Viewpoint-agnostic Image Rendering

Hiroaki Aizawa[1] Hirokatsu Kataoka[2] Yutaka Satoh[2] Kunihito Kato[1]
[1]Gifu University
[2]National Institute of Advanced Industrial Science and Technology (AIST)
aizawa@cv.info.gifu-u.ac.jp, {hirokatsu.kataoka, yu.satou}@aist.go.jp, kkato@gifu-u.ac.jp

Abstract

*Rendering an any-viewpoint image is extremely difficult for Generative Adversarial Networks. This is because conventional GANs do not understand 3D information underlying a given viewpoint image such as an object shape and relationship between viewpoint and objects in 3D space. In this paper, we present how to perform a **V**iewpoint-**A**gnostic **I**mage **R**endering (VAIR), equipping a conditional GAN with a mechanism to reconstruct 3D information of the input view. VAIR realizes any-viewpoint image generation by manipulating a viewpoint in 3D space where the reconstructed instance shape is arranged. In addition, we convert the reconstructed 3D shape into a 2D representation for image-based conditional GAN, while preserving detail 3D information. The representation consists of a depth image and 2D semantic keypoint images, which are obtained by rendering the shape from a viewpoint. In the experiment, we evaluate using a CUB-200-2011 dataset, which contains few-samples biased a viewpoint such that covers only part of the target appearance. As a result, our VAIR clearly renders an any-viewpoint image.*

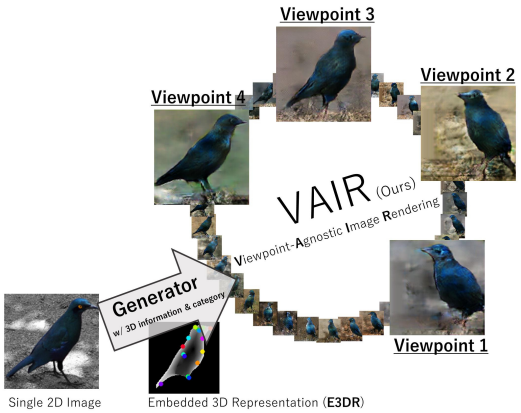


Figure 1. Overview of viewpoint-agnostic image rendering (VAIR), which enables us to render a bird's image with few samples and fine-grained dataset. We found that the following three mechanisms are necessary for any-viewpoint image generation. First, category conditioning is required to keep an object category in view rotation. Second, the self-attention mechanism mainly contributes to generating a clear foreground. Third, 3D information helps the generative models to synthesize the bird's parts and backside of birds. More detailed visual results are shown in Figures 6 and 7.

1. Introduction

Generative adversarial networks (GANs) [7] synthesize a realistic image. Recent research efforts on GANs [1, 8, 21] have improved the image-based generative model. This research trend encourages us to tackle novel problems for realistic image generation. Among them, one of the most interesting directions is the image generation and manipulation tasks by conditional GANs (cGANs). By conditioning, we are able to generate desired images based on a class [28, 25, 59], an image [11, 63, 19], and a natural language [32, 60, 54].

The cGANs are capable of controlling the generated image with object attributes and semantic labels. However, as a common limitation of these methods, 3D image manipulation and its conditioning are very difficult. Especially, generating an arbitrary-viewpoint image from an input image and a target viewpoint is challenging. This is because cGANs must generate plausible images with geometric consistency based on viewpoint transformation and preserve the identity of the instance in viewpoint changes. Moreover, the training of image-based cGANs [11] requires large amounts of viewpoint-unbiased images with 3D supervision of a camera pose and a 3D shape. In contrast, human visual perception can easily imagine a target object from many viewpoints. Because we human understands the fact that images are projections from 3D space onto a 2D plane, we first imagine a target shape and then describe other-viewpoint image by observing the shape from an imaginative viewpoint.

We believe that an explicit inference on the instance's shape is done by humans when imagining an arbitrary viewpoint image. Therefore, inspired by the above-mentioned

idea, we propose a framework that consists of a 3D reconstruction process that infers the shape from the input view, and a viewpoint hallucination process that synthesizes an arbitrary viewpoint image using 3D clues which are predicted. We call our framework **V**iewpoint-**A**gnostic **I**mage **R**endering (VAIR). VAIR makes it possible to render an any-viewpoint image naturally by manipulating the viewpoint in 3D space where the reconstructed shape is arranged. In addition, in order to connect these processes between 2D and 3D smoothly, we convert the reconstructed 3D mesh into a 2D representation while preserving detail 3D information. This representation is termed as Embedded 3D Representation (E3DR). E3DR consists of a depth image and 2D semantic keypoint images, which are obtained by rendering the 3D shape from the given viewpoint description with a differentiable renderer [15]. E3DR removes shape and viewpoint factors from the viewpoint hallucination network, and allow it to focus on synthesizing the texture of a target category. Therefore, VAIR synthesizes a viewpoint-agnostic image from viewpoint-biased few samples and a fine-grained dataset.

As another advantage on VAIR, we are able to make a better use of the potential of previous works for VAIR's two process. In this work, we utilize category-specific mesh reconstruction (CMR) [14] as a 3D reconstruction process because CMR is able to infer 3D shape, camera pose, and semantic keypoints from image collections. In a viewpoint hallucination process, we design encoder-decoder self-attention GAN in the setting conditioned on E3DR and a category label.

Undoubtedly, viewpoint-agnostic image rendering is extremely difficult with a cGAN even though some conditions are considered in the image generation. In our earlier trial, we employ viewpoint-biased and fine-grained CUB-200-2011 dataset [47] for the evaluation. In the experiment, we investigate a structure and a condition setting of VAIR. Note that we refer to 200 categories for birds in the CUB-200-2011 dataset as *category* rather than *class* in this paper. As a result, we show that VAIR with a self attention mechanism [59] generates an realistic image while controlling the given viewpoint (see Figure 1) by the E3DR-condition setting. Moreover, we reveal a difficulty in viewpoint change by cGAN and suggest open problems.

We summarize our contributions as follows:

- **Image generation with reconstructed 3D information.** We incorporate a conditional GAN with reconstructed 3D properties to acquire a concept of viewpoint change. The properties are converted into the representation for cGAN, which consists of a depth image and a 2D semantic keypoints image. VAIR enables the rendering of a viewpoint-agnostic image from a single 2D image taken from any viewpoint, while retaining the instance's shape in the input image and category-specific information such as color, shade and locations of bird's parts.

- **View-agnostic image rendering with few samples and fine-grained dataset.** VAIR was trained on the CUB-200-2011 dataset. The training from a viewpoint-biased image with few samples (e.g. back side of bird) output an ill-conditioned image. VAIR synthesizes such a self-occluded region using predicted 3D clues.

2. Related work

2.1. Generative Adversarial Networks

Since Goodfellow *et al.* first published their study on GANs [7], research on generative models has been very active. One more important development, the cGAN, has been implemented by using conditions in the Discriminator network [23, 32], adding tasks with generated image classification [28], and embedding conditions [25]. The cGANs have been applied to various problems, such as domain adaptation [11], normal-to-image translation [49], and text-to-image translation [32, 60, 54]. Especially in multi-class image generation, spectral normalization GAN [25], self-attention GAN [59], and BigGAN [2] succeeded in generating diverse objects included in the ImageNet dataset [3].

Recently, several researchers have studied GAN equipped with geometric knowledge [17, 26, 9, 27]. Especially, HoloGAN [26] and PLATONICGAN [9] have achieved to disentangle the identity and a viewpoint by randomly rigid transformation in feature space. Noguchi *et al.* [27] have proposed a camera-conditioned GAN, which is trained from a consistency loss based on the optical flow between images generated from different camera poses. While these sophisticated works are able to generate any-viewpoint images from a latent code and a target viewpoint, they cannot rotate a given input image.

Here, 3D knowledge must be considered as a geometric property to understand and generate a realistic object. We believe that the 3D knowledge enables us to successfully execute more precise and viewpoint-agnostic image generation. Unlike the conventional generative approaches without any geometric property, our proposed method employs a depth image and a 2D semantic keypoints image as geometric properties. The geometric properties are reconstructed from a given input image and are used as cues to synthesize any-viewpoint images.

2.2. 3D Reconstruction

Thanks to 3D conceptualization, we understand geometric information such as "what shape does the object have?" and "where do we see from?" The traditional yet important problem in computer vision has already been explored

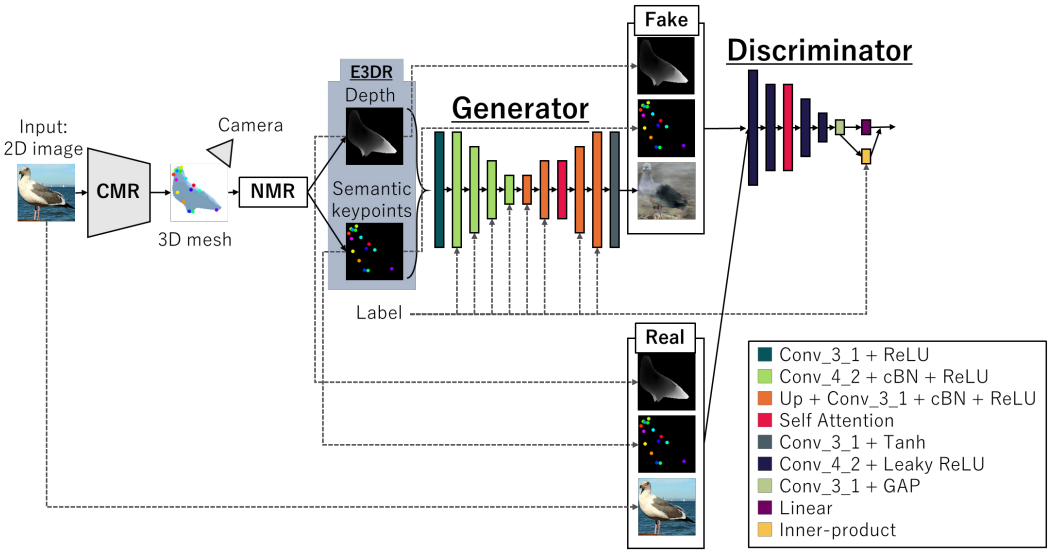


Figure 2. Flowchart of VAIR process and network architecture.: The convolution layer is denoted as Conv_{filter_size}_{stride} (e.g. Conv_3_1). We set a channel size at each convolution layer with VGG-like structure. GAP means global average pooling layer. We a assign conditional batch normalization (cBN) to embed a category into feature maps excluding related to input and output. The feature maps are upsampled using nearest neighbor interpolation (UP). In a discriminator, we incorporate projection [25] into PatchGAN's discriminator [11] to embed a category. Spectral normalization [24] is applied at all convolutional and fully-connected layers in both generator and discriminator. We initialize parameters of cGAN with kaiming uniform.

in various directions, such as image editing [57], data augmentation with synthetic images [31, 39], and domain adaptation by computer graphics [35]. We have explored how to extract 3D information. Over the past couple of years, 3D reconstruction with DNN has been progressing with voxel usage [56], multi-view masking [34, 45, 44], weakly supervised matching [13], and 3D GAN [53]. Using a neural 3D mesh renderer [15], Kato *et al.* explored the potential for calculating training loss between an image reprojected from a 3D shape and a genuine 2D silhouette. Thus, the 3D mesh renderer has become a key technology for bridging the gap between 2D and 3D rendering.

To the best of our knowledge, CMR [14] is the work closest to ours. The purpose of CMR learns to reconstruct a category-specific 3D mesh along with a texture from a 2D image input. CMR is able to render the reconstructed 3D mesh into a 2D image without a background. While Kanazawa *et al.* aim at 3D reconstruction from a single image, we mainly focus on generating viewpoint-agnostic images with a background using 3D information obtained from the 2D image input.

2.3. Novel View Synthesis

Novel view synthesis is a task that generates a new another view image including the same object or scene as in the observed input view. It is the computational equivalent of mental rotation [37]. Given single or multiple images with camera poses, in order to synthesize novel views, several works have proposed sophisticated methods such as disentangled representations [18], warping pixels [62], complementing invisible regions [30], integrating multiple views [42, 29], implicit view-invariant 3D representations [43, 6, 41, 55], and explicit reconstructed 3D shapes [64]. In addition to view synthesis on a static image, challenging tasks, such as novel view action synthesis retaining action in an input view [46, 36] and aerial-to-street view synthesis [33, 20], have been tackled. In addition, if abundant multiple views are available, we can synthesize realistic novel view from learned 3D representations by geometric rendering [40, 22, 38, 52, 58].

Our work is to synthesize any-viewpoint images by providing the concept of viewpoint rotation to GANs without 3D supervision and multiple views. To achieve this, we incorporate cGANs with the 3D reconstruction framework, which is able to train from only 2D image collections. Our VAIR explicitly infers the 3D shape from the input instance and utilize it for image generation, and aims at the CUB dataset with a background.

3. Proposed method

Figure 2 shows the flowchart of our VAIR. Our goal is to render an image with an arbitrary viewpoint from the single

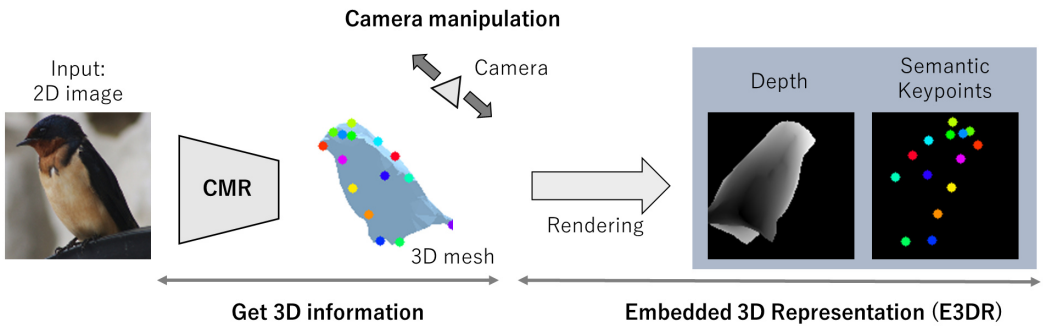


Figure 3. Variation of projected depth and semantic keypoints depending on camera viewpoint.

image. To achieve this goal, the framework must address the following problems:

- **Problem 1)** We must redesign conventional generative models effectively to represent 3D information for view-independent image generation.
- **Problem 2)** The rendered images must simultaneously preserve categories and shape in the input view.

To overcome problem 1), we introduce a 3D representation in which 3D information is embedded in a 2D image to arrange the structure and orientation of the object. To address problem 2), we condition our renderer on the reprojected silhouette in order to preserve categories and shape.

3.1. 3D Reconstruction Process

The purpose of the 3D reconstruction process is to obtain 3D information for viewpoint-agnostic image generation. From the result of a preliminary study, we found that a 3D mesh, camera pose, and 3D semantic keypoints are required to describe object's orientation and parts. We apply CMR [14] to obtain the descriptions as 3D information from a single input RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, where W is the width and H is the height.

3.2. Embedded 3D Representation

To bridge a gap of representation between the 3D reconstruction process and the following viewpoint hallucination process, we convert the 3D properties into a reprojected image with Embedded 3D Representation (E3DR). It consists of a depth image and a 2D semantic keypoints image. These images are obtained by rendering the 3D shape from the given viewpoint description with Neural Mesh Renderer (NMR) [15]. Therefore, E3DR is a viewer-oriented representation. We project estimated 3D semantic keypoints into a black background image (see Figure 3). These images with embedded 3D properties are concatenated along the channel axis and fed into the cGAN, which hallucinates an any-viewpoint image in the following process from its E3DR and the category label. E3DR with embedded 3D information is defined as $\mathbf{S}' \in \mathbb{R}^{H \times W \times c}$, where W is the width, H is the height, and c is the total number of the depth and semantic keypoint channels (c is set as four). We input the representation into our GANs proposed in Section 3.3. Throughout the camera pose manipulation and re-rendering, the cGAN acquires the concept of viewpoint changes.

3.3. Viewpoint Hallucination Process

The goal of the viewpoint hallucination process is to synthesize an arbitrary viewpoint image from a category label and E3DR. In other words, the objective of this process is to learn a function G to map from the object categories y and our 3D representation (E3DR) $\mathbf{S}$ to an RGB image $\mathbf{I}'$ taken from a manipulated viewpoint. This mapping G is defined as follows:

$$\mathbf{I}' = G\left(\mathbf{S}, y\right). \tag{1}$$

We represent its problem as conditional GANs setting, and design a viewpoint-controllable generative model by E3DR to acquire the concept of viewpoint by extending class-conditioned GANs.

Inserting categories and 3D representation. Due to the condition setting, we have to insert 3D information as E3DR $\mathbf{S}$ and the object categories y into cGAN. Therefore, we extend its generator and discriminator, respectively, based on the category-conditioned self-attention GAN (Zhang *et al.* [59]) and image-conditioned pix2pix (Isola *et al.* [11]).

We changed our generator from a self-attention GAN to the encoder-decoder structure similar to pix2pix because it can receive our E3DR $\mathbf{S}$ that is reprojected to a 2D image as an input. In addition, the object categories are represented by a one-hot tensor and then embedded in our generator using conditional batch normalization [5]. Conditional batch normalization enables us to control the categories of the generated image by adjusting the distribution of feature maps with learnable category-independent affine parameters. With these mechanisms, we can generate an arbitrary

Figure 4. Examples of artifacts. $L1$ loss between the real image and the reconstructed image creates artifacts when rendering the manipulated viewpoint. This result means that the model does not acquire generalization with respect to the viewpoint.

viewpoint image from the object categories and 3D information obtained from the edited viewpoint.

As shown in Figure 2, we employed a projection discriminator [25] that is modified so as to receive pairs of viewpoint images and E3DR as the input, based on PatchGAN's discriminator [11].

Condition setting. Technically speaking, for our condition setting, our VAIR outputs image $\mathbf{I}'$ for every possible viewpoint while conditioning 3D information as E3DR $\mathbf{S}$ and categories y. The cGANs also consist of generator G and discriminator D. The objective of our VAIR is to train cGANs through a dataset $\{(\mathbf{I}_i, \mathbf{S}_i, y_i)\}_{i=1}^{N}$ with N samples comprising image, 3D information, and categories. The generator G, which is parameterized by θ_g, reconstructs an image $\mathbf{I}' \in \mathbb{R}^{H \times W \times 3}$. The reconstructed image $\mathbf{I}'$ is realistic because the distribution in generator $p_g(\mathbf{I} \mid \mathbf{S}, y; \theta_g)$ is fitted onto conditional probability $p_{data}(\mathbf{I} \mid \mathbf{S}, y)$. The discriminator D is trained to distinguish between a real image $\mathbf{I}$ from a dataset and a generated fake image $\mathbf{I}'$. We formulate the objective function in the conditional adversarial learning as follows:

$$L_D(G, D) = -\mathbb{E}_{\mathbf{I},\mathbf{S},y}\left[min\left(0, -1 + D(\mathbf{I}, \mathbf{S}, y)\right)\right] - \mathbb{E}_{\mathbf{S},y}\left[min\left(0, -1 - D(\mathbf{I}', \mathbf{S}, y)\right)\right], \tag{2}$$

$$L_G(G, D) = -\mathbb{E}_{\mathbf{S},y}\left[D(\mathbf{I}', \mathbf{S}, y)\right]. \tag{3}$$

The goal of our cGAN is for the generator and discriminator to minimize the above adversarial losses L_G and L_D, respectively. The conditions cause the image to be generated from several viewpoints to obtain a more sophisticated representation. The shape and orientation is created from 3D information. Moreover, conditioned categories allow the generated image to retain its object category when the viewpoint rotates.

Improvement of image rendering. Through experiments, we confirmed that loss L_{L1} [11] between the real images $\mathbf{I}$ and reconstructed images $\mathbf{I}'$ creates artifacts when rendering the manipulated viewpoint, as shown in Figure 4. Therefore, to mitigate these artifacts, we employed feature matching loss [48] and perceptual loss [12, 4], which calculate the error of the feature space instead of the image space. The reconstruction loss L_{recon} that combines these losses is defined as follows:

$$L_{recon}(G, D) = \mathbb{E}_{\mathbf{I},\mathbf{S},y}\left[\sum_{i=1}^{T_D} \frac{1}{M_i} \left\| D^i(\mathbf{I}) - D^i(\mathbf{I}') \right\|_1\right] + \mathbb{E}_{\mathbf{I},\mathbf{S},y}\left[\sum_{i=1}^{T_F} \frac{1}{N_i} \left\| F^i(\mathbf{I}) - F^i(\mathbf{I}') \right\|_1\right], \tag{4}$$

where the first term on the right-hand side of equation (4) is feature matching loss, $D^i(\cdot)$ is feature maps on the i-th layer of the discriminator, T_D is the number of applied layers, and M_i is the number of elements in each layer. The second term on the right-hand side of equation (4) is perceptual loss, $F^i(\cdot)$ is feature maps on the i-th layer of the pretrained VGG network, T_F is the number of applied layers, and N_i is the number of elements in each layer.

To achieve the goal, our generator and discriminator minimized the following full objectives $L_{optimize_G}$ and $L_{optimize_D}$, respectively:

$$\begin{aligned} L_{optimize_G} &= \lambda_{GAN} L_G + \lambda_{recon} L_{recon}, \\ L_{optimize_D} &= \lambda_{GAN} L_D, \end{aligned} \tag{5}$$

where λ_{GAN} and λ_{recon} adjust the balance between the adversarial loss and the reconstruction loss.

4. Evaluation

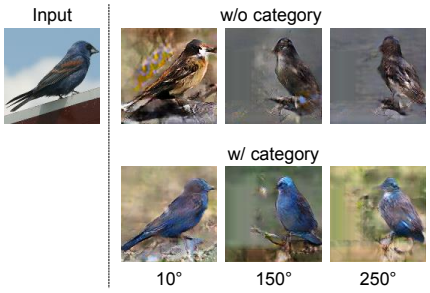
We perform experiments from two perspectives: (i) We qualitatively and quantitatively evaluate images which are generated by our proposed methods using metrics based on similarity between images of the same viewpoint, (ii) we qualitatively compare our proposed model with conventional ones in terms of image quality.

4.1. Experimental Settings

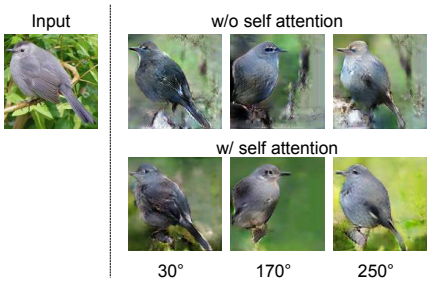
Caltech-UCSD Birds 200 (CUB-200-2011) dataset [47]. The dataset contains 200 bird species with fine-grained categories. The total number of image is 11,788. Based on the CMR setting [14], we divide the original CUB-200-2011 dataset into training 5,964, validation 2,874 and test 2,874. We can access a bounding box and 14 semantic keypoints per image. The dataset has few-training samples per category and the bias on viewpoint for each category. Concretely, our model is forced to learn a concept of viewpoint change from only 30 viewpoint-biased training images per category. In the dataset, therefore, we have

Table 1. Exploration study with SSIM/MS-SSIM and LPIPS. (SSIM/MS-SSIM: higher is better, LPIPS: lower is better. SA: self-attention, Recon: reconstruction loss with feature matching loss and perceptual loss, L1: L1 loss, Depth: w/ depth image, Category: w/ conditioned category, KP: w/ semantic keypoints, Mask: w/ binary mask image.)

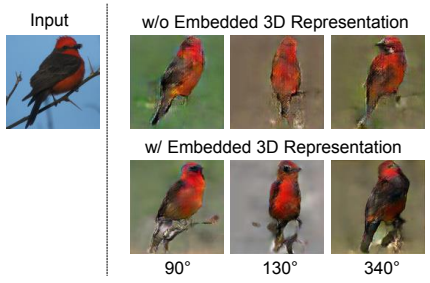
Configuration						SSIM	MS-SSIM	LPIPS
			Recon		SA	.067	.447	.304
Mask	Category		Recon		SA	.073	.486	.295
Depth			Recon		SA	.068	.451	.305
Depth	Category		Recon		SA	.073	.489	.281
Depth	Category	KP		L1	SA	.068	.471	.315
Depth	Category	KP	Recon	L1	SA	.070	.474	.308
Depth	Category	KP	Recon			**.077**	.491	.287
Depth	Category	KP	Recon		SA	.075	**.493**	**.278**



(a) Category conditioning



(b) Self-attention mechanism [59]



(c) 3D information

Figure 5. Build-up image rendering with several parameters. In the exploration study, we reveal that (a) w/ category, (b) w/ self attention and (c) w/ 3D information (depth image and semantic keypoints by adjusting camera pose) are important configuration for viewpoint-agnostic image rendering.

to acquire a concept of viewpoint change from only 30 training images per category.

Implementation details. In the 3D-based representation, we refer to the CMR [14] for category-specific 3D mesh, camera pose and 3D semantic keypoints. We render 128×128 images to connect with a cGAN. Therefore, we modify the number of embedding dimensions which is fed into the fully-connected layers from a 4096-D vector to 1024-D vector. We train the modified CMR in the fully supervised manner. The trained parameters are fixed during training of VAIR.

We optimize the full objectives (Equation (5)) with Adam [16], where $\beta_1 = 0.0$ and $\beta_2 = 0.999$. We apply two-timescale update rule (TTUR) [10], as with self-attention GAN [59] to stabilize the training. Learning rates in the generator and discriminator are 0.0001 and 0.0004 respectively. We update a discriminator 5 times at each generator's update. We stop the training of VAIR after 150 epochs. We set $\lambda_{GAN} = 1.0$ and $\lambda_{recon} = 10.0$. The batch size is set as 32. Training with a generative model roughly takes 2 days using a NVIDIA Quadro P6000.

Performance measurement with cGANs. We quantitatively evaluate the effectiveness of our build-up generative models in terms of image quality. To verify view-agnostic images in object category and shape, we employ Structured Similarity (SSIM) [50], Multiscale SSIM (MS-SSIM) [51], and Learned Perceptual Image Patch Similarity (LPIPS) [61] for evaluation methods. SSIM/MS-SSIM evaluate a similarity-driven image quality between two different images. These metrics mainly focus on quality for a pixel value, contrast, and structure. On one hand, LPIPS is claimed as a metric which represents a human-like perception [61]. The evaluation method takes a lower value when a view-agnostic image preserves the same object category and shape as the input image. We calculate these similarity between the input image and the generated image taken

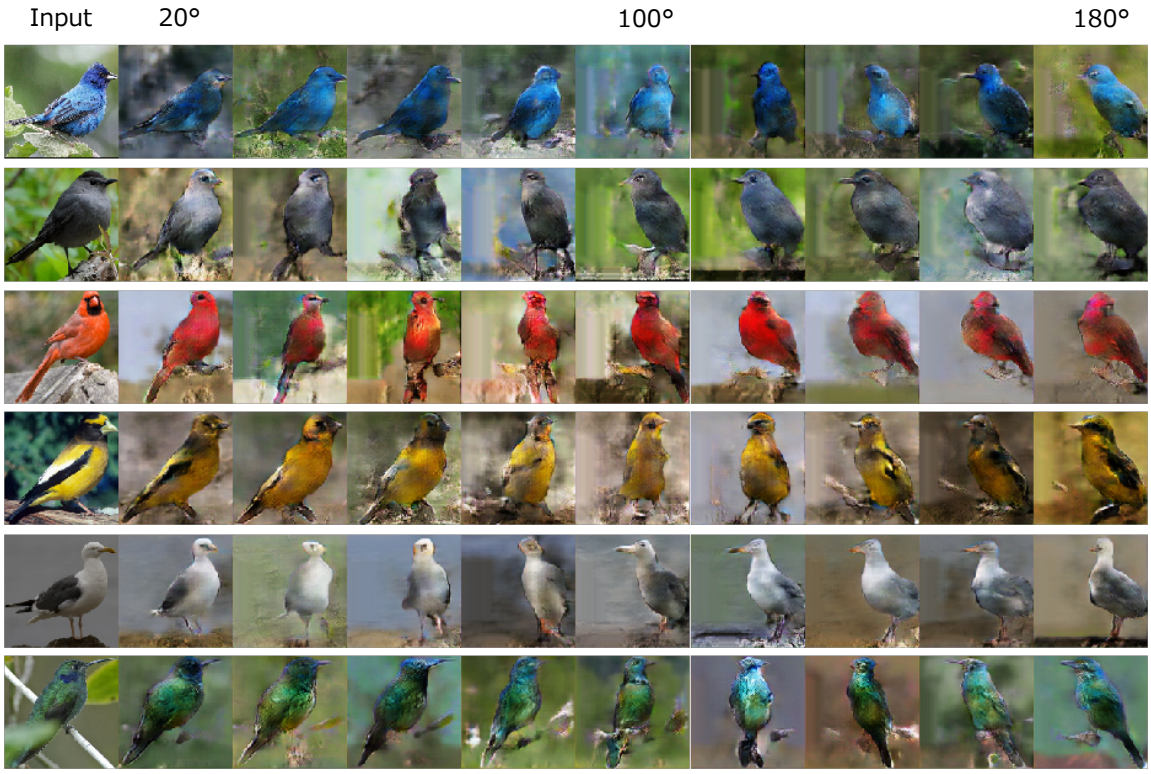


Figure 6. Visual results. The figure indicates rotated images from 0 to 180 [degree] at each 20 degree.

from predicted camera poses using these metrics because we do not have any ground truth images.

4.2. Experimental Results

Results. Table 1 shows exploration study with SSIM/MS-SSIM and LPIPS as evaluation metrics. According to the table, the configuration with {self attention, feature matching loss, perceptual loss, depth, category, key points} is the best score (lower is better) in LPIPS. The similar way without self-attention mechanism is the best in terms of SSIM which achieved .077 in the experiment. Basically, we have increased the evaluation scores in LPIPS scores, we confirmed that our model conditioned on 3D information and the category achieved .278 which is the best score at all models. Therefore, our modification effects an improvement from a conventional image-conditioned GAN. Moreover, Figures 5 visually shows our build-up results with (w/) or without (w/o) conditioned category, self-attention mechanism and 3D information based on our E3DR (depth image and semantic keypoints). The details are described as follows.

Category conditioning. We compared the unconditioned model and category-conditioned one. These results are shown in Figure 5(a) which teaches us to utilize a conditioned category into a cGAN. Without a conditioned category, a generated result cannot be maintained the category in an input image. In contrast, the category-conditioned generative model (w/ category) keeps the category from the input image.

Self attention. We confirmed the improvement of image sharpness by the self attention mechanism in Figure 5(b). As Figure 5(b) indicates, the generative model without self attention focuses the whole image including background, therefore the birds contain blurred edges inside of foreground areas. In contrast, because the generative model with self attention mainly generates the birds in foreground areas, the generated image has a sharpness. As the result of self-attention mechanism, the LPIPS score was improved from .287 to .278 (lower is better in LPIPS).

Embedded 3D representation. We compared a binary mask (w/o 3D information) and with our 3D representation (w/ 3D information) to verify our contribution based on E3DR. The experiment results are shown in Figure 5(c) and Table 1. Figure 5(c) tells that our 3D representation with depth image and semantic keypoints enables a model to generate a viewpoint-agnostic image. As we can see in Figure 5(c), our rendering successfully reconstructs an any-

Figure 7. Comparison of Our VAIR vs. CMR [14]. Compared to CMR with CG-like artifacts, our VAIR generates more realistic images with a background.

viewpoint image including the back side of bird. The model without a 3D representation is narrowed down to render a blurred image. We can understand that E3DR is required in order to get a concept of an object's rotation and locate object's parts. Moreover, our E3DR also contributed to improve the image quality in addition to the any-viewpoint rendering. The results clearly show that 3D information is necessary to achieve the goal of viewpoint-agnostic image rendering.

Feature matching loss [48] and perceptual loss [12, 4] without L1 loss. We confirmed that the score with L1 loss clearly decreased from the score without the loss. As shown in Figure 4, an effective loss function is one of the solutions for a sharp image generation. The scores in Table 1 suggest that L1 loss drops a performance of realistic image generation. The main cause of the results is degradation on viewpoint variation by the pixel-level loss calculation. These evidences with bad results understand us that L1/L2 distance is not suitable for viewpoint-agnostic image generation.

Visual results with VAIR (ours). The rendering images with our VAIR are shown in Figure 6. The figure reports rotated images from 0 to 180 [degree] at each ten degree. Our VAIR successfully enables to generate images which are trained with 30 training instances per category in the dataset. The proposed method can render any-viewpoint images including front, side and backside views.

Comparison with CMR. We visually compare our VAIR and conventional CMR in Figure 7. Once look at the visualization, the rendered images with CMR seem to have a so-called CG-like boundary around edges. Unlike the CMR, our VAIR along with background enables to generate a clear foreground. Although the merit of CMR is color mapping with a texture of input image, however, the characteristic means that an occluded object is directly projected into a rendered image. The both methods cannot effectively generate a background area. Our VAIR slightly acquires how to render a background, but a more realistic scene rendering will be achieved in the future. Hereafter we are ready for building up a generative model to render view-independent images.

5. Conclusion

We proposed Viewpoint-Agnostic Image Rendering (VAIR) that contributes a successful view-independent image generation from a single 2D image. We have achieved a clear and view-rotated image rendering through an embedded 3D representation (E3DR) with a depth image and semantic keypoints, even though the acquiring a concept of viewpoint change is undoubtedly difficult in image generation. On the hard issue, we started with collapsed image generation with a representative conditional GAN (e.g. pix2pix), we have greatly improved the rendering results by assigning E3DR and some techniques like self-attention mechanism. In the experimental section, we have conducted exploration study how to improve the rendering in viewpoint changes. The visualization results described that our VAIR enables to generate a clearer rendering image.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

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

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

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

[5] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017.

[6] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[7] 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.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, 2017.

[9] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *ICCV*, 2019.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

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

[13] Angjoo Kanazawa, David W Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016.

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

[15] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[17] Jean Kossaifi, Linh Tran, Yannis Panagakis, and Maja Pantic. Gagan: Geometry-aware generative adversarial networks. In *CVPR*, 2018.

[18] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015.

[19] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[20] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *CVPR*, 2020.

[21] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, 2017.

[22] 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.

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

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

[25] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *ICLR*, 2018.

[26] 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.

[27] Atsuhiro Noguchi and Tatsuya Harada. Rgbd-gan: Unsupervised 3d representation learning from natural image datasets via rgbd image synthesis. In *ICLR*, 2020.

[28] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017.

[29] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. Transformable bottleneck networks. In *ICCV*, 2019.

[30] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, 2017.

[31] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *ICCV*, 2015.

[32] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

[33] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *CVPR*, 2018.

[34] Danilo Jimenez Rezende, SM Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. In *NIPS*, 2016.

[35] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016.

[36] Kara Schatz and Erik Quintanilla. A recurrent transformer network for novel view action synthesis. In *ECCV*, 2020.

[37] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

[38] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020.

[39] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017.

[40] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019.

[41] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NIPS*, 2019.

[42] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *ECCV*, 2018.

[43] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016.

[44] Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *CVPR*, 2018.

[45] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017.

[46] Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction. In *ECCV*, 2020.

[47] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[48] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

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

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[51] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003.

[52] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, 2020.

[53] 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.

[54] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.

[55] Xiaogang Xu, Ying-Cong Chen, and Jiaya Jia. View independent generative adversarial network for novel view synthesis. In *ICCV*, 2019.

[56] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*, 2016.

[57] Shunyu Yao, Tzu Ming Harry Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, Bill Freeman, and Joshua Tenenbaum. 3d-aware scene manipulation via inverse graphics. In *NIPS*, 2018.

[58] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *CVPR*, 2020.

[59] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[60] Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[61] 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.

[62] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016.

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

[64] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *NIPS*, 2018.