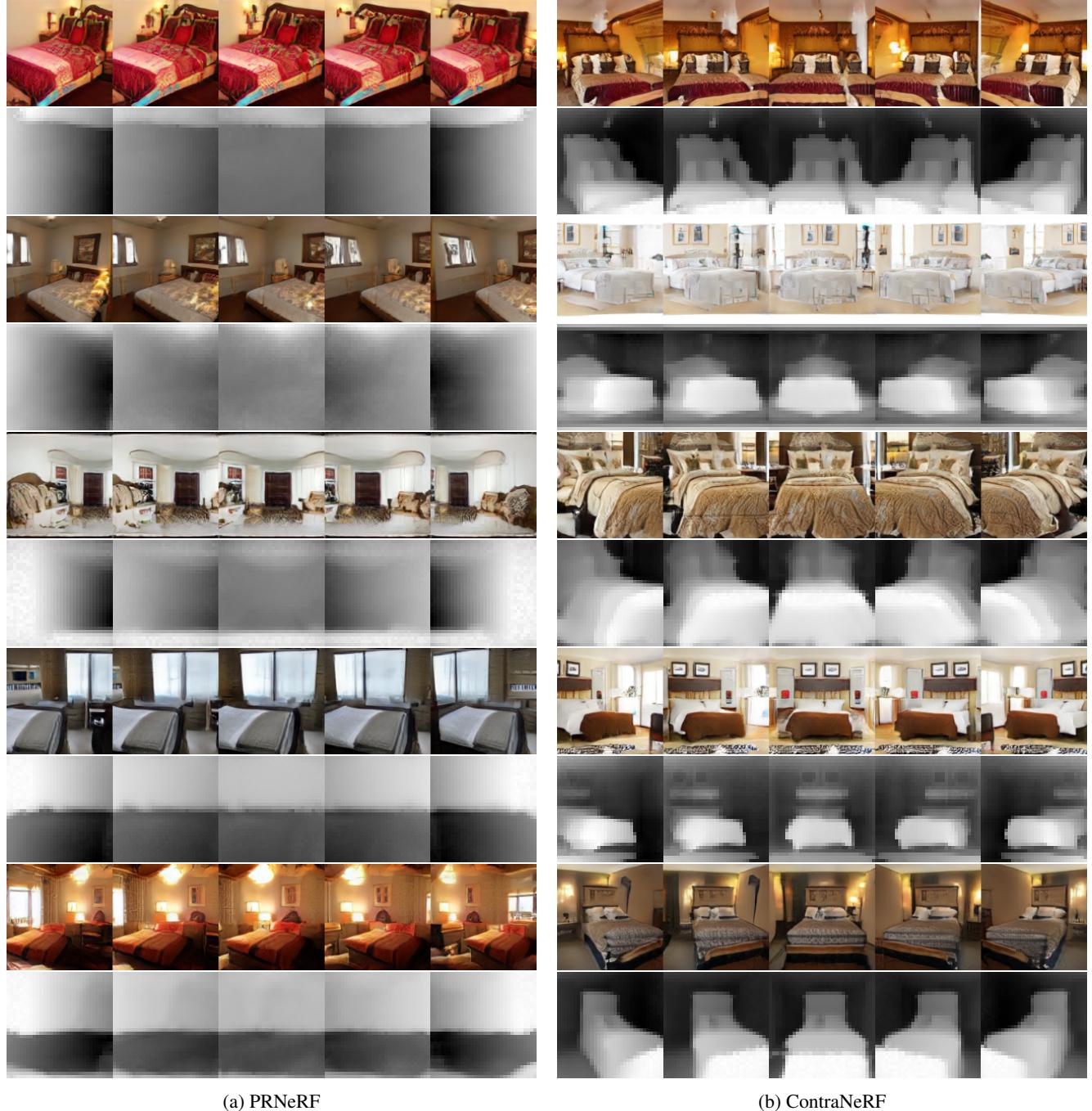


Figure H: Qualitative results of PRNeRF and ContraNeRF on the LSUN Bedroom dataset. We visualized the rendering results with RGB image and its depth map. The view angles of each scene are rotated at regular intervals: -40° , -20° , 0° , 20° , and 40° .

F.2. LSUN Church

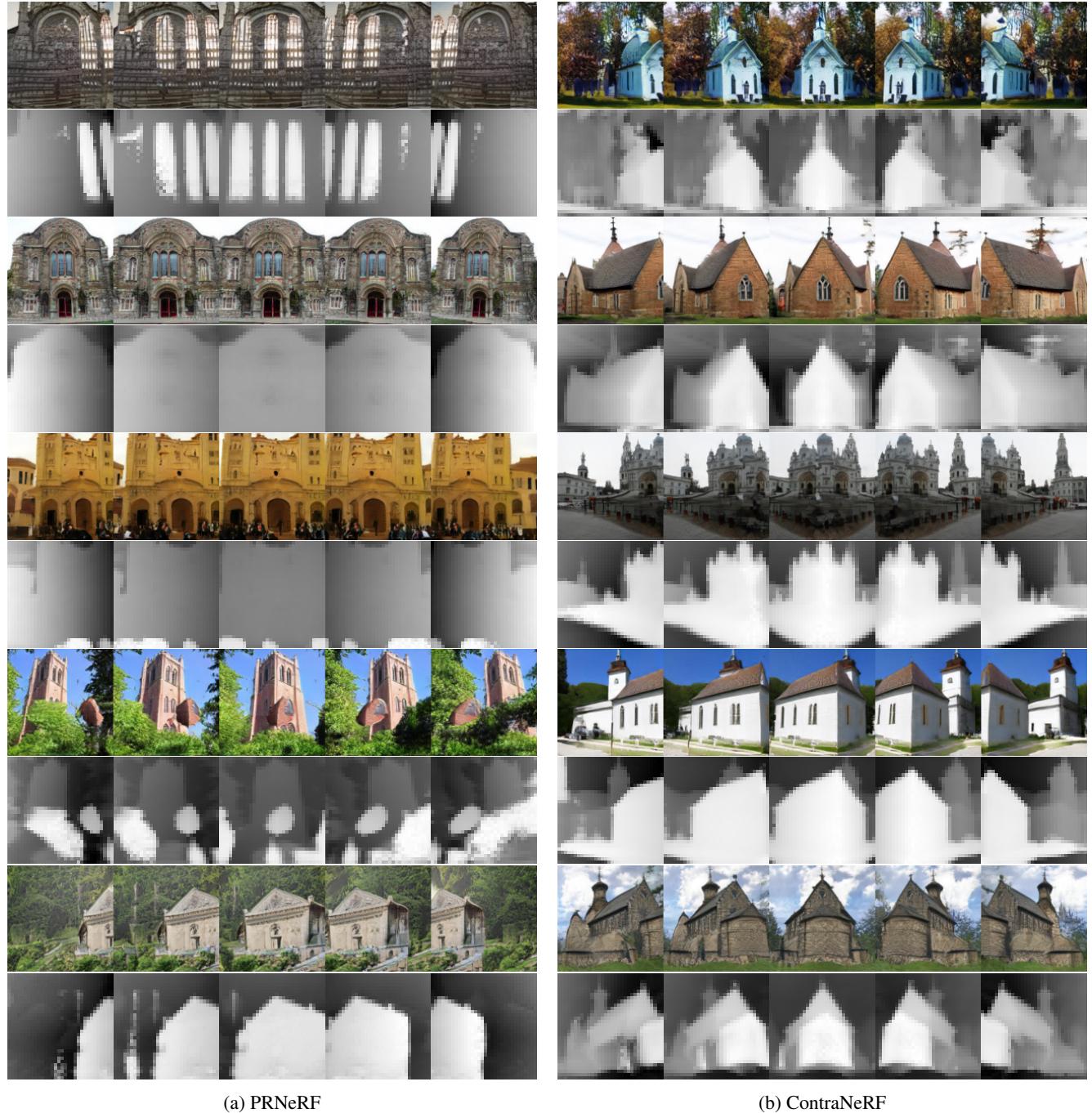
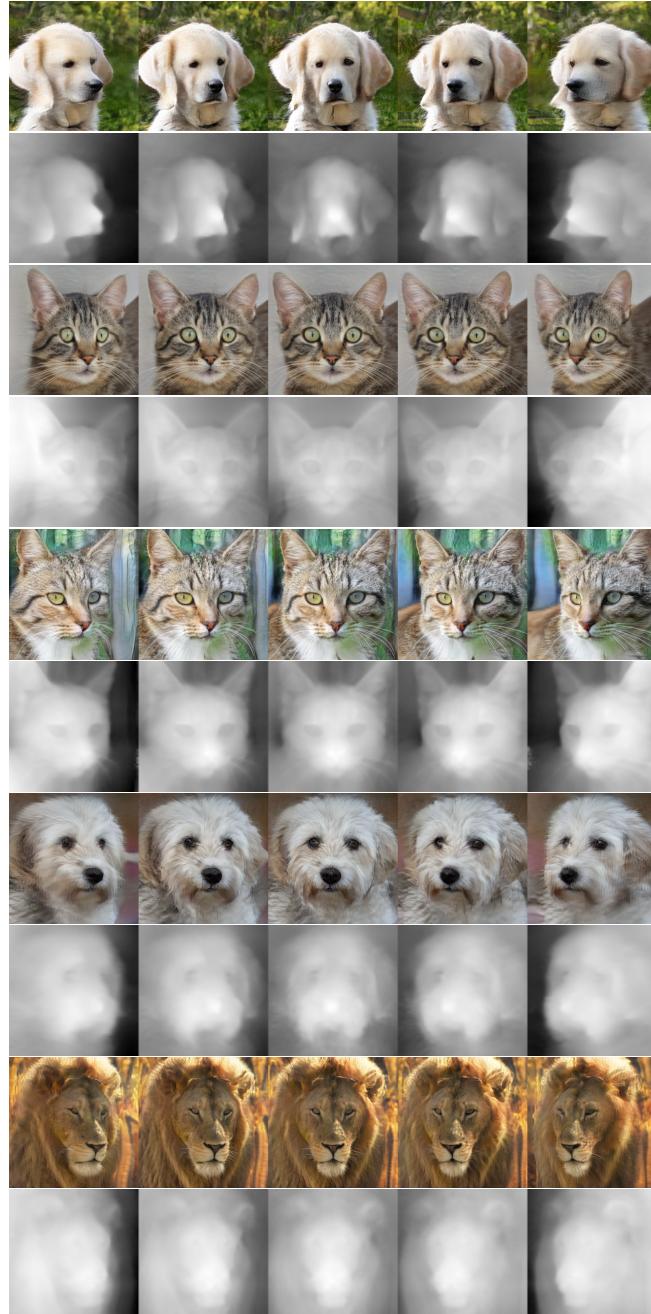
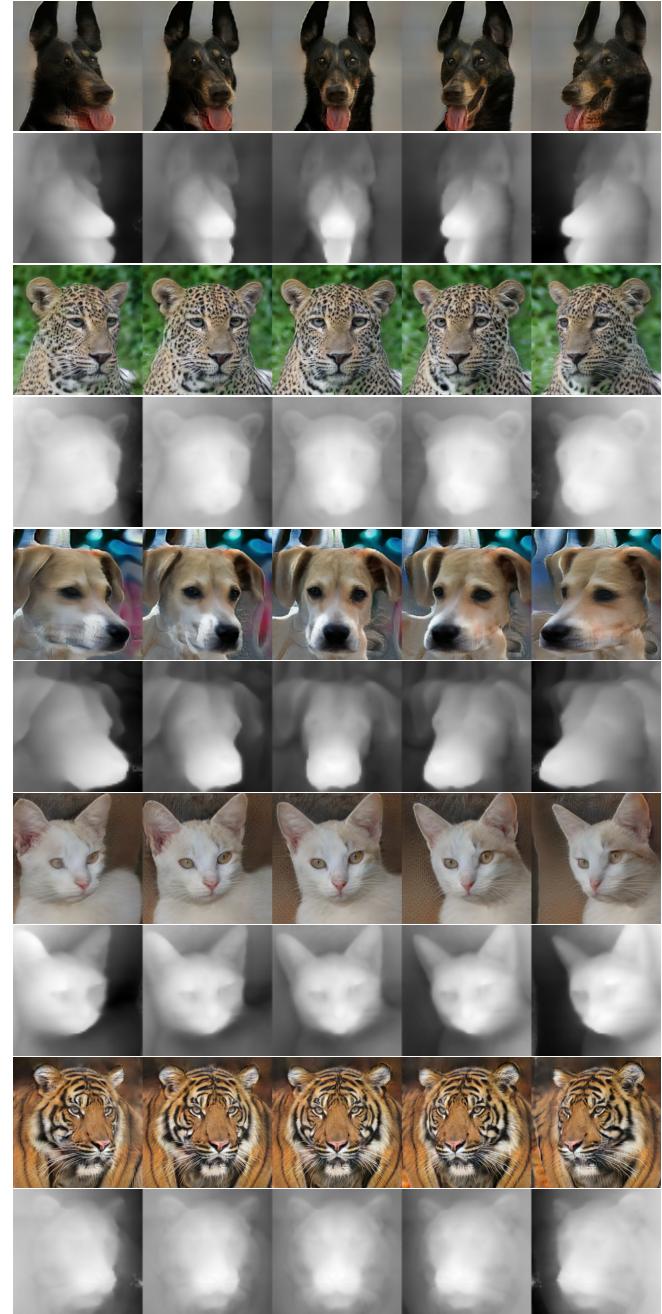


Figure I: Qualitative results of PRNeRF and ContraNeRF on the LSUN Church dataset. We visualized the rendering results with RGB image and its depth map. The view angles of each scene are rotated at regular intervals: -40° , -20° , 0° , 20° , and 40° .

E.3. AFHQ



(a) PRNeRF



(b) ContraNeRF

Figure J: Qualitative results of PRNeRF and ContraNeRF 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° .

F.4. CUB

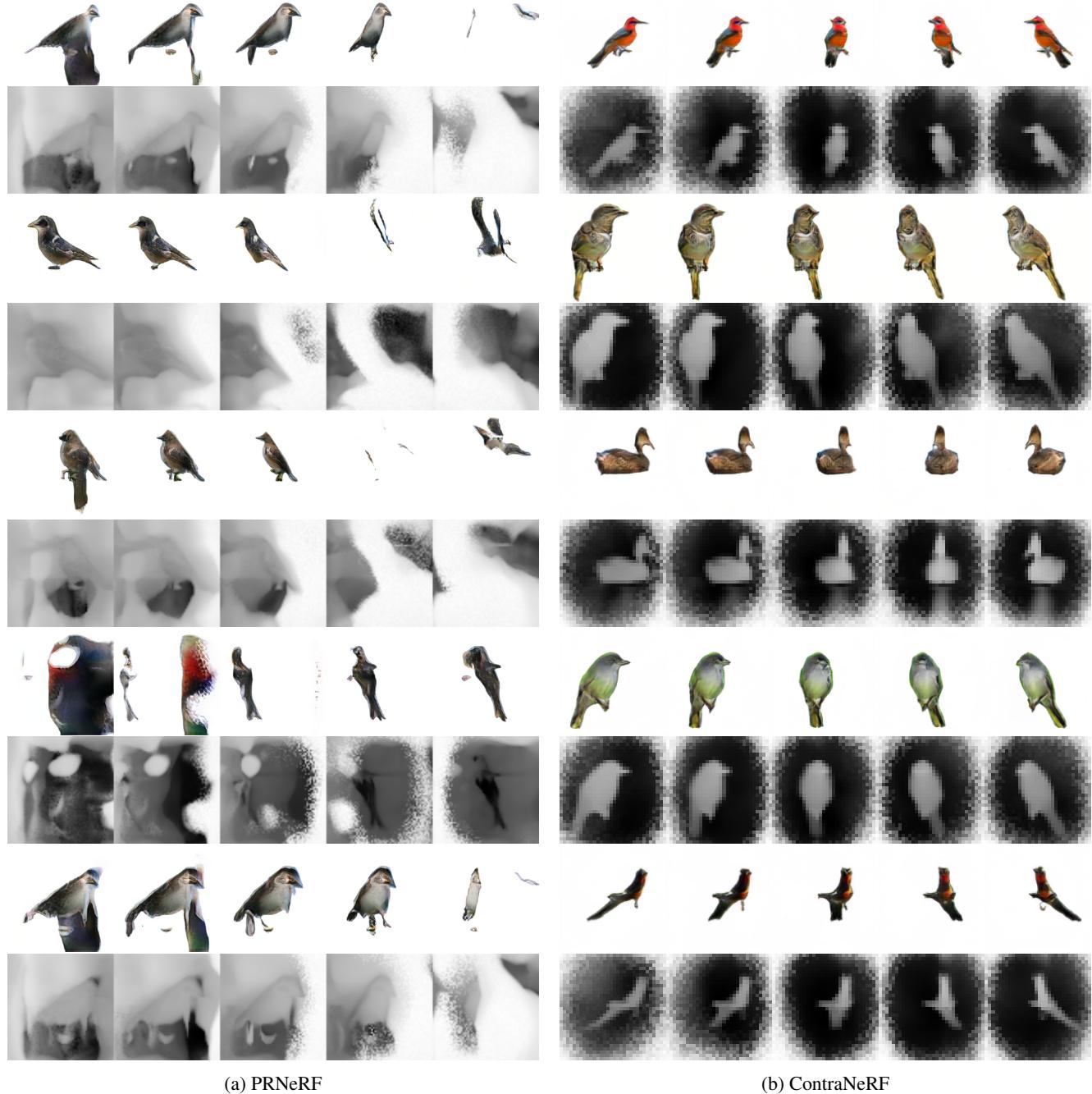


Figure K: Qualitative results of PRNeRF and ContraNeRF on the CUB dataset. We visualized the rendering results with RGB image and its depth map. The view angles of each scene are rotated at regular intervals: -40° , -20° , 0° , 20° , and 40° .