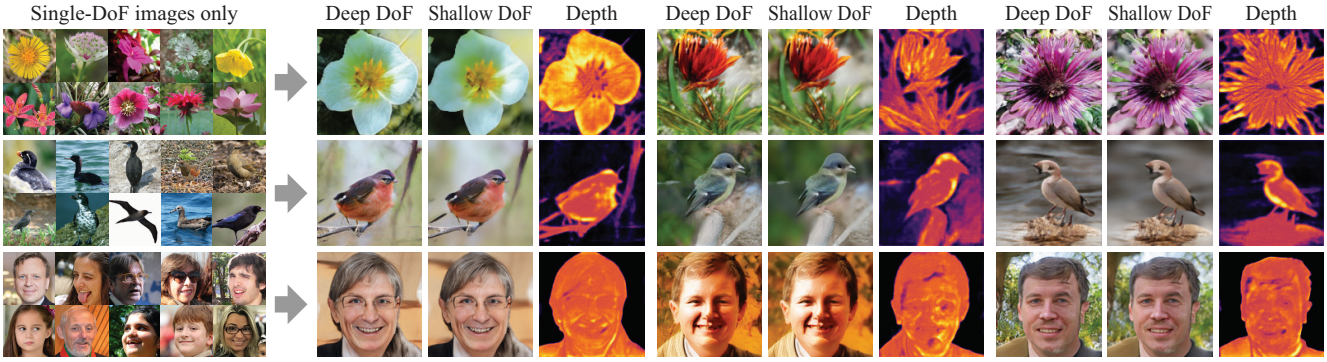
Unsupervised Learning of Depth and Depth-of-Field Effect from Natural Images with Aperture Rendering Generative Adversarial Networks

Takuhiro Kaneko

NTT Communication Science Laboratories, NTT Corporation



(a) Training data (b) Generated data

Figure 1. **Unsupervised learning of depth and depth-of-field (DoF) effect from unlabeled natural images.** (a) In training, we adopt *only* a collection of single-DoF images *without* any additional supervision (e.g., ground-truth depth, pairs of deep and shallow DoF images, and pretrained model). (b) Once trained, our model can synthesize tuples of deep and shallow DoF images and depths from random noise. The generated data are beneficial in learning a shallow DoF renderer, which also requires *no* external supervision. The project page is available at http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/ar-gan/.

Abstract

Understanding the 3D world from 2D projected natural images is a fundamental challenge in computer vision and graphics. Recently, an unsupervised learning approach has garnered considerable attention owing to its advantages in data collection. However, to mitigate training limitations, typical methods need to impose assumptions for viewpoint distribution (e.g., a dataset containing various viewpoint images) or object shape (e.g., symmetric objects). These assumptions often restrict applications; for instance, the application to non-rigid objects or images captured from similar viewpoints (e.g., flower or bird images) remains a challenge. To complement these approaches, we propose aperture rendering generative adversarial networks (AR-GANs), which equip aperture rendering on top of GANs, and adopt focus cues to learn the depth and depth-of-field (DoF) effect of unlabeled natural images. To address the ambiguities triggered by unsupervised setting (i.e., ambiguities between smooth texture and out-of-focus blurs, and between foreground and background blurs), we develop DoF mixture learning, which enables the generator to learn real image distribution while generating diverse DoF images. In addition, we devise a center focus prior to guiding the learning direction. In the experiments, we demonstrate the effectiveness of AR-GANs in various datasets, such as flower, bird, and face images, demonstrate their portability by incorporating them into other 3D representation learning GANs, and validate their applicability in shallow DoF rendering.

1. Introduction

Natural images are 2D projections of a 3D world. Addressing the inverse problem, i.e., understanding the 3D world from natural images, is a fundamental challenge in computer vision and graphics. Owing to its diverse applications in various fields, such as in robotics, content creation, and photo editing, this challenge has been actively studied.

A direct solution to challenge is learning a 3D predictor in a supervised manner using 2D and 3D data pairs or multiview image sets. However, obtaining such data is often impractical or time-consuming. To eliminate this process, several studies have attempted to learn 3D representations from single-view images (i.e., with only a single view per training instance). However, owing to the ill-posed nature, several studies required auxiliary information, such as 2D keypoints [49, 23] or 2D silhouettes [17, 6, 32, 13], to align object positions or extract a target object from the background. Other studies required predefined category-specific shape models (e.g., 3DMM [3] and SMPL [36]) [22, 56, 11, 43, 44] to obtain clues for reconstruction. Although they have exhibited promising results, collecting auxiliary information still requires a laborious annotation process, and a shape model requires additional preparation costs and restricts applicable objects.

To eliminate these disadvantages, fully unsupervised learning methods that enable 3D representation learning from single-view images *without* additional supervision and shape models have been devised. Although this is a se-

vere setting, previous studies have addressed this challenge by imposing assumptions for viewpoint distribution (e.g., a dataset including various viewpoint images) [38, 47, 40] or object shape (e.g., symmetric objects) [59]. The first assumption is required to learn 3D representations by sampling diverse viewpoint images. The second assumption is required to perform stereo reconstruction using a pair of mirrored images. Although these assumptions are practical for objects of a specific class (e.g., human faces), several objects do not satisfy these assumptions. For example, these methods are difficult to apply to non-rigid objects or images captured from similar viewpoints (e.g., flower or bird images).

To broaden the application without contradicting previous achievements, in this study, we consider complementary cues inherent in photos that have not been actively used in previous deep generative models (including those above). In particular, we focus on *focus cues*, in other words, we consider the learning depth[1] and the depth-of-field (DoF) effect in the defocus process. Specifically, instead of imposing an assumption on the *viewpoint distribution*, we do so on the *DoF distribution* (i.e., a dataset including various DoF images), and as shown in Figure 1, we attempt to learn 3D representations (particularly *depth* and *DoF effect*) from a collection of single-DoF images (i.e., images with solely a single DoF setting per training instance).

To achieve this, we propose a novel family of generative adversarial networks (GANs) [14], referred to as *aperture rendering GAN (AR-GAN)*, which equips aperture rendering (e.g., light field aperture rendering [46]) on top of GANs. Specifically, AR-GAN initially generates a pair of a deep DoF image and depth from a random noise, and then renders a shallow DoF image from the generated deep DoF image and depth via aperture rendering. With this mechanism, we can synthesize various DoF images using a virtual camera with an optical constraint on the light field.

When AR-GAN is learned in an unsupervised manner using single-DoF images, two non-trivial challenges are ambiguity between the smooth texture and out-of-focus blurs and ambiguity between the foreground and background blurs, as we cannot obtain explicit supervision of these relationships. For the first problem, we introduce *DoF mixture learning*, which enables the generator to learn the real image distribution while generating various DoF images. This learning ensures that the generated images (deep and shallow DoF images) are in a real distribution, and facilitates the learning of the depth, which is a source of connecting deep and shallow DoF images. For the second problem, based on the observed tendency to focus on the center object when a focused image is considered, we impose a *center focus prior*, which facilitates the focusing of the center while guiding the surroundings to be behind the focal plane. In practice, we adopt prior solely at the beginning of training to guide the learning direction.

[1]In this study, we use depth and disparity interchangeably to indicate disparity across a camera aperture.

To evaluate the effectiveness of AR-GAN, we first conducted experiments with comparative and ablation studies on diverse datasets, including flower (Oxford Flowers 102 [39]), bird (CUB-200-2011 [53]), and face (FFHQ [26]) datasets. A significant property of AR-GAN is its portability, which we validated by incorporating AR-GAN into other 3D representation learning GANs (i.e., HoloGAN [38] and RGBD-GAN [40]). Another significant property of AR-GAN is its ability to synthesize a tuple of deep and shallow DoF images and depth from a random noise, after training. We utilize this property to train a shallow DoF renderer and empirically demonstrate its utility.

Overall, our contributions are summarized as follows:

- We provide *unsupervised learning of depth and DoF effect from unlabeled natural images*. This is noteworthy because it does not impose assumptions on the viewpoint distribution or object shape, which are required in conventional unsupervised 3D representation learning.
- To achieve this, we propose a novel GAN family (*AR-GAN*), which generates a deep DoF image and depth from a random noise and renders a shallow DoF image from them via aperture rendering.
- To address ambiguities caused by a fully unsupervised setting, we devise *DoF mixture learning* to enable the generator to learn real image distribution using generated diverse DoF images, and develop a *center focus prior* to determine the learning direction.
- We validate the effectiveness, portability, and applicability of AR-GANs via extensive experiments. The project page is available at `http://www.kecl.ntt.co.jp/people/kaneko.takuhiro/projects/ar-gan/`.

2. Related work

Generative adversarial networks. GANs [14] have achieved remarkable success in 2D image modeling via a series of advancements (e.g., [5, 26, 27]). A substantial property of GANs is their ability to mimic data distribution in a random sampling process without explicitly defining the data distribution. This allows GANs to learn various distribution types. For example, recent studies [57, 58, 38, 17, 47, 40, 33] have made it possible to learn a 3D-aware image distribution via 3D GAN architectures or 3D representations. Among them, HoloGAN [38] and RGBD-GAN [40] share a similar motivation with us in terms of learning 3D representations from natural images in a fully unsupervised manner; however, the major difference is that they adopt *viewpoint cues*, whereas we employ *focus cues*. We empirically demonstrate this difference in Section 5.2. Owing to this difference, the previous and present models are not exchangeable but complementary. We verify their compatibility in Section 5.4 by combining AR-GAN with HoloGAN and RGBD-GAN.

Another related topic is the application of GANs for un-

supervised learning of the foreground and background [51, 62]. Although previous and present studies are relevant in terms of learning image compositions, they decompose the image *discretely*, whereas we learn the *continuous depth*. Furthermore, we can learn the *DoF effect*, which has not been achieved in previous studies.

Other relevant GANs are GANs with measurements [4, 41, 31, 24, 25], which apply measurements (e.g., mask and noise) before matching a generated image with a real image. Our aperture rendering functions similarly to those measurements. However, in the previous work, applicable measurements were limited to those in a *2D image plane*, and effectiveness was solely demonstrated on *synthetically corrupted* images. In contrast, AR-GAN can learn a DoF effect, which yields a *3D space*. In the experiments, we verify that this effect can be learned from images taken in *real* scenarios (Section 5).

Unsupervised 3D representation learning. As mentioned in Section 1, the learning of 3D representations from single-view images has garnered attention owing to its data collection advantage. To address this challenge, several studies have employed auxiliary information as clues for reconstruction, such as 2D keypoints [49, 23], 2D silhouettes [17, 6, 32, 13], or shape models [22, 56, 11, 43, 44]. In contrast, we attempt to address this challenge with *no* additional supervision and *no* predefined model to reduce costs from laborious annotation and model preparation.

Recently, some studies [38, 47, 40, 59] have addressed this; however, their assumptions and objectives differ from ours. They impose assumptions for the *viewpoint distribution* or *object shape*, whereas we impose an assumption for the *DoF distribution*. Owing to this assumption difference, they can learn 3D meshes [47], depth [40, 59], albedo [59], texture [47], light [59], and viewpoints [38, 47, 40, 59], whereas AR-GAN can learn the depth and DoF effect. Therefore, AR-GAN can be considered a model that can complement (not replace) previous models, and we validate this statement by incorporating AR-GAN into HoloGAN [38] and RGBD-GAN [40] (Section 5.4).

Monocular depth estimation. Monocular depth estimation involves predicting the depth when a single image is given. A successful approach involves learning a depth predictor using paired or consecutive data, such as image and depth pairs [8, 34, 30, 29, 61, 9], stereo pairs [10, 12, 60], and videos [65, 63, 54]. Although this approach is a promising solution, collecting such data is often impractical or time-consuming.

In another direction, some studies [46, 15] have proposed the adoption of focused and all-in-focus image pairs, including learning the depth in the process of reconstructing the focused image from an all-in-focus image. Although this study is inspired by their success, the main difference is that they require paired supervision between focused and all-in-focus images, whereas ours does not need it. However, owing to this difference, our task is very challenging; therefore, in this study, we did not attempt to achieve high-quality depth estimation comparable to supervised methods. Instead, in the experiments, we compared AR-GAN with a previous fully unsupervised depth estimation model (i.e., RGBD-GAN [40]) and demonstrated the utility of AR-GAN in this challenging setting (Section 5.2).

DoF rendering. The DoF or Bokeh effect is a popular photography technique, and its synthesis has garnered considerable interest in computer vision and graphics. To achieve this without prior knowledge of geometry and lightning, previous studies adopted stereo pairs [1], a stack of images taken in different camera settings [20, 16], and a segmentation mask [45, 52] to determine the degree of blur. Although they have exhibited remarkable results, they are limited owing to their general dependence on a manually defined DoF renderer. To address this limitation, end-to-end supervised learning methods [46, 55, 19, 42], which learn a DoF renderer using pairs of shallow and deep DoF images, were devised. Recently, an unpaired learning method [66] was also proposed. This method eliminates the requirement for paired supervision; however, set-level supervision (i.e., supervision of whether each image is a deep or shallow DoF image) remains necessary. In contrast, we focus on learning a DoF renderer in a *fully* unsupervised manner. We demonstrate the effectiveness of our approach in Section 5.5.

3. Preliminaries

3.1. GANs

We briefly introduce two previous works on which our model is based. The first is GAN [14], which learns data distribution using the following objective:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = &\mathbb{E}_{I^r\sim p^r(I)}[\log C(I^r)] \\ &+ \mathbb{E}_{z\sim p(z)}[\log(1 - C(G(z))],\end{aligned} \tag{1}$$

where, given a random noise z, a generator G generates an image $I^g = G(z)$ that can deceive a discriminator C by minimizing this objective, whereas C distinguishes I^g from a real image I^r by maximizing this objective. Here, superscripts r and g denote the real and generated data, respectively. Using this min-max game, the generative distribution $p^g(I)$ approaches the real distribution $p^r(I)$.

3.2. Light field aperture rendering

Light field aperture rendering [46] is a type of differentiable aperture rendering.[2] Its objective is to learn an aperture renderer R that synthesizes a shallow DoF image $I_s(\mathbf{x}) = R(I_d(\mathbf{x}), D(\mathbf{x}))$, given a deep DoF image $I_d(\mathbf{x})$ and depth $D(\mathbf{x})$.[3] Here, $\mathbf{x}$ represents the spatial coordinates

[2] Another representative aperture rendering is compositional aperture rendering [46], which discretely models disparities using a stack of blur kernels. In the initial experiments, we determined that light field aperture rendering, which models the light field within a camera, is more compatible with our unsupervised learning. This is possibly because the learning clues are few in our unsupervised learning; therefore, a camera constraint via light field aperture rendering works sufficiently.

[3] In the original study [46], $D(\mathbf{x})$ is estimated from $I(\mathbf{x})$. However, this estimation is not adopted in AR-GAN; hence, we omitted it here.

of the light field on the image plane. When $I_d(\mathbf{x})$ is directly warped into the viewpoint in the light field based on $D(\mathbf{x})$, holes can appear in the resulting light field. Instead, a trainable neural network T is adopted to expand $D(\mathbf{x})$ into a depth map $M(\mathbf{x}, \mathbf{u})$ for each view in the light field:

$$M(\mathbf{x}, \mathbf{u}) = T(D(\mathbf{x})), \tag{2}$$

where $\mathbf{u}$ denotes the angular coordinates of the light field on the aperture plane. Subsequently, $I_d(\mathbf{x})$ is warped into each view in the light field using the depth map $M(\mathbf{x}, \mathbf{u})$:[4]

$$L(\mathbf{x}, \mathbf{u}) = I_d(\mathbf{x} + \mathbf{u}M(\mathbf{x}, \mathbf{u})), \tag{3}$$

where $L(\mathbf{x}, \mathbf{u})$ is the simulated camera light field. Finally, it is integrated to render a shallow DoF image $I_s(\mathbf{x})$:

$$I_s(\mathbf{x}) = \sum_{\mathbf{u}} A(\mathbf{u}) L(\mathbf{x} + \mathbf{u}, \mathbf{u}), \tag{4}$$

where $A(\mathbf{u})$ is an indicator that represents the disk-shaped camera aperture. Hereafter, for simplicity, we omit $\mathbf{x}$ and $\mathbf{u}$ when they are not required.

4. Aperture rendering GANs: AR-GANs

4.1. Problem statement

We begin by defining the problem statement. We consider a fully unsupervised setting in which we cannot obtain any supervision or pretrained model except for an image collection. As discussed in Section 2, typical end-to-end focus-based monocular depth estimation methods (e.g., [46, 15]), and DoF rendering methods (e.g., [46, 19, 42, 66]) achieve their objectives using a conditional model (i.e., a deep DoF image is used as the input, and a depth or shallow DoF image is estimated based on it). However, in our fully unsupervised setting, we cannot employ this formulation as we cannot determine whether each image is either a deep or shallow DoF image.

Alternatively, we aim to learn an unconditional generator $G(z)$ that can generate a tuple of a deep DoF image, depth, and shallow DoF image, i.e., (I_d^g, D^g, I_s^g), from a random noise. When the training images are extremely biased in terms of the DoF (e.g., all images are all-in-focus), it is difficult to determine focus cues from the images; hence, we impose the following assumption on an image distribution:

Assumption 1 *The DoF setting is different for each image, and the dataset includes various DoF images.*

Note that we do not have to collect a set of various DoF images for each training instance. We observed that this assumption is satisfied by typical natural image datasets (e.g., flower [39], bird [53], and face [26] datasets shown in Figure 1). Under this assumption, we aim to learn the above-mentioned generator in a *wisdom of crowds* approach.

[4]The depth of the focal plane can be learned explicitly by adding the parameterized offset $\hat{m}$ to M in Equation 3. However, we do not do so under the assumption that it is determined per image I_d and internally represented and optimized in D, which is used in Equation 2. In this case, the focal plane exists at $D = 0$, while out-of-focus occurs in $|D| > 0$.

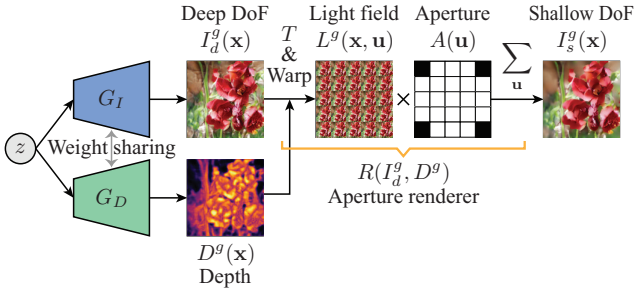


Figure 2. **Overall pipeline of AR-GAN generator.** The AR-GAN generator first generates a deep DoF image I_d^g and depth D^g from a random noise z, and then renders a shallow DoF image I_s^g from I_d^g and D^g using the aperture renderer R.

4.2. Overall pipeline

The overall pipeline of the AR-GAN generator is illustrated in Figure 2. When given a random noise z, we first generate the deep DoF image I_d^g and depth D^g as follows:

$$I_d^g = G_I(z), D^g = G_D(z). \tag{5}$$

In practice, we share the weights between G_I and G_D except for the last layer because the image and depth exhibit high correlation. A previous study [35] demonstrated that this kind of weight sharing is beneficial in learning a joint distribution between relevant domains. Subsequently, we render a shallow DoF image I_s^g from the generated I_d^g and D^g using the aperture renderer R described in Section 3.2.

Typical GANs apply a discriminator C to the final output of the generator (i.e., I_s^g in our case). However, in AR-GAN, both generators (i.e., G_I and G_D) and R are trainable. Hence, without constraints, they could compete roles. Consequently, they can drift into an extreme solution (e.g., R learns strong out-of-focus, and G_I learns an over-deblurred image). To address this, we develop *DoF mixture learning*, which is detailed in the next section.

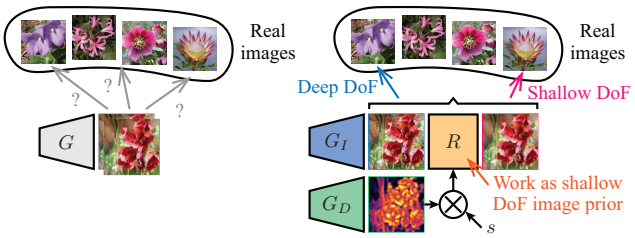
4.3. DoF mixture learning

A possible solution to this problem is regularizing G_I using an explicit distance metric (e.g., L1, L2, or perceptual loss [21, 7]) such that I_d^g is approximate to I_s^g. However, this solution disrupts the depth learning (Section 5.3.1).

Alternatively, we introduce *DoF mixture learning*. Figure 3 illustrates the comparison between standard and DoF mixture learning. In standard GAN training, the generator attempts to cover the real image distribution using images without constraints. In contrast, in the DoF mixture learning, the generator attempts to represent the real image distribution using diverse DoF images whose extent is adjusted by a scale factor s. More precisely, in our AR-GAN, the GAN objective (Equation 1) is rewritten as follows:

$$\begin{aligned}\mathcal{L}_{\text{AR-GAN}} &= \mathbb{E}_{I^r \sim p^r(I)}[\log C(I^r)] \\ &+ \mathbb{E}_{z \sim p(z), s \sim p(s)}[\log(1 - C(R(G_I(z), sG_D(z))))], \end{aligned} \tag{6}$$

where $s \in [0, 1]$; when $s = 0$, a deep DoF image (almost equal to I_d^g) is rendered, whereas when $s = 1$, a shallow



(a) Standard learning (baseline) (b) DoF mixture learning (proposed)

Figure 3. **Comparison of standard and DoF mixture learning.**

DoF image (i.e., I_s^g) is rendered. Intuitively, the aperture renderer R, which has an optical constraint on the light field, functions as a shallow DoF image prior. Under Assumption 1 (a real image distribution $p^r(I)$ includes both deep I_d^r and shallow I_s^r DoF images), this prior encourages the generated deep I_d^g and shallow I_s^g DoF images to be approximate to I_d^r and I_s^r, respectively. This also facilitates the learning of D^g, which is a source of the I_d^g and I_s^g connection.

In practice, we determined that sampling s from a binomial distribution, i.e., $p(s) = B(1, p_s)$, works optimally, where p_s indicates a probability of $s = 1$. In Section 5.3, we examine the effect of the p_s value. It was manually determined for simplicity; however, optimizing it in a data-driven approach is a potential direction for future work.

4.4. Center focus prior

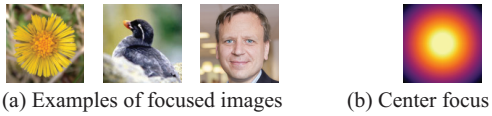
Another challenge unique to unsupervised depth and DoF effect learning is to the difficulty in distinguishing foreground and background blurs without any constraint or prior knowledge. Although not all images satisfy this, focused images tend to be captured when the main targets are positioned at the center, as shown in Figure 4(a). Based on this observation, we impose a center focus prior defined by

$$D_p = \begin{cases} 0 & (r <= r_{\text{th}}) \\ -g \cdot (r - r_{\text{th}}) & (r > r_{\text{th}}), \end{cases} \tag{7}$$

where r indicates the distance from the center of the image, and r_{th} and g denote the hyper-parameters that define the focused area and depth gain, respectively. We visualize this prior in Figure 4(b). As shown in this figure, the prior facilitates the center area focus while promoting the surrounding area to be behind the focal plane. We apply this prior to the generated depth D^g as follows:

$$\mathcal{L}_p = \lambda_p \| D^g - D_p \|_2^2, \tag{8}$$

where λ_p represents a weighting parameter. In practice, we apply this only at the beginning of training to mitigate the negative effect triggered by the gap between D^r and D_p.



(a) Examples of focused images (b) Center focus prior

Figure 4. **Examples of focused images and center focus prior.** In (b), light color indicates the foreground.

5. Experiments

5.1. Experimental settings

We conducted four experiments to verify the effectiveness of AR-GANs from multiple perspectives: a comparative study on unsupervised 3D representation learning (Section 5.2), ablation studies on DoF mixture learning and center focus prior (Section 5.3), portability analysis (Section 5.4), and application to shallow DoF rendering (Section 5.5). Here, we explain the common settings and present the details of each in the following sections.

Datasets. We evaluated AR-GANs on three natural image datasets that cover various objects: Oxford Flowers 102 [39] (8189 flower images with 102 categories), CUB-200-2011 [53] (11788 bird images with 200 categories), and FFHQ [26] (70000 face images). To efficiently examine various cases, we resized the images to 64×64. We also experimented on 128×128 images in some cases to confirm the dependency on image resolution (e.g., Figure 1).

Metrics. To evaluate the visual fidelity of the generated images, we adopted the *kernel inception distance (KID)* [2],[5] which computes the maximum mean discrepancy between real and generated images within the Inception model [48]. When calculating scores, we generated 20000 images from each model. Measuring depth and DoF accuracy directly is non-trivial because we aim to learn an unconditional model from unpaired and unlabeled natural images, and cannot obtain the ground truth. Alternatively, we evaluated the depth accuracy by (1) learning the depth estimator using pairs of images and depths generated by GANs, (2) predicting the depths of real images using the learned depth estimator, and (3) comparing the obtained results with the depths predicted by a state-of-the-art monocular depth estimator [60], which is trained using stereo pairs in an external dataset.[6] We used *scale-invariant depth error (SIDE)* [8] to measure the difference. In both metrics, the performance increased as the score decreased. In all the experiments, we report the mean score with the standard deviation over three training runs.

Implementation. We implemented the model based on HoloGAN [38]. The generator has a StyleGAN-like architecture [26]. In AR-GANs, 3D convolution used in HoloGAN is not required; hence, we replaced it with 2D convolution. The discriminator has instance [50] and spectral normalizations [37]. The networks were trained using the Adam optimizer [28] with a non-saturating GAN loss [14].

5.2. Comparative study

First, we conducted a comparative study to clarify the difference between AR-GAN and previous fully unsupervised 3D representation learning. For comparison, we used

[5]We used the KID because it has an unbiased estimator and complements the flaws of other representative metrics (i.e., Fréchet inception distance (FID) [18] and inception score (IS) [18]).

[6]We used the pretrained model provided by the authors: https://github.com/KexianHust/Structure-Guided-Ranking-Loss.

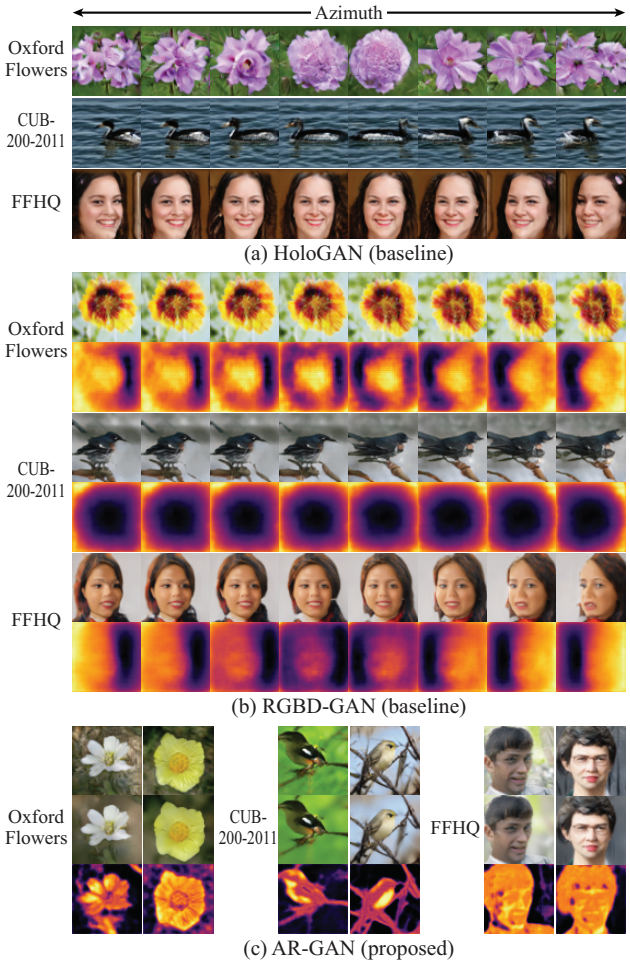


Figure 5. **Qualitative comparison of HoloGAN, RGBD-GAN, and AR-GAN.** HoloGAN, RGBD-GAN, and AR-GAN generate images, image and depth pairs, and tuples of deep and shallow DoF images and depths, respectively.

HoloGAN [38] and *RGBD-GAN* [40], which are representative models in this category, as well as the standard *GAN* [14]. As discussed in Section 2, HoloGAN/RGBD-GAN learns 3D representations using *viewpoint cues*, whereas AR-GAN achieves this with *focus cues*. Hence, the applicable datasets are different, which we verified by applying the models to the three datasets.

Results. Examples of the generated images are presented in Figure 5. Here, the obtainable 3D representations and applicable datasets differ among the GANs. Although HoloGAN and RGBD-GAN succeed in learning viewpoint-aware representations in FFHQ, they fail to do so in Oxford Flowers and CUB-200-2011, where viewpoint distributions are biased and viewpoint cues do not exist sufficiently. In contrast, AR-GAN succeeds in learning the depth and DoF effect in all datasets because it can employ focus cues, which are present across all datasets.

The KID comparison is summarized in Table 1. We found that AR-GAN achieved comparable performance and did not incur a negative effect across all datasets.

KID $\times 10^3 \downarrow$	Oxford Flowers	CUB-200-2011	FFHQ
GAN	11.71 ± 0.68	15.04 ± 0.14	6.97 ± 0.30
HoloGAN	11.30 ± 0.37	14.68 ± 0.51	6.89 ± 0.38
RGBD-GAN	12.04 ± 0.35	14.92 ± 0.49	6.73 ± 0.26
AR-GAN	**11.23** ± 0.36	**14.30** ± 0.56	**5.75** ± 0.19

Table 1. **Comparison of KID $\times 10^3 \downarrow$ among different GANs.**

SIDE $\times 10^2 \downarrow$	Oxford Flowers	CUB-200-2011	FFHQ
RGBD-GAN	7.01 ± 0.81	7.06 ± 0.02	5.81 ± 0.40
AR-GAN	**4.46** ± 0.03	**3.58** ± 0.04	**4.21** ± 0.15

Table 2. **Comparison of SIDE $\times 10^2 \downarrow$ between RGBD-GAN and AR-GAN.** GAN and HoloGAN are not listed here because they cannot generate depth along with an image.

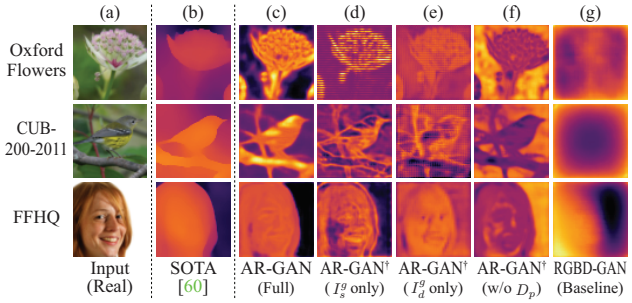


Figure 6. **Examples of predicted depths.** † indicate the ablated models. (c–g) Results obtained in a fully unsupervised setting.

The SIDE comparison is presented in Table 2. We found that AR-GAN outperforms RGBD-GAN in all datasets. Examples of the predicted depths are presented in Figure 6. Although the predicted depths exhibit a lower resolution that those predicted by the supervised methods (e.g, [60]),[7] we found that AR-GAN (c) improves the details (e.g., flower details and tree branches) that disappear in [60] (b) and RGBD-GAN (g), thus benefiting from focus cues.

5.3. Ablation study

5.3.1 Ablation study on DoF mixture learning

Metrics. We first evaluated the importance of the DoF mixture learning. As mentioned in Section 5.1, measuring depth and DoF accuracy directly is non-trivial; therefore, we further adopted two metrics alongside KID and SIDE: *(1) Learned perceptual image patch similarity (LPIPS)* [64] computes the distance between two images in the CNN feature space and is demonstrated to exhibit a high correlation with human perceptual similarity [64]. We adopted LPIPS to measure the perceptual similarity between pairs of I_d^g and I_s^g. LPIPS is expected to be moderately small because the content is preserved before and after the application of aperture rendering. *(2) Depth standard deviation (DSD)* is the standard deviation of the generated depths. Our objective is to learn a meaningful depth that can yield a plausible DoF effect. When depth learning is successful, DSD is expected to be sufficiently large.

[7] Note that 64×64 is the standard resolution for fully unsupervised learning methods (e.g., HoloGAN and RGBD-GAN), and applications to images with complex objects/backgrounds are challenging for them.

Oxford Flowers	KID×10^3↓	SIDE×10^2↓	LPIPS↓	DSD↑
I_s^g only $p_s = 1$	12.36 ±0.59	5.48 ±0.20	0.229 ±0.027	0.157 ±0.063
Mixture $p_s = 0.75$	10.97 ±0.26	4.81 ±0.06	0.023 ±0.001	0.657 ±0.006
Mixture $p_s = 0.5$	10.69 ±0.48	4.65 ±0.05	0.022 ±0.000	0.771 ±0.022
Mixture $p_s = 0.25$	11.23 ±0.36	4.46 ±0.03	0.028 ±0.001	1.007 ±0.025
I_d^g only $p_s = 0$	11.58 ±0.37	4.56 ±0.20	0.113 ±0.013	0.446 ±0.065
L1	11.66 ±0.72	5.26 ±0.65	0.033 ±0.001	0.387 ±0.116
Double discriminators	9.74 ±0.31	6.79 ±2.21	0.000 ±0.001	0.032 ±0.046
CUB-200-2011	KID×10^3↓	SIDE×10^2↓	LPIPS↓	DSD↑
I_s^g only $p_s = 1$	13.62 ±0.53	4.63 ±0.50	0.125 ±0.037	0.354 ±0.021
Mixture $p_s = 0.75$	12.68 ±0.61	3.75 ±0.08	0.037 ±0.003	0.748 ±0.072
Mixture $p_s = 0.5$	13.14 ±0.03	3.55 ±0.02	0.043 ±0.003	0.959 ±0.075
Mixture $p_s = 0.25$	14.30 ±0.56	3.58 ±0.04	0.059 ±0.002	1.175 ±0.017
I_d^g only $p_s = 0$	14.58 ±0.56	5.94 ±0.70	0.115 ±0.019	0.193 ±0.012
L1	12.54 ±0.32	5.75 ±1.26	0.042 ±0.001	0.725 ±0.195
Double discriminators	12.50 ±0.12	4.33 ±0.34	0.000 ±0.000	0.001 ±0.000
FFHQ	KID×10^3↓	SIDE×10^2↓	LPIPS↓	DSD↑
I_s^g only $p_s = 1$	5.75 ±0.44	6.00 ±0.35	0.097 ±0.011	0.296 ±0.018
Mixture $p_s = 0.75$	5.67 ±0.23	4.38 ±0.10	0.009 ±0.001	0.757 ±0.177
Mixture $p_s = 0.5$	5.75 ±0.19	4.21 ±0.15	0.009 ±0.001	0.769 ±0.119
Mixture $p_s = 0.25$	6.17 ±0.08	4.68 ±0.33	0.010 ±0.001	0.583 ±0.071
I_d^g only $p_s = 0$	6.85 ±0.13	4.77 ±0.13	0.028 ±0.006	0.202 ±0.003
L1	5.82 ±0.21	4.82 ±0.09	0.015 ±0.004	0.466 ±0.045
Double discriminators	6.20 ±0.08	5.20 ±0.47	0.000 ±0.000	0.000 ±0.000

Table 3. **Comparison of KID×10^3↓, SIDE×10^2↓, LPIPS↓, and DSD↑ among AR-GANs with different learning methods.**

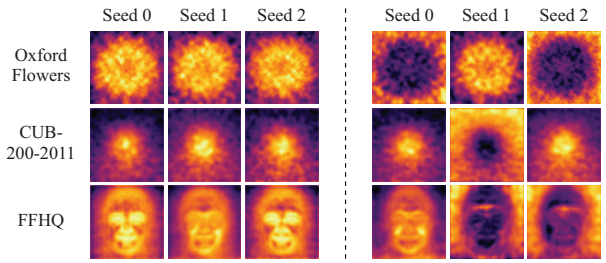
Comparison models. We conducted the analysis from two perspectives. (1) We evaluated the effect of the p_s value, which indicates the rate of using shallow DoF images in the DoF mixture learning (Equation 6). (2) We tested two possible alternatives: *L1*, which uses L1 loss to guide I_d^g closer to I_s^g, and *double discriminators*, which adopts two discriminators, for I_d^g and I_s^g, respectively. This facilitates both $p^g(I_d)$ and $p^g(I_s)$ to coincide with the overall real distribution $p^r(I)$.

Results. A comparison of the scores is summarized in Table 3. Our main findings are two-folds:

(1) Effect of value of p_s. We found that some fluctuations exist in the KID; however, in all cases, the scores are comparable to those of the other GANs presented in Table 1. This indicates that AR-GANs can generate plausible images regardless of p_s. In contrast, SIDE, LPIPS, and DSD are affected by p_s. SIDE tends to improve when the DoF mixture learning is adopted.[8] This is because in the DoF mixture learning, we can encourage I_d^g and I_s^g to be approximate to I_d^r and I_s^r, respectively, as well as facilitate D^g learning, which is the source that connects them. Examples of predicted depths (Figure 6 (c–e)) also validate the effectiveness of DoF mixture learning. Regarding LPIPS and DSD, as LPIPS increases, and DSD successively decreases when $p_s = 1$ or $p_s = 0$. This indicates that the DoF mixture learning is required to manage LPIPS and DSD. Among AR-GANs with DoF mixture learning (i.e., $p_s \in \{0.25, 0.5, 0.75\}$), there is a trade-off and dataset dependency relative to LPIPS and DSD. Consider the score balance, we set $p_s = 0.25$ for Oxford Flowers and CUB-200-2011 and $p_s = 0.5$ for FFHQ in other experiments.

(2) Comparison with alternatives. Although L1 achieves reasonable LPIPS, it deteriorates SIDE and DSD more than those in the DoF mixture learning. This result indicates that this method disrupts depth learning. Double discriminators facilitate both $p^g(I_s)$ and $p^g(I_d)$ to coincide with $p^r(I)$. Consequently, LPIPS and DSD approach zero, and SIDE depreciates. This result verifies the importance of mixing I_d^g and I_s^g when learning $p^r(I)$.

[8] The sole "I_d^g only" case in Oxford Flowers is an exception. In this case, D^g is not regularized by aperture rendering; however, weight sharing between G_I and G_D (Section 4.2) aids the depth learning. This strategy exhibits dataset dependency and fails in the other datasets.



(a) With D_p (b) Without D_p

Figure 7. **Comparison of AD with and without D_p.**

Oxford Flowers	KID×10^3↓	SIDE×10^2↓	LPIPS↓	DSD↑
W/ D_p	11.23 ±0.36	4.46 ±0.03	0.028 ±0.001	1.007 ±0.025
W/o D_p	10.69 ±0.24	6.78 ±1.58	0.026 ±0.002	0.915 ±0.137
CUB-200-2011	KID×10^3↓	SIDE×10^2↓	LPIPS↓	DSD↑
W/ D_p	14.30 ±0.56	3.58 ±0.04	0.059 ±0.002	1.175 ±0.017
W/o D_p	13.96 ±0.63	4.86 ±1.84	0.062 ±0.004	1.183 ±0.059
FFHQ	KID×10^3↓	SIDE×10^2↓	LPIPS↓	DSD↑
W/ D_p	5.75 ±0.19	4.21 ±0.15	0.009 ±0.001	0.769 ±0.119
W/o D_p	5.72 ±0.10	6.70 ±1.88	0.009 ±0.001	0.851 ±0.057

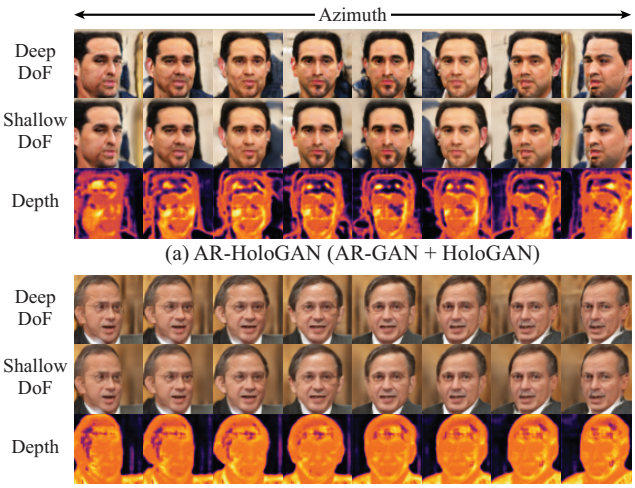
Table 4. **Comparison of KID×10^3↓, SIDE×10^2↓, LPIPS↓, and DSD↑ among AR-GANs with and without D_p.**

5.3.2 Ablation study on center focus prior

Metrics. We evaluated the necessity of the center focus prior D_p. To assess the overall tendency of each pixel to represent the foreground or background blur, we calculated the *average depth (AD)*, i.e., the pixel-wise average of the generated depths. To validate the learning consistency, we compared the results over three training runs.

Results. The results are presented in Figure 7. We found that we can obtain constant results across training runs when D_p is adopted. In all results, the center is focused, while the surroundings are behind the focal plane. In contrast, when we eliminate D_p, the foreground and background are turned over, depending on the initialization. These results indicate that D_p is beneficial in determining the learning direction.

The comparison of the scores is summarized in Table 4. We found that KID, LPIPS, and DSD are comparable across all datasets. We deduce that D_p is solely adopted at the beginning of training; therefore, it does not disrupt the entire training. In contrast, SIDE depreciates when D_p is not implemented. This occurs because the foreground and background are reversed, as shown in Figure 6(f).



(a) AR-HoloGAN (AR-GAN + HoloGAN)

(b) AR-RGBD-GAN (AR-GAN + RGBD-GAN)

Figure 8. **Examples of images generated using AR-HoloGAN and AR-RGBD-GAN.** The viewpoint change in the horizontal direction is obtained by the HoloGAN/RGBD-GAN function, while the DoF change and depth in the vertical direction are obtained by the AR-GAN function.

5.4. Portability analysis

As presented in Section 5.2, the obtainable representations differ between HoloGAN/RGBD-GAN and AR-GAN. An interesting approach would be learning these representations jointly by combining them. A significant property of AR-GAN is portability, i.e., it is easy to incorporate into other GANs. Specifically, we can achieve this simply by adding aperture rendering on top of HoloGAN/RGBD-GAN and training it with the DoF mixture training and a center focus prior. One requirement is that a dataset should satisfy the assumptions of both models; i.e., a dataset should include diverse viewpoint images along with various DoF images. Among the datasets described above, only FFHQ satisfies this requirement. Hence, we solely evaluated *AR-HoloGAN* (AR-GAN + HoloGAN) and *AR-RGBD-GAN* (AR-GAN + RGBD-GAN) on this dataset.

Results. Examples of generated images are shown in Figure 8. As shown, we can jointly control both viewpoints and the DoF effect with the HoloGAN/RGBD-GAN and AR-GAN functions. As a reference, we also calculated the KID scores. The scores for AR-HoloGAN and AR-RGBD-GAN were 5.70 ± 0.32 and 5.43 ± 0.22, respectively. These are better than the scores of the original AR-GAN, HoloGAN, and RGBD-GAN (Table 1).

5.5. Application in shallow DoF rendering

Finally, we demonstrate the applicability of AR-GAN in shallow DoF rendering. After training, AR-GAN can synthesize tuples of (I_d^g, D^g, I_s^g) from random noise. By utilizing this, we learn a shallow DoF renderer $I_d \rightarrow I_s$ using pairs of (I_d^g, I_s^g). We call this approach *AR-GAN-R*. As another approach, we learn a depth estimator $I_d \rightarrow D$ using pairs of (I_d^g, D_g). By employing the learned depth estimator, we estimate D from I_d and then render I_s from (I_d, D) using R in AR-GAN. We call this approach *AR-GAN-DR*.

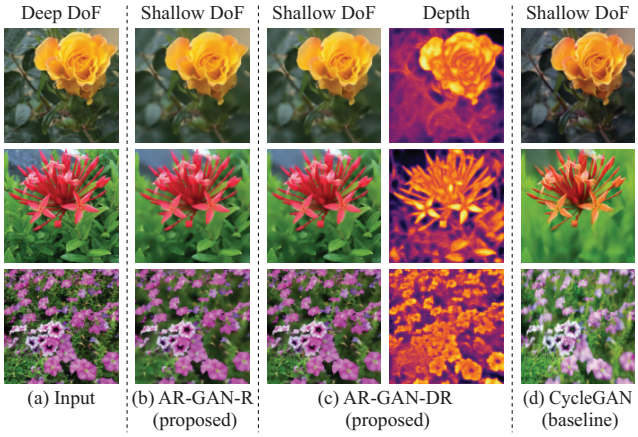


Figure 9. **Examples of shallow DoF rendering.**

Comparison model. To the best of our knowledge, no previous method can learn the DoF effect from natural images in the same setting as ours (i.e., *without* additional supervision and a predefined model). Therefore, as a baseline, we used CycleGAN [66], which can learn a shallow DoF renderer $I_d \rightarrow I_s$ using *set-level* supervision (i.e., supervision of whether each image is a deep or shallow DoF image).[9]

Dataset. We used Oxford Flowers and AR-GAN-generated images to train AR-GAN and AR-GAN-R/DR, respectively. To confirm generality, we conducted a test on a different dataset, including flower photos taken by smartphones, which were used in the CycleGAN study [66].

Results. Examples of the rendered images are presented in Figure 9. We found that CycleGAN often yields unnecessary changes (e.g., color change), whereas AR-GAN-R/DR does not. We infer that the aperture rendering mechanism in AR-GAN contributes to this phenomenon. In addition, AR-GAN-DR can estimate the depth simultaneously.

6. Conclusion

We proposed a novel family of GANs, AR-GANs, which can learn depth and DoF effect from unconstrained natural images. To achieve this, we incorporated aperture rendering into GANs and developed DoF mixture learning and a center focus prior to address the ambiguities triggered by the unsupervised setting. Via comparative and ablation studies, we elucidated the differences from previous GANs and the significance of the proposed techniques. We demonstrated that AR-GANs are compatible and complementary to previous GANs by combining AR-GANs with HoloGAN/RGBD-GAN. Finally, we demonstrated the applicability of AR-GANs in shallow DoF rendering. Despite their applications in photos, several deep generative models do not utilize focus cues. In the future, we expect that our findings will facilitate further studies on such models.

[9] We used the pretrained model provided by the authors: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

References

[1] Jonathan T Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. In *CVPR*, 2015. 3

[2] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 5

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, 1999. 1

[4] Ashish Bora, Eric Price, and Alexandros G Dimakis. AmbientGAN: Generative models from lossy measurements. In *ICLR*, 2018. 3

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[6] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3D objects with an interpolation-based differentiable renderer. In *NeurIPS*, 2019. 1, 3

[7] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NIPS*, 2016. 4

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 3, 5

[9] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 3

[10] Ravi Garg, Vijay Kumar B G, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 3

[11] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *CVPR*, 2019. 1, 3

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 3

[13] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, 2020. 1, 3

[14] 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, 3, 5, 6

[15] Shir Gur and Lior Wolf. Single image depth estimation trained via depth from defocus cues. In *CVPR*, 2019. 3, 4

[16] Samuel W Hasinoff and Kiriakos N Kutulakos. A layer-based restoration framework for variable-aperture photography. In *ICCV*, 2007. 3

[17] Philipp Henzler, Niloy Mitra, and Tobias Ritschel. Escaping Plato's cave using adversarial training: 3D shape from unstructured 2D image collections. In *ICCV*, 2019. 1, 2, 3

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *NIPS*, 2017. 5

[19] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *CVPR Workshops*, 2020. 3, 4

[20] David E Jacobs, Jongmin Baek, and Marc Levoy. Focal stack compositing for depth of field control. *Stanford Computer Graphics Laboratory Technical Report*, 1(1):2012, 2012. 3

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 3

[23] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 1, 3

[24] Takuhiro Kaneko and Tatsuya Harada. Noise robust generative adversarial networks. In *CVPR*, 2020. 3

[25] Takuhiro Kaneko and Tatsuya Harada. Blur, noise, and compression robust generative adversarial networks. In *CVPR*, 2021. 3

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 4, 5

[27] 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

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[29] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, 2017. 3

[30] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 3

[31] Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. MisGAN: Learning from incomplete data with generative adversarial networks. In *ICLR*, 2019. 3

[32] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3D reconstruction via semantic consistency. In *ECCV*, 2020. 1, 3

[33] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *CVPR*, 2020. 2

[34] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):2024–2039, 2015. 3

[35] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016. 4

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):1–16, 2015. 1

[37] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 5

[38] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *ICCV*, 2019. 2, 3, 5, 6

[39] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 2, 4, 5

[40] Atsuhiro Noguchi and Tatsuya Harada. RGBD-GAN: Unsupervised 3D representation learning from natural image datasets via RGBD image synthesis. In *ICLR*, 2020. 2, 3, 6

[41] Arthur Pajot, Emmanuel de Bezenac, and Patrick Gallinari. Unsupervised adversarial image reconstruction. In *ICLR*, 2018. 3

[42] Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, and Jian Cheng. BGGAN: Bokeh-glass generative adversarial network for rendering realistic bokeh. *arXiv preprint arXiv:2011.02242*, 2020. 3, 4

[43] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*, 2019. 1, 3

[44] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In *ECCV*, 2020. 1, 3

[45] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *ECCV*, 2016. 3

[46] Pratul P Srinivasan, Rahul Garg, Neal Wadhwa, Ren Ng, and Jonathan T Barron. Aperture supervision for monocular depth estimation. In *CVPR*, 2018. 2, 3, 4

[47] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3D shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019. 2, 3

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, 2016. 5

[49] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *CVPR*, 2018. 1, 3

[50] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[51] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NIPS*, 2016. 3

[52] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Trans. Graph.*, 37(4):1–13, 2018. 3

[53] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 4, 5

[54] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, 2018. 3

[55] Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. DeepLens: Shallow depth of field from a single image. *ACM Trans. Graph.*, 37(6):1–11, 2018. 3

[56] Mengjiao Wang, Zhixin Shu, Shiyang Cheng, Yannis Panagakis, Dimitris Samaras, and Stefanos Zafeiriou. An adversarial neuro-tensorial approach for learning disentangled representations. *Int. J. Comput. Vis.*, 127(6-7):743–762, 2019. 1, 3

[57] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 2

[58] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *NIPS*, 2016. 2

[59] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3D objects from images in the wild. In *CVPR*, 2020. 2, 3

[60] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *CVPR*, 2020. 3, 5, 6

[61] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017. 3

[62] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. LR-GAN: Layered recursive generative adversarial networks for image generation. In *ICLR*, 2017. 3

[63] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018. 3

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[65] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 3

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3, 4, 8