Spherical Image Generation from a Single Image by Considering Scene Symmetry

Takayuki Hara[1], **Yusuke Mukuta**[1, 2], **Tatsuya Harada**[1, 2]
[1]The University of Tokyo
[2]RIKEN
{hara, mukuta, harada}@mi.t.u-tokyo.ac.jp

Abstract

Spherical images taken in all directions (360°×180°) allow the full surroundings of a subject to be represented, providing an immersive experience to viewers. Generating a spherical image from a single normal-field-of-view (NFOV) image is convenient and expands the usage scenarios considerably without relying on a specific panoramic camera or images taken from multiple directions; however, achieving such images remains a challenging and unresolved problem. The primary challenge is controlling the high degree of freedom involved in generating a wide area that includes all directions of the desired spherical image. We focus on scene symmetry, which is a basic property of the global structure of spherical images, such as rotational symmetry, plane symmetry, and asymmetry. We propose a method for generating a spherical image from a single NFOV image and controlling the degree of freedom of the generated regions using the scene symmetry. To estimate and control the scene symmetry using both a circular shift and flip of the latent image features, we incorporate the intensity of the symmetry as a latent variable into conditional variational autoencoders. Our experiments show that the proposed method can generate various plausible spherical images controlled from symmetric to asymmetric, and can reduce the reconstruction errors of the generated images based on the estimated symmetry.

Introduction

Spherical images capturing all possible directions (horizontal 360° × vertical 180°) are used in various domains such as surveillance systems, construction industry, tourism, autonomous cars, and entertainment. They can capture environments around the main subject and represent the entire space. Furthermore, when viewed through a head-mounted display, spherical images allow one to enjoy a scene in a more immersive manner. However, capturing spherical images is not an easy task, as doing so requires a specific panoramic camera or specific software that stitches together images taken from multiple directions. Therefore, it would be more convenient to generate a spherical image from a single normal-field-of-view (NFOV) image which is a perspective projection image taken by a normal camera. Furthermore, it will considerably expand the usage scenarios. For example, the background of virtual reality content can be created using a single conceptual photograph, which can reduce production time and costs dramatically. In addition, generating peripheral views of images taken can allow viewers to enjoy content with a feeling of immersion (Kimura and Rekimoto 2018). Finally, estimating an illumination environment map will allow rendering an object with realistic relighting for use in augmented reality (Gardner et al. 2017; Song and Funkhouser 2019). In these applications, it is better to reconstruct the actual scene as it is, however it is worthwhile if we can obtain an image that is perceptually plausible to humans, even if it is different from the actual scene.

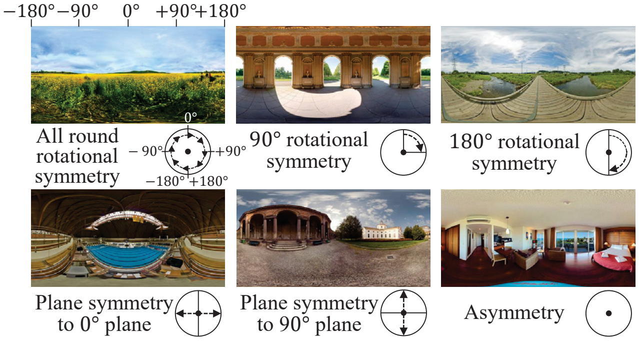

Figure 1: Symmetry types of spherical images in SUN360 dataset (Xiao et al. 2012). Images are represented through a equirectangular projection (see Figure 2), where the arrows in the circles correspond to the viewpoint transitions that do not significantly change the appearance.

However, because an NFOV image captures only a small portion of the whole view [1], two main challenges exist while generating a spherical image from a single NFOV image: generating an image corresponding to the spherical structure, and controlling the high degree of freedom involved in generating a wide area, which includes all directions of a plausible spherical image.

Generating a spherical image from a single image is related to the task of image completion, which predicts an entire image from a partial one. However, conventional image-

[1] We assume a FOV of about 120 degrees even on the wide-angle side in reference to commercial cameras.

completion methods (Li et al. 2017; Iizuka, Simo-Serra, and Ishikawa 2017; Zheng, Cham, and Cai 2019; Zhao et al. 2020) are unsuitable for generating spherical images because such methods are designed for planar images and cannot handle a spherical distortion or continuity. A spherical-image-completion method using a single image was recently proposed (Gardner et al. 2017; Song and Funkhouser 2019; Akimoto et al. 2019). This method employs a two-dimensional (2D) convolutional neural network (CNN) in an equirectangular projection (Figure 2) of a spherical image; however, there are two problems with this approach: (1) they have a discontinuity at the left and right ends of the equirectangular projection, and (2) they cannot control the content of the generated regions. Therefore, unlike conventional methods, we generate a spherical image without a discontinuity and control the aforementioned degree of freedom to obtain plausible variations of the desired spherical image.

There are various factors controlling spherical image generation. Among these factors, *scene symmetry* is a basic property of a global structure of a spherical image. As pointed out in (Xiao et al. 2012), a 360° view has a specific type of symmetry structure for place categories. We interpret it as the scene symmetry that specific geometric operations such as a rotation and flip do not change the appearance significantly. In Figure 1, the typical types of symmetry are depicted, such as rotational symmetry, plane symmetry, and asymmetry. Furthermore, since the scene rarely changes dramatically within a single spherical image, we consider that there are many spherical images whose global structure has a certain level of symmetry combination. Based on this consideration, we control the intensities of the various type of scene-symmetry to improve the quality and diversity of the generated images.

The proposed method incorporates the intensity of the symmetry as a latent variable in a conditional variational autoencoder (CVAE) (Sohn, Lee, and Yan 2015) to estimate the distribution of possible symmetry from a single image and control the symmetry of a generated spherical image. In this study, we target the goal of spherical image generation into the following two points: (1) reconstructing the actual scene and (2) generating plausible variations of the scene, which includes the input image. For (1), we reconstruct the spherical image with the mean of the estimated distribution of symmetry, and for (2), we generate a plausible spherical image with the value sampled from the distribution. We employ a CNN on an equirectangular projection with a circular padding that eliminates the discontinuity, and scene symmetry is implemented using both a circular shift and a flip of the hidden variables.

The contributions of this paper can be summarized as follows.

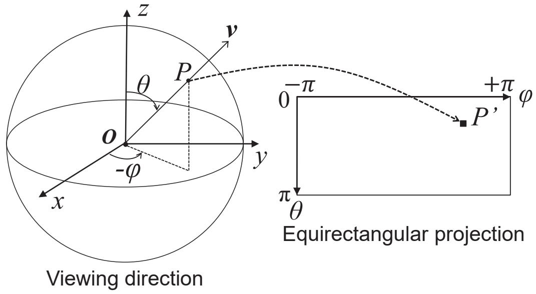

Figure 2: Equirectangular projection. Point P on a 2D sphere is converted into point P' on the θ-ϕ plane.

- We propose a novel spherical image generation method from a single NFOV image, which improves the reproducibility and plausibility of the generated images by leveraging the symmetry of the scene.
- We propose a network architecture that employ CNNs on an equirectangular projection with a circular padding to generate a spherical image without a discontinuity.
- We design a new architecture that can estimate and control the symmetry of the image by using a circular shift and a flip of the hidden variables of a CNN.
- We demonstrate that our proposed method can generate multiple spherical images, controlled from symmetric to asymmetric.

Related Work

Image Completion

Several image completion technologies have been proposed thus far for predicting the missing regions of an image. Traditionally, numerous diffusion-based methods (Ballester et al. 2001; Bertalmio et al. 2000) diffuse the information of the visible regions into the missing regions, and multiple patch-based methods (Criminisi, Perez, and Toyama 2003; Barnes et al. 2009) complete the missing regions by matching, copying, and realigning using the visible regions. Both methods assume that the missing regions contain information correlated with the visible regions.

Generative models, which are trained using large-scale datasets, such as a variational autoencoder (VAE) (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014) and generative adversarial networks (GANs) (Goodfellow et al. 2014), have experienced a significant boost, and both of these models have been adopted for image completion. Li et al. (2017) directly generated the contents of missing regions using CNNs (Fukushima and Miyake 1982; LeCun et al. 1989) with a combination of the reconstruction loss, semantic parsing loss, and two adversarial losses. Furthermore, Iizuka, Simo-Serra, and Ishikawa (2017) employed global and local context discriminators in the framework of adversarial learning to improve the naturalness and consistency of the completed regions.

While most image-completion methods produce only one result for each input, Zheng, Cham, and Cai (2019); Zhao et al. (2020) presented a method for generating multiple and diverse plausible solutions for image completion using CVAEs; however, these methods are designed for planar images and are not suitable for spherical images. In addition, these methods cannot explicitly control the generated contents. Therefore, unlike any of the previous methods, we pro-

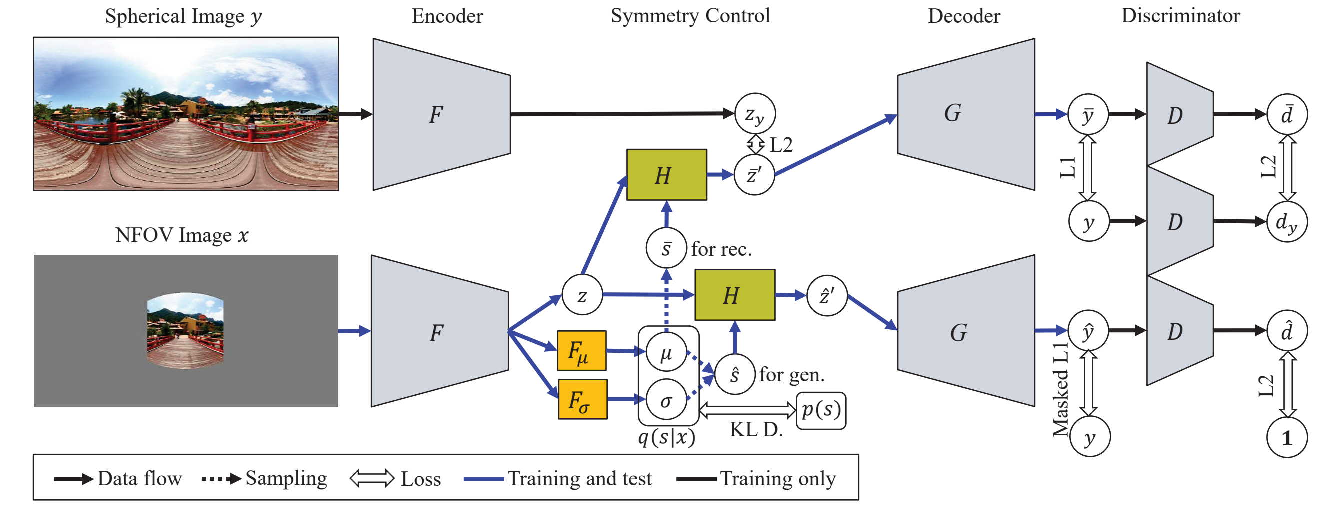

Figure 3: Structure of our proposed method for spherical-image generation. During training, the spherical image y and the NFOV image x are used. The accents $\bar{\cdot}$ and $\hat{\cdot}$ represent the variables in the reconstruction setting and the generation setting respectively. During testing, a spherical image $\bar{y}$ or $\hat{y}$ is generated from a single x.

pose a spherical image generation method that can explicitly control the symmetry.

Panoramic and Spherical Image Generation

Zhang et al. (2013) proposed extrapolating an NFOV image into a panoramic image using the panoramic image whose scene category is same to that of input image as a guide. In Kimura and Rekimoto (2018), the authors proposed a peripheral-image-generation method based on pix2pix (Isola et al. 2017); however, the field of view of the generated image is limited. In Sumantri and Park (2020), a spherical-image-generation method is proposed that requires a set of images captured from multiple directions as an input. Panoramic three-dimensional structure prediction methods have also been proposed (Song et al. 2018; Srinivasan et al. 2020), although such methods require RGB-D data or stereo pairs of images. In Gardner et al. (2017); Song and Funkhouser (2019); Akimoto et al. (2019), the authors proposed a spherical image completion method using a single NFOV image. However, such methods have discontinuity in the generated image and cannot control the content of the generated regions. Unlike these conventional research approaches, we generate a spherical image without discontinuity and control the aforementioned degree of freedom to obtain plausible variations of the desired spherical image.

CNN for Spherical Signals

Ordinal 2D CNNs are designed based on the translational equivariance on a plane; thus, they cannot be directly applied to spherical images. With a CNN-based approach for spherical signals, planar CNNs are applied on the projected plane, such as the equirectangular projection (Su and Grauman 2017; Assens et al. 2018) depicted in Figure 2, tangent planes (Tateno, Navab, and Tombari 2018), and cube mapping (Cheng et al. 2018). Other approaches that do not use 2D planar CNNs, such as generalized FFT-based CNNs on a sphere (Cohen et al. 2018; Esteves et al. 2018), distortion-aware sampling (Lee et al. 2018; Benjamin, Condurache, and Geiger 2018), and graph convolution on a sampled grid on a sphere (Khasanova and Frossard 2017; Perraudin et al. 2019; Yang et al. 2020), have also been proposed.

Although projection distortion is caused when using methods based on planar CNNs, some realistic spherical images were generated by using an equirectangular projection, though from multiple images (Sumantri and Park 2020). Furthermore, using an equirectangular projection offers advantages such as a symmetrical operation around the gravity axis, which can be realized using a circular shift and a flip with low computational requirements. In this study, we employ equirectangular-projection-based CNNs.

Proposed Method

This section describes our proposed method, which generates a spherical image y from an NFOV image x by using a symmetry intensity $s \in \mathbb{R}^C$. The term s corresponds to the intensity of C types of rotational symmetry around the gravity-axis, and plane symmetry to vertical planes (the gravity-axis is aligned with the negative direction of the z-axis in Figure 2). To generate diverse images conditioned with an NFOV image, we employ CVAEs as a base framework and incorporate the symmetry intensity as a latent variable. The entire network structure is depicted in Figure 3. The details of our method are as follows.

Controlling Scene Symmetry

Similar to previous studies (Song and Funkhouser 2019; Akimoto et al. 2019; Sumantri and Park 2020), we employ an encoder–decoder based spherical image generation method using CNNs on an equirectangular projection, which takes an NFOV image x as an input and a spherical image y as an output. In this framework, we obtain hidden variables

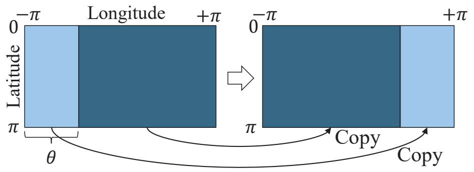

Figure 4: Circular shift. A circular shift is used for estimating and controlling the rotational symmetry.

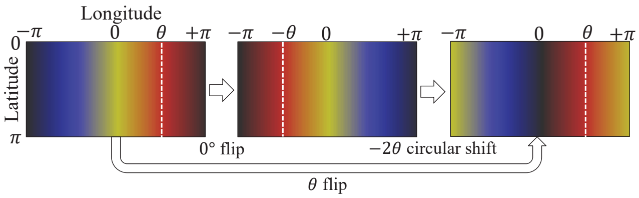

Figure 5: Flip on the θ axis. This operation combines a 0° flip and -2θ circular shift. The flip is used for estimating and controlling the plane symmetry.

$z = F(x) \in \mathbb{R}^{H_z \times W_z \times D_z}$ and generate $y = G(z)$, where F is an encoder and G is a decoder. Because both F and G consist of full CNNs, the hidden variables z have the location information of x and y. We focus on this property and control the symmetry of the generated image y by applying a symmetry-control function to the hidden variables z.

We define the symmetry-control function H, which takes the weighted linear sum of the symmetric transformations $\{T_i\}$ of z as follows:

$$H(z, s) = \frac{w \odot z + \sum_{i=1}^{C} \zeta(s_i) T_i(w \odot z)}{w + \sum_{i=1}^{C} \zeta(s_i) T_i(w)}, \tag{1}$$

where $s = (s_1, s_2, \cdots, s_C) \in \mathbb{R}^C$ is a symmetry intensity, $\zeta : \mathbb{R} \to (0, 1)$ is a sigmoid function, $w \in \mathbb{R}^{H_z \times W_z \times D_z}$ denotes a weight vector, $\odot$ is an elementwise product, and the quotient is elementwise. The symmetric transformations $\{T_i\}$ are rotations about the gravity-axis and flip along the vertical plane. In an equirectangular projection, a rotation about the gravity-axis corresponds to a horizontal circular shift, as depicted in Figure 4. The horizontal circular shift of the m elements can be written as $S^{(m)} : (z_{i,j,k}) \mapsto (z_{i,(j+m)\mathrm{mod} W_z + 1, k})$. Flip along the longitude θ plane corresponds to a 0° flip and a -2θ circular shift, as depicted in Figure 5. In this manner, we represent both rotational and plane symmetry-control functions through a circular shift and a flip of the hidden variables, respectively.

As mentioned above, we obtain a symmetry-controlled hidden variable $z' = H(z, s)$, and generate a spherical image $y = G(z')$.

CVAEs with Symmetry Intensity

Next, we describe how to incorporate the intensity of the symmetry as a latent variable in a CVAE. Figure 6 shows the assumed graphical model which indicates the causal relationship between variables, and the joint distribution is expressed as follows:

$$p(x, y, s) = p(y|x, s) p(x) p(s). \tag{2}$$

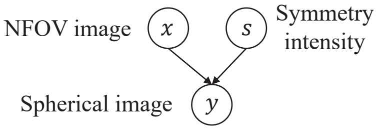

Figure 6: The graphical model indicates the causal relationship of a spherical image, an NFOV image, and a symmetry intensity.

Under this assumption, the variational lower bound of the conditional log-likelihood of a NFOV image x for a spherical image y is derived as follows:

$$\begin{aligned} \log p(y|x) \geq & -\mathbb{KL}(q_\psi(s|x) || p(s)) \\ & + \mathbb{E}_{q_\psi(s|x)}[\log p_\omega(y|x, s)]. \end{aligned} \tag{3}$$

In this formulation, we restrict the conditional variables of the variational posterior distribution q_ψ to x to learn the estimator of s from x. We obtain the conditional distribution $p_\omega(y|x, s)$ and $q_\psi(s|x)$, maximizing the variational lower bound of the likelihood for the training data $\{y^{(n)}, x^{(n)}\}_{n=1}^{N}$ (where s is not required).

Probability Distribution Settings

The prior and posterior distributions for the latent variable s are set as follows:

$$p(s) = \mathcal{N}(s|0, I) \tag{4}$$

$$q_\psi(s|x) = \mathcal{N}(s|F_\mu \circ F(x), \nu \mathrm{diag}(F_\sigma \circ F(x))^2) \tag{5}$$

where $\mathcal{N}(\cdot|\mu, \Sigma)$ denotes the probability density function of a normal distribution with the mean vector μ and the covariance matrix Σ. Here, F_μ, F_σ denote the functions that estimate the mean and variance of s, respectively; ν is a hyper parameter; and $\mathrm{diag}(a)$ is a diagonal matrix whose components are vector a.

In addition, the likelihood function is set as follows:

$$\begin{aligned} \log p_\omega(y|x, s) = & \alpha || m \odot (y - G(z')) ||_1 \\ & + \beta || F(y) - z' ||_2 + \mathrm{const}. \end{aligned} \tag{6}$$

where $z' = H(F(x), s)$ is a symmetry controlled hidden variable, $G(z')$ is a generated spherical image, $\alpha, \beta \leq 0$ are hyper parameters, and m is a masking vector. The likelihood is high when the generated spherical image $G(z')$ and the symmetry-controlled variable z' are close to the ground truth.

After training, we obtain $q_\psi(s|x)$ and $p_\omega(y|x, s)$, which means that functions F, F_μ, F_σ, and G are determined.

Settings for Reconstruction and Generation

We employ the combination of two kinds of settings for reconstruction and generation, as inspired by PICNet (Zheng, Cham, and Cai 2019). However, there is a definite difference in that PICNet uses two types of inputs, i.e., a spherical image y and an NFOV image x, whereas we use two types

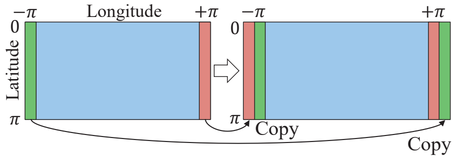

Figure 7: Circular padding. A circular padding eliminates the discontinuity between the left and right edges of the equirectangular image for a convolution.

of hyper-parameter settings of the distributions $q_\psi(s|x)$ and $p_\omega(y|x,s)$. In the reconstruction setting, $\nu \to 0$, $\alpha = \alpha_{\rm rec}$, $\beta = \beta_{\rm rec}$, and all elements of m are fixed to 1. In the generation setting, $\nu = 1$, $\alpha = \alpha_{\rm gen}$, $\beta = 0$, and m is set to a constant vector whose elements take a 1 in the overlapping area of y and x, and a 0 in the other area. The aim of the reconstruction setting is to reconstruct the actual scene using the mean of $q(s|x)$, and the aim of the generation setting is to generate plausible variations of the scene, which includes the input image using the sampled value of $q(s|x)$.

We denote the negative lower bound of Eq. (3) with the reconstruction setting as $\mathcal{L}_{\rm rec}(y|x)$, and with the generation setting as $\mathcal{L}_{\rm gen}(y|x)$. We minimize the following loss function by adding each function with the ratio of $\eta \in [0,1]$. We therefore have the following:

$$\mathcal{L}(y|x) = \eta\mathcal{L}_{\rm rec}(y|x) + (1-\eta)\mathcal{L}_{\rm gen}(y|x). \quad (7)$$

Adversarial Learning

Because a VAE tends to produce blurred images, we employ adversarial learning combined with a VAE, as used in (Larsen et al. 2015; Bao et al. 2017; Zheng, Cham, and Cai 2019). We add adversarial losses, $\mathcal{L}^{\rm ad}_{\rm rec}$ and $\mathcal{L}^{\rm ad}_{\rm gen}$, to the reconstruction loss $\mathcal{L}_{\rm rec}$ and the generation loss $\mathcal{L}_{\rm gen}$, respectively, as follows:

$$\begin{aligned} \mathcal{L}^{\rm ad}_{\rm rec} &= \gamma||D(y) - D(G(z'))||_2 \\ \mathcal{L}^{\rm ad}_{\rm gen} &= \gamma||\mathbf{1} - D(G(z'))||_2^2 \end{aligned} \quad (8)$$

where D is a function that outputs confidence to discriminate real images for multiple regions, $\mathbf{1}$ is a constant vector whose elements all take a value of 1, and γ is a hyper prameter.

First, we learn F, F_μ, F_σ, and G to maximize the evaluation function of Eq. (7), including Eq. (8). Subsequently, based on a LSGAN (Mao et al. 2017) for each mini-batch data, we learn D to minimize the following loss function:

$$\mathcal{L}_D = {\rm E}_{p(y)}[||\mathbf{1} - D(y)||_2^2] + {\rm E}_{p(y|x,s)}[||D(y)||_2^2]. \quad (9)$$

Network Structure

Functions F, F_μ, F_σ, G, and D are implemented using full CNNs. To eliminate the discontinuity between the left and right ends of the equirectangular image, we employ circular padding before each convolution layer. As depicted in Figure 7, the circular padding copies $(k-1) \bmod 2$ columns from the left and right edges of the image to each opposite side in the equirectangular image, where k denotes the kernel size of the convolution.

In addition, Figure 3 depicts the entire network structure of the proposed method. The input spherical image y and NFOV image x are represented using equirectangular projections (see Figure 2). The encoder F calculates the feature vectors z and z_y for x and y, respectively. In addition, the estimators F_μ and F_σ calculate $\bar{s}, \hat{s}$ using the *reparameterization trick* (Kingma and Welling 2013) for each reconstruction and generation setting, respectively. The symmetry controller H outputs a symmetry controlled variable $\bar{z}', \hat{z}'$, and the decoder G generates spherical images $\bar{y}, \hat{y}$. During the training phase, the loss functions are minimized as described above.

Experiments

We conducted experiments to verify the effectiveness of the proposed method. To this end, we used the Sun360 dataset (Xiao et al. 2012), which includes various spherical images, both symmetric and asymmetric, from indoor to outdoor scenes. The data were divided into 50,000 images for training, 10,000 images for testing, and 5,000 images for validation. The spherical image was an RGB image of the equirectangular format with a resolution of 256×512 pixel. Furthermore, a partial (NFOV) image was cropped from a spherical image with a 30° to 120° field of view and an aspect ratio of 1:1, following which the viewpoint direction was randomly set on the sphere and projected onto the equirectangular image, the margin of which was filled with gray values.

Implementation Details

We trained the networks from scratch using the Adam optimizer (Kingma and Ba 2014) with a fixed learning rate of 10^{-4} and a mini-batch size of 8. During the optimization, the weighting factors of the loss function were set as $\alpha_{rec} = \alpha_{gen} = -5.0 \times 10^{-4}$, $\beta_{rec} = -3.6 \times 10^{-3}$ and $\gamma = 0.27$; the element of weight vector w corresponding to position v on the unit sphere is $w(v) = \exp(3\langle c, v\rangle)$, where $\langle c, \cdot\rangle$ is an inner product of the center position c of the input image; and the mixture ratio of the two approximations was set to $\eta = 0.5$. Furthermore, we considered $C = 5$ types of symmetry, namely, the 90°, 180°, and 270° rotational symmetries and plane symmetries along the 0° and 90° axes.

For a fair comparison, the functions F, G, and D have the same configuration with PICNet (Zheng, Cham, and Cai 2019), with the following two exceptions: First, we employ circular padding before each convolution layer, and second, D outputs a confidence for the entire image using an additional four-layer ResBlock (Zheng, Cham, and Cai 2019) as a global discriminator, and not just partial regions.

We do not use latent variables obtained from F of PICNet and instead use only features. The functions F_μ and F_σ that estimate the symmetry intensity as latent variables are specific to our network and consist of three-layer ResBlock.

Comparison with Related Studies

We compare our method with PICNet and 360IC (Akimoto et al. 2019). PICNet is a state-of-the-art image completion method, the source code of which is publicly available. Because we made a comparison with the same network con-

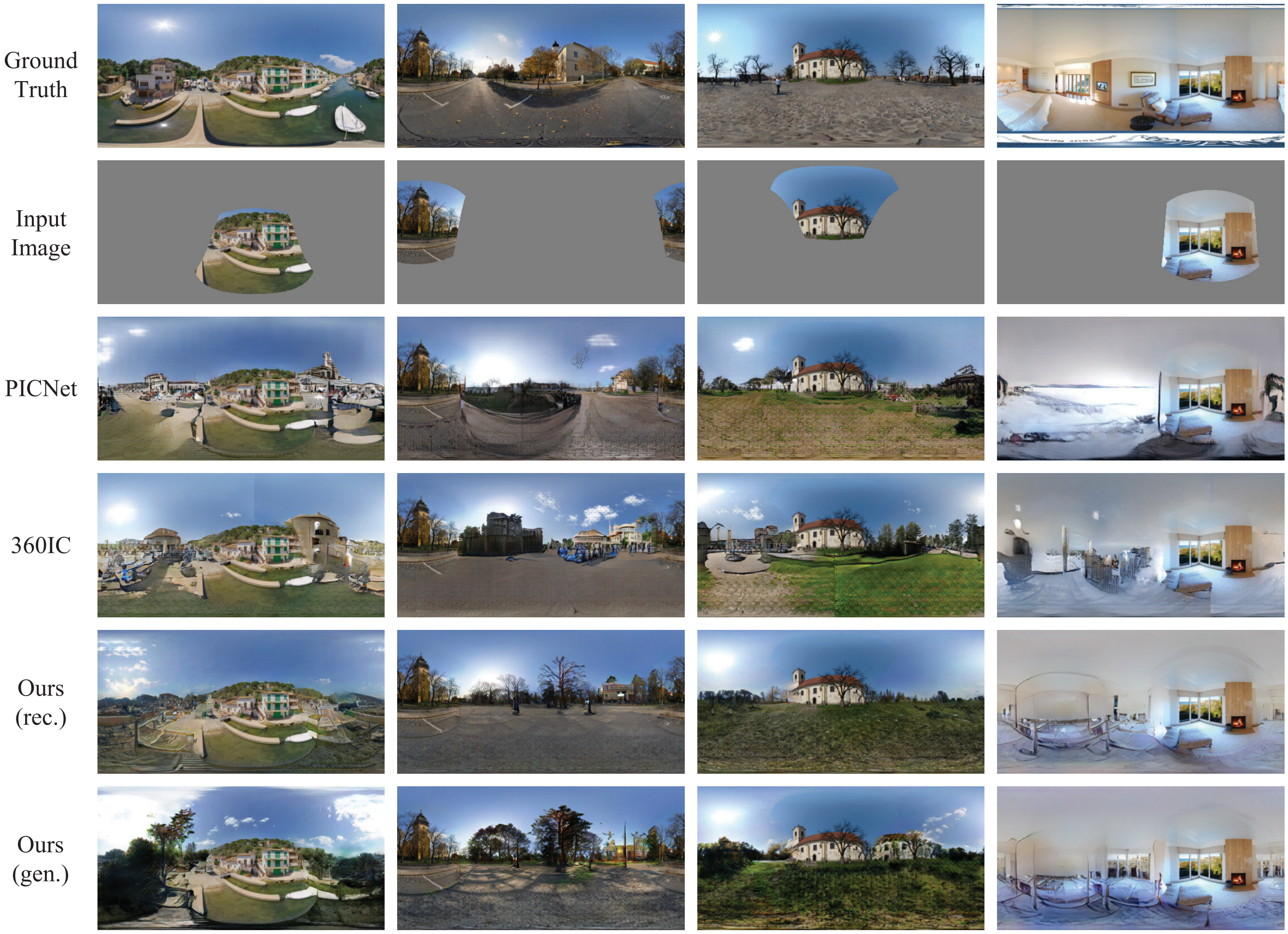

Figure 8: Qualitative comparison between spherical images generated using the proposed method and those generated using previous methods.

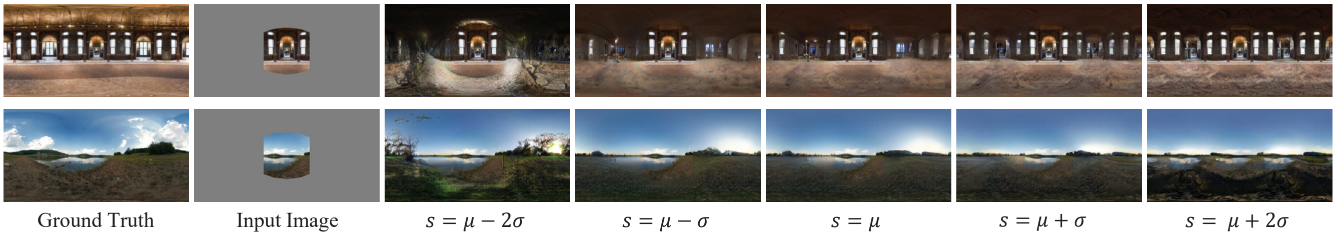

Figure 9: Generated images with the intensity of the symmetry controlled at specific values.

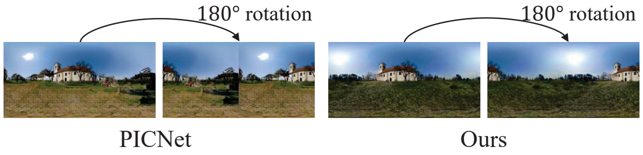

Figure 10: 180° rotated the generation results of PICNet and our approach for the same ground truth (the third image from the left in Figure 8). In PICNet, the left and right ends are discontinuous, whereas they are continuous in our method.

Method	FID (↓)	L1 (↓)	PSNR (↑)
PICNet	28.3	0.173	12.5
360IC	27.5	0.170	12.6
Ours (rec.)	28.1	**0.138**	**14.3**
Ours (gen.)	**24.4**	0.172	12.4

Table 1: Qualitative evaluation with each method.

figuration, we implemented 360IC by adding parallel dilated convolutions in ResBlocks and input the rearrangement steps into PICNet. In addition, we omitted the refined network from the original paper for better results of the preliminary experiment.

Qualitative Evaluation Figure 8 shows examples of spherical images generated from NFOV images by using each method. With our method, we show the generated images in the reconstruction and generation settings. Images generated with our method are finer than those of PICNet. In addition, 360IC incurs a discontinuity in the upper and lower regions from the input image through a rearrangement of the steps. As shown in Figure 10, PICNet also has a discontinuity on the left and right ends of the image. Our method can avoid such a discontinuity using the circular padding. In our generation setting, biased symmetry images are naturally generated (e.g., the leftmost image approaches asymmetry, and the third image reaches closer to symmetry, as shown in Figure 8).

Next, we show the generated images with the intensity of the symmetry controlled to specific values in Figure 9. We set such intensity from $\mu - 2\sigma$ to $\mu + 2\sigma$ in increments of σ, where $\mu = F_\mu \circ F(x), \sigma = F_\sigma \circ F(x)$. In these examples, our method can generate plausible spherical images controlled from asymmetric to symmetric, and can reconstruct the image close to the symmetry of the ground truth within the control ranges.

Quantitative Evaluation We evaluated each method by generating 10,000 images from partial images with randomized viewpoints and the FOV for the test dataset by using the Frechet inception distance (FID) score (Heusel et al. 2017), L1 error, and peak-to-signal-noise-ratio (PSNR). The FID measures the distance on the feature distribution, whereas L1 and PSNR measure the difference in the pixel values. With our method, we deal with two approaches, i.e., reconstruction and the generation settings. The evaluation results are presented in Table **??**. Our method with the generation setting displayed a superior FID score than those obtained by the other methods, meaning that our method can generate plausible images by capturing the distribution of high-order features of the spherical images. One possible interpretation of this result is that an image generated from a symmetric combination of input image features is more plausible as a whole scene than an image generated from estimated features in the unobserved region. When using the reconstruction setting, our method is superior in terms of L1 and the PSNR, which indicates that it allows us to reconstruct spherical images more accurately than conventional methods.

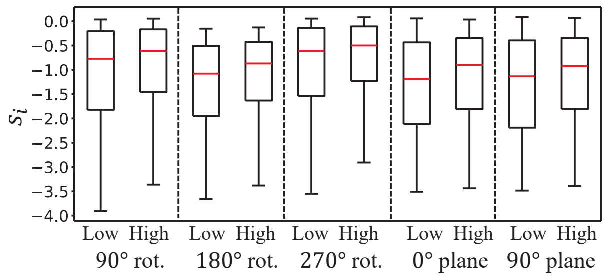

Figure 11: Median, quartiles, maximum, and minimum of the estimates of s_i on low and high SEM datasets for each symmetric transformation. Index i corresponds to the type of symmetric transformation for which the SEM is calculated.

Evaluation of Symmetric Control

We now quantitatively validate the symmetric control. Here, we use a normalized correlation between an original image and a symmetric-transformed image whose pixel values are normalized in $[-1, 1]$, and this is termed as the symmetry-evaluation metric (SEM). Note that the SEM is defined on each symmetric transformation T_i.

Initially, we evaluate the symmetry estimation. We divided the test images into 2,000 images with a higher SEM and 2,000 images with a lower SEM for each of the five symmetric transformations, which were used in previous experiments. In Figure 11, the authors show that the distribution of s_i estimated by our method using the reconstruction setting corresponds to each symmetric transformation. For all symmetry types, the estimates for the higher SEM images are higher than those of the lower SEM images (p-value $< 10^{-11}$ in Welch's t-test).

Next, we evaluate the reproducibility of the symmetry of the ground truth. In Figure 12, a scatter plot of the SEM is shown between the ground truth and images generated by our method using the reconstruction setting in the test dataset for each symmetric type. Note that we omit the results for a 270° rotation because it has the same values as that of a 90° rotation. There are some positive correlations whose correlation coefficients are 0.59, 0.54, 0.56, and 0.57, whereas these are 0.46, 0.38, 0.42, and 0.42 in PICNet, and 0.50, 0.42, 0.46, and 0.46 in 360IC. This proves that our method can generate spherical images to estimate and reconstruct the symmetry of the ground truth. Most of the images do not have an SEM near zero, and thus estimating and controlling the symmetry can reduce the reconstruction errors.

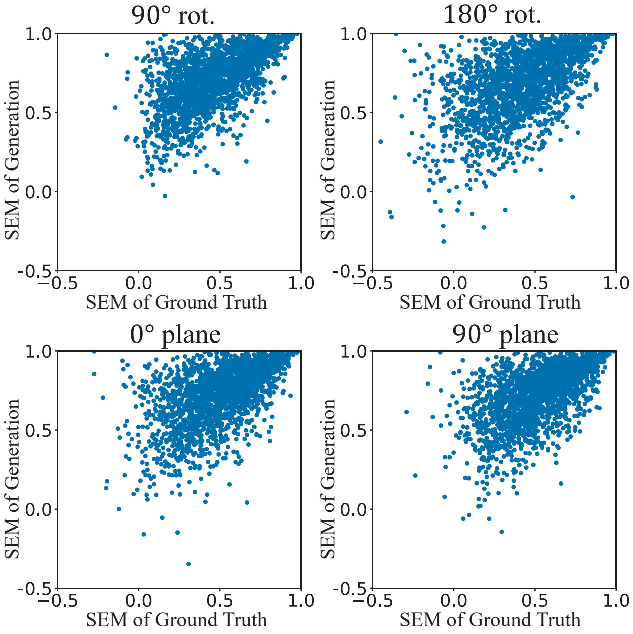

Figure 12: Scatter plot of the SEM between the ground truth and generated images obtained using our method with a reconstruction setting.

Method	FID (↓)	L1 (↓)	PSNR (↑)
Ours w/o GD (gen.)	27.6	0.171	12.4
Ours w/o GD (rec.)	46.8	0.135	14.6

Table 2: Evaluation results without global discriminator.

Ablation Study

We evaluate the quality of the generation without a global discriminator (GD) using the confidence of the entire image. The evaluation results are shown in Table **??**. This indicates that the GD reduces the FID, but does not improve in terms of the L1 or PSNR.

Limitations

Although the performance of the proposed method is promising, it has several limitations. First, when the FOV of input image is extremely small, it becomes difficult to reconstruct original images. Second, while our method controls the symmetry of the global structure, it cannot control local appearance such as placement of individual objects in the foreground. Third, in the case of using images of scenes that are not in the training data as an input, it is difficult to generate plausible images, as with conventional learning-based image completion methods. We hope that those limitations can be properly resolved in the further research.

Conclusion

We proposed a novel method for generating spherical images from a single NFOV image by estimating and controlling the scene symmetry. We incorporated the symmetry intensity into CVAEs as a latent variable, and the symmetry control was implemented as a circular shift of the hidden variables of the neural networks. Furthermore, our experimental results showed that the proposed method can generate various plausible spherical images controlled from symmetric to asymmetric, and we also reduced the reconstruction errors.

Acknowledgments

This work was partially supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, and partially supported by JSPS KAKENHI Grant Number JP19H01115. We would like to thank Antonio Tejero de Pablos, Atsuhiro Noguchi, Li Yang, Lin Gu, Sho Inayoshi, Takuhiro Kaneko, Wataru Kawai and Yusuke Kurose for helpful discussions.

References

Akimoto, N.; Kasai, S.; Hayashi, M.; and Aoki, Y. 2019. 360-degtee Image Completion by Two-stage Conditional-GANs. In *ICIP*.

Assens, M.; Giro-i Nieto, X.; McGuinness, K.; and O'Connor, N. E. 2018. Scanpath and saliency prediction on 360 degree images. In *Signal Processing: Image Communication*.

Ballester, C.; Bertalmio, M.; Caselles, V.; Sapiro, G.; and Verdera, J. 2001. Filling-in by joint interpolation of vector fields and gray levels. In *TIP 10(8)*, 1200–1211.

Bao, J.; Chen, D.; Wen, F.; Li, H.; and Hua, G. 2017. CVAE-GAN: Fine-grained image generation through asymmetric training. In *ICCV*.

Barnes, C.; Shechtman, E.; Finkelstein, A.; and Goldman, D. B. 2009. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ToG*.

Benjamin, A.; Condurache, P.; and Geiger, A. 2018. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *ECCV*.

Bertalmio, M.; Sapiro, G.; Caselles, V.; and Ballester, C. 2000. Image inpainting. In *SIGGRAPH*.

Cheng, H. T.; Chao, C. H.; Dong, J. D.; Wen, H. K.and Liu, T. L.; and Sun, M. 2018. Cube Padding for Weakly-Supervised Saliency Prediction in 360 Videos. In *CVPR*.

Cohen, T. S.; Geiger, M.; Koehler, J.; and Welling, M. 2018. Spherical CNNs. In *ICLR*.

Criminisi, A.; Perez, P.; and Toyama, K. 2003. Object removal by exemplar-based inpainting. In *CVPR*.

Esteves, C.; Allen-Blanchette, C.; Makadia, A.; and Daniilidis, K. 2018. Learning SO(3) Equivariant Representations with Spherical CNNs. In *ECCV*.

Fukushima, K.; and Miyake, S. 1982. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. In *Pattern Recognition 15(6)*, 455–469.

Gardner, M.-A.; Sunkavalli, K.; Yumer, E.; Shen, X.; Gambaretto, E.; Christian, G.; and Lalonde, J.-F. 2017. Learning to predict indoor illumination from a single image. In *TOG*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPs*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; and Nessler, B. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPs*.

Iizuka, S.; Simo-Serra, E.; and Ishikawa, H. 2017. Globally and Locally Consistent Image Completion. In *SIGGRAPH*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

Khasanova, R.; and Frossard, P. 2017. Graph-based classification of omnidirectional images. In *ICCV Workshops*.

Kimura, N.; and Rekimoto, J. 2018. ExtVision: Augmentation of Visual Experiences with Generation of Context Images for a Peripheral Vision Using Deep Neural Network. In *CHI*.

Kingma, D. P.; and Ba, J. L. 2014. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. In *arXiv preprint arXiv:1312.6114*.

Larsen, A. B. L.; Sonderby, S. K.; Larochelle, H.; and Winther, O. 2015. Autoencoding beyond pixels using a learned similarity metric. In *arXiv preprint arXiv:1512.09300*.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. In *Neural computation 1(4)*, 541–551.

Lee, Y. K.; Jeong, J.; Yun, J. S.; Cho, W. J.; and Kuk-Jin, Y. 2018. SpherePHD: Applying CNNs on a Spherical PolyHeDron Representation of 360°. In *arXiv preprint arXiv:1811.08196*.

Li, Y.; Liu, S.; Yang, J.; and Yang, M.-H. 2017. Generative Face Completion. In *CVPR*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Smolley, S. P. 2017. Least squares generative adversarial networks. In *ICCV*.

Perraudin, N.; Defferrard, M.; Kacprzak, T.; and Sgier, R. 2019. DeepSphere: Efficient spherical convolutional neural network with HEALPix sampling for cosmological applications. In *Astronomy and Computing 27*, 130–146.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Auto-encoding variational bayes. In *ICML*.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. In *NeurIPs*.

Song, S.; and Funkhouser, T. 2019. Neural illumination: Lighting prediction for indoor environmentsk. In *CVPR*.

Song, S.; Zeng, A.; Chang, A. X.; Savva, M.; Savarese, S.; and Funkhouser, T. 2018. Im2Pano3D: Extrapolating 360° Structure and Semantics Beyond the Field of View. In *CVPR*.

Srinivasan, P. P.; Mildenhall, B.; Tancik, M.; Barron, J. T.; Tucker, R.; and Snavely, N. 2020. Lighthouse: Predicting Lighting Volumes for Spatially-Coherent Illumination. In *CVPR*.

Su, Y.-C.; and Grauman, K. 2017. Learning spherical convolution for fast features from 360 imagery. In *NeurIPs*.

Sumantri, J. S.; and Park, I. K. 2020. 360 Panorama Synthesis from a Sparse Set of Images with Unknown FOV. In *WACV*.

Tateno, K.; Navab, N.; and Tombari, F. 2018. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV*.

Xiao, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2012. Recognizing Scene Viewpoint using Panoramic Place Representation. In *CVPR*.

Yang, Q.; Li, C.; Dai, W.; Zou, J.; Qi, G.-J.; and Xiong, H. 2020. Rotation Equivariant Graph Convolutional Network for Spherical Image Classification. In *CVPR*.

Zhang, Y.; Xiao, J.; Hays, J.; and Tan, P. 2013. FrameBreak: Dramatic Image Extrapolation by Guided Shift-Maps. In *CVPR*.

Zhao, L.; Mo, Q.; Lin, S.; Wang, Z.; Zuo, Z.; Chen, H.; Xing, W.; and Lu, D. 2020. UCTGAN: Diverse Image Inpainting based on Unsupervised Cross-Space Translation. In *CVPR*.

Zheng, C.; Cham, T.-J.; and Cai, J. 2019. Pluralistic Image Completion. In *CVPR*.