Diverse Plausible 360-Degree Image Outpainting for Efficient 3DCG Background Creation

Naofumi Akimoto Yuhi Matsuo Yoshimitsu Aoki
Keio University
{nakimoto,ymatsuo}@aoki-medialab.jp, aoki@elec.keio.ac.jp

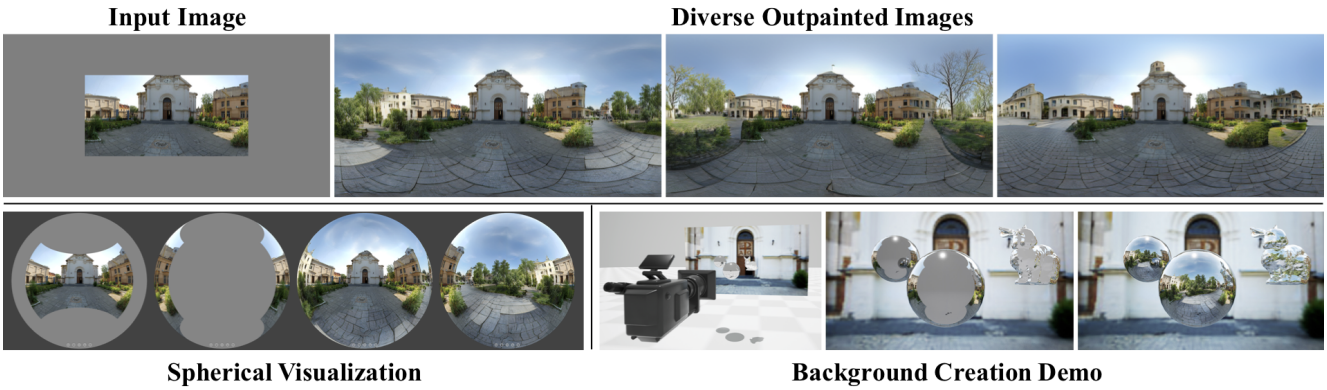


Figure 1. We generate the plausible environment from a narrow field of view image, using a transformer-based outpainting method that considers the nature of 360-degree images, to realize efficient 3DCG scene creation. See also demonstrations in the supplementary video.

Abstract

We address the problem of generating a 360-degree image from a single image with a narrow field of view by estimating its surroundings. Previous methods suffered from overfitting to the training resolution and deterministic generation. This paper proposes a completion method using a transformer for scene modeling and novel methods to improve the properties of a 360-degree image on the output image. Specifically, we use CompletionNets with a transformer to perform diverse completions and AdjustmentNet to match color, stitching, and resolution with an input image, enabling inference at any resolution. To improve the properties of a 360-degree image on an output image, we also propose WS-perceptual loss and circular inference. Thorough experiments show that our method outperforms state-of-the-art (SOTA) methods both qualitatively and quantitatively. For example, compared to SOTA methods, our method completes images 16 times larger in resolution and achieves 1.7 times lower Fréchet inception distance (FID). Furthermore, we propose a pipeline that uses the completion results for lighting and background of 3DCG scenes. Our plausible background completion enables perceptually natural results in the application of inserting virtual objects with specular surfaces.

1. Introduction

In recent three-dimensional computer graphics (3DCG) production, 360-degree images are helpful for efficiently creating lighting and backgrounds. For example, a designer might spend much time creating 3D objects in the near field and creating the background quickly by using 2D images with a narrow field of view (NFoV) images or 360-degree images. However, the production method of creating the background by placing 2D images behind a 3D object cannot fully represent the scenery reflected on the surface of the 3D object. Of course, this problem does not occur if the image surrounds the object in 360 degrees. However, 360-degree images, especially high dynamic range images (HDRI), are generally more expensive to prepare than NFoV images.

This paper addresses the problem of converting an NFoV image into a 360-degree image by complementing its surroundings to obtain a 360-degree environment consistent with the image given as a partial background (Fig. 1). By solving this problem, users can use only a NFoV image to reflect the surrounding environment to objects [1, 24], or in the case of HDRI, to achieve natural shadows and global illumination through Image-Based Lighting [6, 24].

For use by designers, it is desirable to infer NFoV images of any size and to have choices by generating diverse 360-degree images. However, existing methods are deter-

Figure 2. Limitations of prior methods. (a) CNN-based method [1] and (b) transformer-based translation method [4] suffer from overfitting to resolution during training (512×256). Furthermore, (b) has no connection between the ends.

ministic in their estimation and can only correctly infer the trained resolution. For example, as shown in Fig. 2, 360IC [1], trained at 512×256, produces many artifacts at 1024×512. We believe that this is due to distortions caused by the equirectangular projection (ERP). Based on the above, our goal is to realize the outpainting of 360-degree images with the following two properties. (1) Sample diverse outputs for a single input and (2) infer arbitrary resolutions.

The key idea of our approach is to introduce a transformer [29] into an outpainting method for diverse outputs. In previous works, TT [4] is an image-to-image translation method with a transformer that can generate various outputs through sampling from the learned distribution. However, as shown in Fig. 2(b), TT alone is not consistent enough with the input image, and it can only generate images of fixed size. Therefore, we cannot directly introduce TT into our task, and the resolution problem remains. To solve this technical problem of introducing a transformer, we propose an additional network as a second stage. Specifically, we present a framework comprising CompletionNets and AdjustmentNet. (1) CompletionNets is an image completion module that uses the same networks as TT and two novel techniques to be proposed later. (2) AdjustmentNet improves the consistency of color, stitching, and resolution between the output result of CompletionNets and the input image. Because AdjustmentNet adjusts the fixed-size output of CompletionNets to the size of the input image, we can obtain completion results for any image resolution.

Furthermore, because the above framework does not yet sufficiently consider the unique properties of 360-degree images, we propose two novel techniques for this purpose. First, to achieve continuity at both ends of an image, which is a property of 360-degree images, we propose circular inference as a new auto-regressive order for a transformer. It improves the connectivity at both ends of an image at the pixel and semantic levels by performing inference while circulating the image. Second, to further improve the perceptual quality, we propose a WS-perceptual loss function for training of CompletionNets. This loss function reflects that 360-degree images have different information content along the latitudinal direction and improves the performance of 360-degree image modeling by focusing on computing the loss in the information-rich regions.

Our thorough experiments show not only that the proposed method can perform diverse completions at arbitrary resolutions but also that the proposed method outperforms several state-of-the-art methods both qualitatively and quantitatively. For example, in terms of FID score, our method shows 1.7% lower improvement than 360IC and achieves plausible completion for images with 16 times as many pixels (1024×512) as EnvMapNet [24] (256×128).

Moreover, we propose a pipeline to create an HDR environment map from a single NFoV image and use it as lighting and background in 3DCG. Through demonstrations, we show that our method reaches the quality of 360-degree image completion, which can be used for 3DCG and is helpful for efficient background creation.

The proposed method produces a plausible 360-degree image and provides various completion results, allowing designers to choose their preferred result among them. Considering these characteristics, we conclude with a discussion of potential applications.

Our contributions can be summarized as follows:

- We propose AdjustmentNet to introduce a transformer into 360-degree image outpainting, which enables diverse and arbitrary-resolution outputs.
- We propose two novel techniques for acquiring the properties of 360-degree images: WS-perceptual loss for the training of CompletionNets and circular inference for the transformer. These allow us to outperform previous methods both quantitatively and qualitatively.
- We demonstrate that our high-resolution and plausible completion renders natural-looking scenes even when specular virtual objects are close to a camera or the camera views all around on 3DCG scenes.

2. Related Work

Image completion. Image inpainting is the task of filling in missing regions with appropriate pixels [2, 8]. Learning-based image inpainting [10, 13, 19], which is CNN-trained on large datasets, has been extensively studied recently. In addition, attention-based image inpainting has been proposed and shown promising results [16, 38]. However, most of the methods that train CNN with GAN produce deterministic outputs. In other words, these methods output only one result for an input image. PIC [41], in

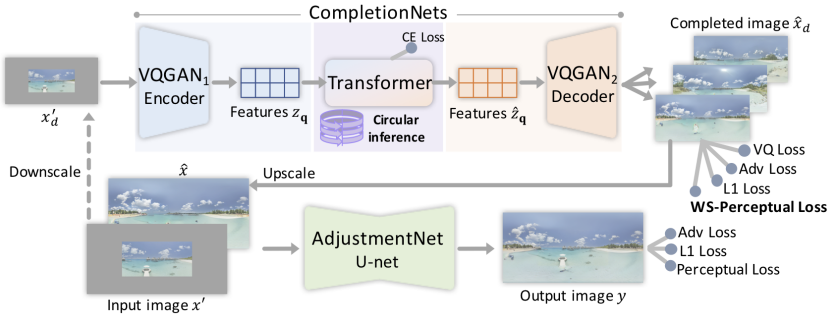


Figure 3. **Framework overview.** Our method comprises two modules: CompletionNets and AdjustmentNet. CompletionNets can sample various completion results from a fixed-size image input. AdjustmentNet improves the consistency of color, stitching, and resolution between the CompletionNets' output and the input image, resulting in completion for any image size.

contrast, outputs multiple results by employing CVAE [23].

Image outpainting is an extrapolation problem that generates the surroundings of an input image. This task can include image extension [27, 32], novel view synthesis [34], infinite landscape generation [15, 36], and panorama generation [14, 22]. A similar task to our work is the generation of panoramas that provide a 360-degree view. However, ERP images have significant distortions at the top and bottom of the images and, therefore, panoramic images cannot be used as ERP images.

Using a transformer for image completion is another recently studied topic [3, 4, 30] and has the following two main advantages. (1) Non-local attention can help generate a global structure and contextual consistency. (2) Sampling from the distribution learned by a transformer leads to more diverse completions than CVAE, as verified by [30]. In contrast, the disadvantages of a transformer are that it requires an enormous computational cost when dealing with large images. Moreover, transformer-based image-to-image translation (Im2Im) [4, 20] resamples the pixels in the conditional input region, resulting in a loss of consistency with the original pixels.

However, the above image completion works designed their approach for planar images. In other words, the generated results lacked the properties of 360-degree images, such as the connection between the two ends of the images and the latitudinal distortion caused by the projection. In contrast, our method can plausibly complete a 360-degree image by introducing a transformer into an image outpainting while considering the 360-degree properties.

360-Degree image outpainting. 360-Degree image outpainting is the task of completing the surroundings of a partial 360-degree image. Inverse rendering [21, 33] and lighting estimation [5, 6, 12, 25] perform the task of completing a 360-degree image as an intermediate task to represent lighting through the 360-degree image (environment map). However, these methods cannot predict high-frequency textures, and the image resolution is small. Similar to our work is the task focused on pixel completion [1, 7, 24]. These are image completion methods that consider the properties of 360-degree images. 360IC and SIG-SS [7] employ techniques to improve the continuity of both ends of a 360-degree image. EnvMapNet trains a network by weighting the pixel loss to account for the difference in latitudinal information density due to the projection. However, except for SIG-SS, which uses CVAE to sample the strength of symmetry, these methods are deterministic outputs. Furthermore, they suffer from overfitting for image resolution during training.

3. Method

We generate a 360-degree image by completing the surrounding area of an NFoV image. In this work, 360-degree images are ERP images. Again, our goal is to obtain multiple and diverse outputs for a single input image and enable inference at arbitrary resolutions different from the training resolution. Our approach is to perform diverse completions of a scene using a transformer. TT has already shown that diverse completion is possible with Im2Im using a transformer. However, as mentioned in Sec. 1 and Fig. 2, the transformer-based Im2Im is not suitable for 360-degree images in the terms of overfitting a training resolution and consistencies with the input pixels. Therefore, we propose a framework extended with AdjustmentNet to solve the problems (Sec. 3.1). Moreover, we propose a new loss, WS-perceptual loss (Sec. 3.2), and a new inference method for the transformer, circular inference (Sec. 3.3), to reflect the properties of 360-degree images to outputs.

Overview. Fig. 3 shows an overview of our proposed framework. The input of the entire framework is an incomplete image $x' \in \mathbb{R}^{H\times W\times 3}$. During training, we crop some regions from the ERP images $x \in \mathbb{R}^{H\times W\times 3}$ and fill the remaining regions with gray values. The output $y \in \mathbb{R}^{H\times W\times 3}$ is a restored image of the complete 360-degree scene.

Our method comprises two modules: CompletionNets and AdjustmentNet. First, we downsize the incomplete input image x' to a fixed size and use it as input to CompletionNets. CompletionNets completes the incomplete image $x'_d \in \mathbb{R}^{h\times w\times 3}$ using a transformer. Because the completed image $\hat{x}_d \in \mathbb{R}^{h\times w\times 3}$ is of fixed size, we restore it by upscaling to the original size of the input image. Next, to enable inference at arbitrary resolutions different from the training resolution, AdjustmentNet uses the completed image $\hat{x} \in \mathbb{R}^{H\times W\times 3}$ and the input image x' to estimate the high-frequency texture of the completed image y according

to the input image x'. It also performs stitching and color correction to obtain the final output.

3.1. Model Architecture

CompletionNets. The primary network structure of CompletionNets is the same as TT: two VQGANs [4] and a transformer. The approach of TT is vector-quantized image modeling, which models a sequence of quantized image tokens. VQGAN is a network that uses a feature quantization mechanism [28] at the bottleneck of the encoder-decoder CNN to obtain the image tokens.

In our CompletionNets, VQGAN_1 encodes the incomplete image, and VQGAN_2 decodes the features complemented by the transformer. Unlike TT, CompletionNets treats images of fixed size as input and output, considering that VQGAN also overfits the training resolution probably due to the inherent distortion of ERP representation, and uses WS-perceptual loss for training and circular inference for inference. Our transformer models the scene of a 360-degree image as a sequence of quantized features and performs diverse image completions by sampling from the learned distribution.

AdjustmentNet. To achieve completion at arbitrary image sizes, we propose AdjustmentNet, a network that improves the consistency between the output of CompletionNets and the input region. In high-resolution image completion methods employing two stages [30, 37], the primary role of the second stage is to refine outputs by adding high-frequency components. However, as shown in Fig. 4, applying only super-resolution is not sufficient for our method. Fig. 4(a) shows a composite image of the upscaled completion image and the input region. Fig. 4(b) shows that even with the SOTA method of super-resolution [31], refinement alone is not sufficient. One of the causes for this is that the transformer resamples not only the completion region but also the input region. As a result, the completed region is predicted to fit the resampled input region and does not match the original input image. In contrast, we adjust the output of CompletionNets to match the input image in terms of color, stitching, and resolution, as shown in (c). The network is a U-net structure implemented in the same CNN structure as VQGAN without the VQ mechanism [28].

3.2. Training

WS-perceptual loss. VQGAN is a network for obtaining quantized vectors of image features, which models local regions of an image using CNN. TT proposes a self-supervised manner using adversarial loss $\mathcal{L}_{\text{GAN}}$, L1 loss $\mathcal{L}_1$, perceptual loss $\mathcal{L}_{\text{Perc}}$, and VQ loss $\mathcal{L}_{\text{VQ}}$. In contrast, we propose a novel loss function, WS-perceptual loss, to suitably model local regions for ERP representation. This loss function reflects the nature of ERP representation that there is a difference in the amount of information in each region along the latitudinal direction. Previous methods [6, 24] weighed pixel-level differential losses, such as L1 loss, to account for their projection onto a sphere. However, high-level features, such as semantics, should also be modeled around the central region. Therefore, WS-perceptual loss is an extension of perceptual loss (LPIPS) [40] to the loss on the unit sphere as follows: Similar to WS-PSNR [26], we prepare the following weights to account for the projection onto a sphere.

$$w'_l(u,v) = \cos((v - H_l/2 + 1/2) \cdot \pi / H_l), \tag{1}$$

where u and v are the positions on the feature (size $H_l \times W_l$) in the lth layer of the feature extractor. We use Eq. 1 to weigh the perceptual loss $\mathcal{L}_{\text{Perc}} = \sum_l \frac{1}{H_l W_l} \sum_{u,v} \|w_l \odot (y^l_{uv} - x^l_{uv})\|_2^2$ at each resolution.

$$\mathcal{L}_{\text{WS-Perc}} = \sum_l \frac{1}{\sum_{u,v} w'_l} \sum_{u,v} w'_l \odot \|w_l \odot (y^l_{uv} - x^l_{uv})\|_2^2. \tag{2}$$

VQGAN. We train both VQGAN_1 and VQGAN_2, which have both encoder and decoder, with

$$\mathcal{L}_{\text{VQGAN}} = \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_1\mathcal{L}_1 + \lambda_{\text{VQ}}\mathcal{L}_{\text{VQ}} + \lambda_{\text{WS-Perc}}\mathcal{L}_{\text{WS-Perc}}. \tag{3}$$

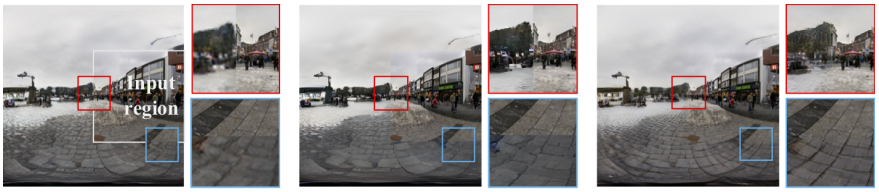
VQGAN_1 learns to reconstruct 360-degree images with missing regions to obtain quantized features $z_{\mathbf{q}} \in \mathbb{R}^{h_{\mathbf{q}} \times w_{\mathbf{q}} \times n_z}$ of the incomplete input image. VQGAN_2, in contrast, learns to reconstruct a complete 360-degree image to obtain a decoder that obtains a complete 360-degree image from quantized features $\hat{z}_{\mathbf{q}} \in \mathbb{R}^{h_{\mathbf{q}} \times w_{\mathbf{q}} \times n_z}$.

Transformer. We train the transformer to model a 360-degree scene and to perform completion. Using the transformation from $z_{\mathbf{q}}$ to $\hat{z}_{\mathbf{q}}$ as supervision, the model learns to predict the distribution of the next index after indices $s_{<i}$, conditional on indices c, by using the following equation:

$$\mathcal{L}_{\text{Transformer}} = \mathbb{E}_{x \sim p(x)} \left[-\log p(s|c) \right], \tag{4}$$

where $p(s|c) = \prod_i p(s_i|s_{<i}, c)$, and the transformer does not directly deal with the quantized features ($z_{\mathbf{q}}$ and $\hat{z}_{\mathbf{q}}$) but treats the sequence of indices (c and s) assigned to them.

AdjustmentNet. AdjustmentNet has a simple network architecture but explicitly learns to match the output image of CompletionNets with its input. Therefore, we train AdjustmentNet by restoring the preprocessed input images to the original images (GT images) in a self-supervised manner. In addition, to avoid over-fitting to the top and bottom distortions of the ERP image, we randomly crop only a part from the ERP image and use it as the GT image. The following steps describe the preprocessing: (1) To learn the adjustment of the boundary connection, reconstruct the GT image using the learned VQGAN_2, obtaining a reconstructed image with the difference from the GT image. (2) To learn color adjustment, add color jitter to the input image before reconstructing it with VQGAN_2, which results in a reconstructed image with slightly different colors from the GT image. (3) To learn to adjust the resolution, scale down the



(a) CompletionNets + Bicubic (b) CompletionNets + Real-ESRGAN (c) CompletionNets + AdjustmentNet

Figure 4. Effect of AdjustmentNet. (a) and (b) show that the output of CompletionNets needs to be adjusted not only in resolution but also in color and stitching with the input region.

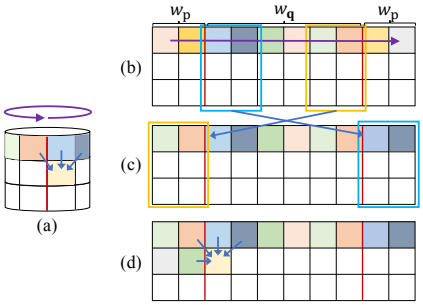


Figure 5. Circular inference.

GT image to be reconstructed by VQGAN_2 in advance. After reconstruction, the image is returned to the original scale using the bicubic method. By composing this reconstructed image with a smaller region of the GT image than this image and using it as the input image for AdjustmentNet, we can learn to adjust color, stitching, and resolution while using the GT region as a hint. Therefore, we use the vanilla perceptual loss instead of the WS-perceptual loss. We use

$$\mathcal{L}_{\text{Adjust}} = \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_1\mathcal{L}_1 + \lambda_{\text{Perc}}\mathcal{L}_{\text{Perc}} \tag{5}$$

for the learning.

3.3. Inference

Circular inference. To obtain completion results that reflect the continuity of 360-degree images, we propose circular inference as an inference order for the transformer. TT proposes a raster ordering, called sliding attention window, as the autoregressive order of the transformer estimation. However, as shown in Fig. 2(b), this method discontinues at both ends of the 360-degree image. SIG-SS proposes circular padding to connect the two ends. However, this padding is for convolution and cannot be applied to estimate the global semantics of 360-degree images using a transformer. Therefore, we propose circular inference so that the continuity of both ends can be accounted for in the estimation stage of the transformer, as shown in Fig. 5(a) and Fig. 5(d). The main idea is to circularly estimate some regions twice by the transformer to generate overlaps. For implementation, we duplicate both ends (with length w_{p}) of the quantized feature map $z_{\mathbf{q}} \in \mathbb{R}^{h_{\mathbf{q}} \times w_{\mathbf{q}} \times n_z}$ on the opposite side in advance (from $h_{\mathbf{q}} \times w_{\mathbf{q}}$ to $h_{\mathbf{q}} \times (w_{\mathbf{q}} + 2w_{\text{p}})$), and the transformer performs an estimation in raster order (Fig. 5(b)). After estimating a row, the estimation results of both ends of the quantized feature map $\hat{z}_{\mathbf{q}} \in \mathbb{R}^{h_{\mathbf{q}} \times w_{\mathbf{q}} \times n_z}$ are copied from the opposite side and replaced (Fig. 5(c)). That is, the length w_{p} from the left is replaced by the estimated result of w_{p} from the right of the original length $w_{\mathbf{q}}$. The same applies to the other side.

As described above, circularly estimating with a transformer allows connecting the two ends in higher-order features. When decoded into an image, the connection at the semantics level is improved, allowing for a more plausible completion as a 360-degree image.

Thus, the overall inference is

$$y = G_{\text{Adjust}}(G_{\text{VQGAN}_2}(T(E_{\text{VQGAN}_1}(x'))), x'), \tag{6}$$

where G_{Adjust}, G_{VQGAN_2}, T, and E_{VQGAN_1} indicate AdjustmentNet, the decoder of VQGAN_2, the transformer, and the encoder of VQGAN_1, respectively.

4. Experiments

We compare our method qualitatively and quantitatively with previous methods to verify the effectiveness of the proposed components. In our supplementary video, we also compare the results in all-around viewing and object insertion applications.

4.1. Experiment Settings

Implementation details. We use Adam [11] as an optimizer whose learning rate = 4.5e-06. $\lambda_1 = 1.0$, $\lambda_{\text{Perc}} = 1.0$, $\lambda_{\text{VQ}} = 1.0$, $\lambda_{\text{WS-Perc}} = 1.0$, and λ_{GAN} is an adaptive weight [4]. We train the transformer with 20 epochs and, for the remaining networks, with 30 epochs. Our method produces 1024×512 images.

Datasets. We use SUN360 [35] and Laval Indoor Dataset [6] as the datasets. We divide SUN360 into 47938 training images and 5000 test images of the "Outdoor" class. In contrast, we divide the Laval Indoor Dataset into 1837 training images and 289 test images, following the provided training and test split. The number of images in the Laval Indoor Dataset is too tiny to train our model, so we fine-tune the model trained by SUN360 on this dataset, just as EnvMapNet uses an additional dataset. For data augmentation, we randomly change the view direction horizontally, as in [1, 7].

Evaluation. As a quantitative metric, we use Fréchet inception distance (FID) [9] to evaluate the quality and diversity of the generated images. We explain the details of the calculation method in Sec. 4.4. When comparing the proposed method with other methods, we set the sequence length of the transformer to 512. When validating each component of the proposed method, we set the sequence length of the transformer to 256 to make network training more efficient.

Baselines. To compare our model with 360IC, we train 360IC on our dataset. The original 360IC introduces a two-

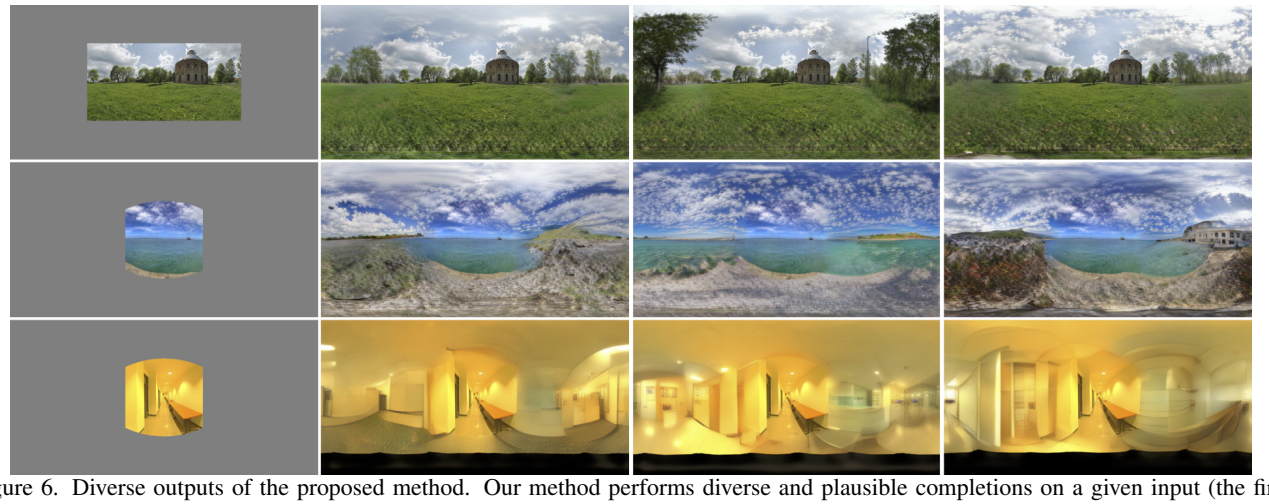


Figure 6. Diverse outputs of the proposed method. Our method performs diverse and plausible completions on a given input (the first column).

stage approach to avoid overtraining on a small training dataset of 600 images. We implement 360IC with a single-stage as in [7], and then we train the model with a sufficient amount of training data. To compare the differences in scene modeling between a transformer and CNN, we use a CNN structure (Fig. 2(a)) similar to that of VQGAN and the same training procedure as ours. Furthermore, we implement the bottleneck by adding their proposed four parallel dilated conv blocks instead of the quantization mechanism.

To compare with SIG-SS, we infer the authors' trained models with our test images. The authors trained their model on SUN360. Note, however, that their train/test splitting is unknown, and they may have trained their model on the test images we prepared.

EnvMapNet does not publish their code, only their evaluation protocol and scripts. We follow their instructions and run their evaluation script on the same train/test split of the Laval Indoor Dataset. For comparison, we quote the resulting images and scores from their paper. Note that the location of the input region and the tone mapping method do not match their experiment and ours.

4.2. Diverse Outputs

Fig. 6 shows that our approach can output multiple and diverse completions. The left column is an input image, and all results are 1024×512 images. The top two rows show the results generated with the same model trained with SUN360 and different input regions. In the top row, the input regions are 180 degrees in the longitudinal direction and 90 degrees in the latitudinal direction. The middle row input regions correspond to the 90-degree angle of view when converted to a perspective image. We can find that the larger the input region, the better the quality of the generated texture. The last row shows the results of an experiment conducted using the Laval Indoor Dataset and that our method can estimate indoor scenes with various structures.

4.3. Qualitative Comparisons

Fig. 7 compares 360IC and the proposed method. 360IC captures the distortions of ERP representation, but there are artifacts in the textures. In contrast, the proposed method can generate the textures and shapes of each object more accurately. Comparing "360IC" and "CompNets Only," we can see the difference in scene modeling between convolution and transformer. Dilated convolution increases the receptive field of the CNN, but this causes the transferred information to be sparse, which may be the cause of artifacts in the texture generation. In contrast, the transformer can generate globally consistent textures and represent distortions in the upper and lower regions of the image, where distortions specific to ERP images occur significantly.

Fig. 8 compares SIG-SS and ours. The input region is the same as in Fig. 6(b); SIG-SS has the results of reconstruction (rec) and sampling (gen). The resolution is 512×256. The reconstruction results show overfitting, where similar objects appear, and the sampling results lack global consistency. In contrast, the proposed method provides results that match the context of the input region.

Fig. 9 shows a comparison between ours and EnvMapNet, where EnvMapNet is a 256×128 completion image, while ours is a 1024×512 completion image. In other words, our method can complete 16 times as many pixels as EnvMapNet, and our results are also better looking. The preprocessing of EnvMapNet, which requires clustering of datasets to stabilize adversarial learning, is not necessary for our method. The results trained on the Laval Indoor Dataset do not generate detailed textures compared to those trained on SUN360. This is because we train the model on a small dataset, which is one of our limitations.

4.4. Quantitative Comparisons

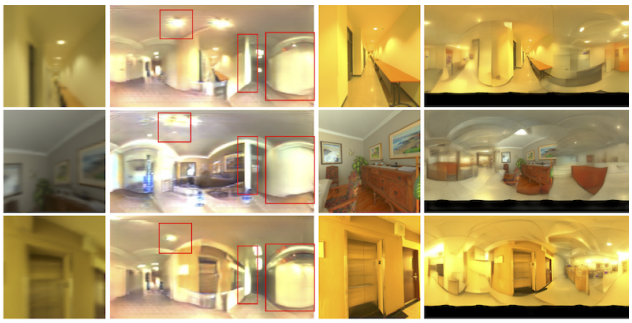
We use FID to evaluate the quality and diversity of the completions of the proposed method. Table 1 uses SUN360

(a) 360IC (b) CompletionNets only (c) Ours

Figure 7. Qualitative comparison with 360IC. The input region is the same as that in the first row of Fig. 6.

(a) SIG-SS (rec) (b) SIG-SS (gen) (c) Ours

Figure 8. Qualitative comparison with SIG-SS. The input region is the same as that in the second row of Fig. 6.



(a) Input Crop (b) EnvMapNet (c) Input Crop (d) Ours

Figure 9. Qualitative comparison with EnvMapNet. The input area is the same as that in the third row of Fig. 6. Note that the inputs of EnvMapNet(a) and ours(c) are not exactly the same.

and Clean-FID [18] as the script to compute the FID. The results show that our method outperforms 360IC. Using the same evaluation method, in Table 2, we compare our method with SIG-SS, showing that our method, which uses a transformer, outperforms their method, which uses CVAE. Table 3 compares our method with EnvMapNet and that of Gardner et al. [6] on the Laval Indoor Dataset. To compute the FID, we follow their evaluation protocol: we convert an image into a Cubemap and remove the top and bottom planes containing little information. In summary, the FID comparisons show that our method is superior in terms of the generated results' quality and diversity.

(a) Raster order (b) Circular padding (c) Circular inference

Figure 10. Circular inference connects both ends of a 360° image, both at the pixel and semantic levels. See Sec. 4.5 for details.

4.5. Analysis

Verification of circular inference. Fig. 10 and Table 4 show that our circular inference can perform consistent estimation as ERP images. In Fig. 10, we show a partially cropped region (256×128) of the output of CompletionNet (512×256), with the image ends aligned such that the completion region is in the image center. We use the same trained model for each method and change only some of the inferences. As shown in Fig. 10(a), the left and right edges are not connected in the raster order, which is the auto-regression of a transformer used in TT. One way to connect the two edges is to use circular padding [7]. Circular padding is a technique that pads the pixels on opposite edges of the ERP image during convolution. We use it to decode the features estimated by our transformer into images. This technique helps to improve the continuity at the pixel level, but at the semantic level, the contents at both ends are different. For example, in Fig. 10(b), the grass and dirt ground are separated in the center. In contrast, circular inference helps to improve the continuity at the semantic level during transformer estimation. As Table 4 shows, the FID scores of circular padding and circular inference are similar, but qualitatively, circular inference contributes to the generation of images that better reflect ERP image properties.

Verification of WS-perceptual loss. The loss considers the difference in pixel level and the difference in high-level features in the information amount along the latitudinal direction of the ERP image. In Table 5 and Table 6, to evaluate the effect more directly, we do not use AdjustmentNet but the output of CompletionNets. To compute the FID only in the completed region and not in the input region, we use only the generated region (256×128) in the image center corresponding to 90 degrees of latitude and 180 degrees of longitude out of the output image (512×256). Table 5 compares the results of WS-perceptual loss with perceptual loss or with WS-L1 loss, which considers only low-level differences, on the proposed network. Table 6 shows that the use of WS-perceptual loss contributes to improving the FID score in training the 360IC network. In summary, WS-perceptual loss contributes to generating images with more ERP image properties by considering the sphere and weighing high-level features.

	360IC [1]	CompletionNets Only	Ours
FID↓	16.44	14.96	**9.52**

Table 1. FID on SUN360 w/ $180^\circ \times 90^\circ$ input.

	SIG-SS(rec) [7]	SIG-SS(gen) [7]	Ours
FID↓	31.91	26.81	**23.13**

Table 2. FID on SUN360 w/ 90° input.

	Gardner *et al.* [6]	EnvMapNet [24]	Ours
FID↓	197.4	52.7	**46.15**

Table 3. FID on Laval Indoor dataset.

	Raster order	Circular padding	Circular inference
FID↓	30.03	**26.33**	26.96

Table 4. Evaluation of circular inference on SUN360 w/ $180^\circ \times 90^\circ$ input.

	Perceptual loss	WS-L1 loss	WS-perceptual loss
FID↓	29.00	35.00	**26.96**

Table 5. Evaluation of WS-perceptual loss on the proposed network on SUN360 w/ $180^\circ \times 90^\circ$ input.

	Perceptual loss	WS-perceptual loss
FID↓	67.47	**50.87**

Table 6. Evaluation of WS-perceptual loss on the 360IC network on SUN360 w/ $180^\circ \times 90^\circ$ input.

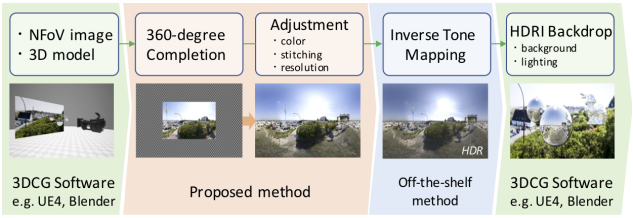


Figure 11. Pipeline for using the completion results for lighting and background of 3DCG scenes.

5. Application

We implement an application that uses a completed 360-degree image as a background and for lighting. We propose a pipeline to realize image-based lighting on the 3DCG software Unreal Engine 4 (UE4), as shown in Fig. 11. To use an HDRI Backdrop plugin, we need to convert images to HDR images beforehand; we convert low dynamic range (LDR) to high dynamic range (HDR) by inverse tone mapping using existing methods [17]. By introducing this process, our method can generate HDR environment maps from an LDR NFoV image.

Using this pipeline, we demonstrate the background creation and lighting for specular objects placed in a 3DCG scene in the bottom row of Fig. 1 and the supplementary video. Because of our proposed method's high resolution and plausible 360-degree image completion, we can demonstrate moving a camera to look all around. To the best of our knowledge, no other work has achieved such demonstration. When an NFoV image is composited behind 3DCG objects, the specular surfaces cannot be represented. Previous works [24, 33], which have often been limited to indoor environments, have generated environment maps to represent specular reflections; however, they have much lower resolution than our method. Therefore, there is no other method to insert specular objects as close to the camera as our results.

6. Discussion

This paper addresses the problem of completing 360-degree images from a narrow field of view. We first reveal that the previous methods had the limitations of overfitting the training resolution and having deterministic outputs. Next, to propose a framework to solve these problems, we introduce a transformer-based diverse Im2Im; however, the resolution problem remains. Thus, we propose AdjustmentNet. Moreover, we propose two novel techniques for obtaining completion results with improved properties of 360-degree images. Finally, our demonstrations show that, unlike others, the proposed method can provide designers with a new workflow for 3DCG creation by efficiently delivering an all-surrounding background.

6.1. Limitation

Inference time and computational memory. Our method takes approximately 30 s for one completion on a single 2080Ti, mainly because of a transformer. However, since 3DCG designers, the subject of our work, are likely to use PCs and servers with high-end GPUs, the performance of our method may be sufficient for practical use.

Controllability. The proposed method does not control what is generated in the completion region. One possible solution is to paste an object that appears directly in the completion region and complete it so that it is smoothly connected.

6.2. Potential Impact

This method is helpful for photo-based background representations in **virtual production**. For example, to project a background on an LED wall, designers occasionally create a background by compositing images obtained from stock photo sites instead of using a 3D environment composed of 3D models. However, it is possible to estimate the situations in areas the images do not photograph and express reflections on an inserted object.

Due to the limited number of photos available on **HDRI's stock photo** site, designers often lack originality as they use the same images. Even a popular website [39] provides only approximately 490 images, which is very few compared to our 5000 test results. Our method solves this problem by generating various new 360-degree HDRIs.

By streaming only the area the user is looking at out of the 360 degrees, we can reduce the communication capacity of video streaming in virtual reality spaces, such as **the Metaverse**. However, when the end-user's avatar holds a mirror in its hand, it may not reflect anything. In this case, our method can render the estimated scene.

Negative impact. Our completion results have room for improvement compared to real 360-degree images. However, if the generation results are improved, it could be a kind of Deepfake. Realistic generation of non-existent scenes or compositing of fake objects can mislead people.

References

[1] Naofumi Akimoto, Seito Kasai, Masaki Hayashi, and Yoshimitsu Aoki. 360-degree image completion by two-stage conditional gans. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4704–4708. IEEE, 2019. 1, 2, 3, 5, 8

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2

[3] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 3

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021. 2, 3, 4, 5

[5] Marc-Andre Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagne, and Jean-Francois Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[6] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2017. 1, 3, 4, 5, 7, 8

[7] Takayuki Hara, Yusuke Mukuta, and Tatsuya Harada. Spherical image generation from a single image by considering scene symmetry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 3, 5, 6, 7, 8

[8] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH 2007)*, 26(3), 2007. 2

[9] 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. *Advances in neural information processing systems*, 30, 2017. 5

[10] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, 2017. 2

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[12] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5918–5928, 2019. 3

[13] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3911–3919, 2017. 2

[14] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Cocogan: Generation by parts via conditional coordinating. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4512–4521, 2019. 3

[15] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 3

[16] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4170–4179, 2019. 2

[17] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 8

[18] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021. 7

[19] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[20] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors, 2021. 3

[21] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[22] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14144–14153, October 2021. 3

[23] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015. 3

[24] Gowri Somanath and Daniel Kurz. Hdr environment map estimation for real-time augmented reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11298–11306, 2021. 1, 2, 3, 4, 8

[25] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. *Proceedings of 33th IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[26] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE signal processing letters*, 24(9):1408–1412, 2017. 4

[27] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10521–10530, 2019. 3

[28] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, 2017. 4

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[30] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4692–4701, October 2021. 3, 4

[31] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021. 4

[32] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019. 3

[33] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 3, 8

[34] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 3

[35] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702, 2012. 5

[36] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10561–10570, 2019. 3

[37] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020. 4

[38] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2

[39] Greg Zaal. *Poly Haven*, 2021. 8

[40] 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. 4

[41] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 2