Autoregressive Omni-Aware Outpainting for Open-Vocabulary 360-Degree Image Generation

Zhuqiang Lu[1], Kun Hu[1,*], Chaoyue Wang[2], Lei Bai[3], Zhiyong Wang[1]
[1]The University of Sydney
[2]JD.com
[3]Shanghai AI Laboratory
zhuqiang.lu@sydney.edu.au, kun.hu@sydney.edu.au, chaoyue.wang@outlook.com, baisanshi@gmail.com, zhiyong.wang@sydney.edu.au

Abstract

A 360-degree (omni-directional) image provides an all-encompassing spherical view of a scene. Recently, there has been an increasing interest in synthesising 360-degree images from conventional narrow field of view (NFoV) images captured by digital cameras and smartphones, for providing immersive experiences in various scenarios such as virtual reality. Yet, existing methods typically fall short in synthesizing intricate visual details or ensure the generated images align consistently with user-provided prompts. In this study, autoregressive omni-aware generative network (AOG-Net) is proposed for 360-degree image generation by outpainting an incomplete 360-degree image progressively with NFoV and text guidances joinly or individually. This autoregressive scheme not only allows for deriving finer-grained and text-consistent patterns by dynamically generating and adjusting the process but also offers users greater flexibility to edit their conditions throughout the generation process. A global-local conditioning mechanism is devised to comprehensively formulate the outpainting guidance in each autoregressive step. Text guidances, omni-visual cues, NFoV inputs and omni-geometry are encoded and further formulated with cross-attention based transformers into a global stream and a local stream into a conditioned generative backbone model. As AOG-Net is compatible to leverage large-scale models for the conditional encoder and the generative prior, it enables the generation to use extensive open-vocabulary text guidances. Comprehensive experiments on two commonly used 360-degree image datasets for both indoor and outdoor settings demonstrate the state-of-the-art performance of our proposed method. Our code is available at https://github.com/zhuqiangLu/AOG-NET-360.

Introduction

A 360-degree (omni-directional) image offers a comprehensive spherical view of a scene and provides users the freedom to explore any direction from a singular view point. They have revolutionized the way that users consume, interact with, and produce visual content. Yet, the exclusive reliance on specialized cameras to capture these images poses significant challenges for their widespread adoption, limiting the scalability and accessibility of creating immersive content for broader audiences. In contrast, given the vast quantity of Narrow Field of View (NFoV) images captured daily via mobile phones and digital cameras, there has been a growing interest in transforming these conventional images into 360-degree panoramic visuals. By converting these NFoV images, extensive visual databases can be leveraged and enable more immersive experiences for the applications in Virtual Reality and Augmented Reality across various domains such as tourism, entertainment and education.

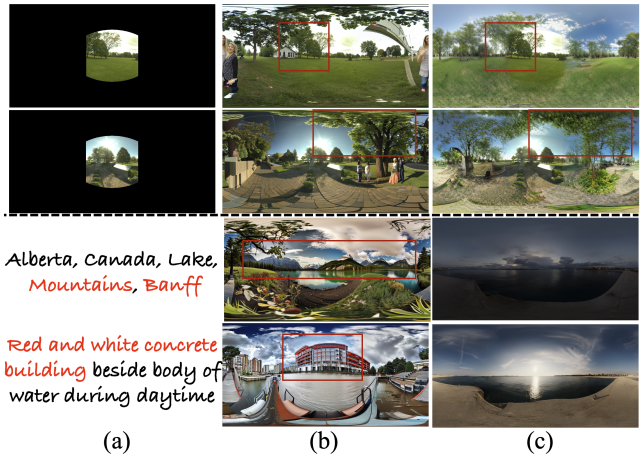

Figure 1: Examples of 360-degree image generation, showcasing the limitation of existing methods compared to ours. The top part above the dashed line depicts an NFoV-guided example and the bottom part below the dashed line is for a text-guided example. (a) Input condition. (b) Ours (AOG-Net). (c) Top - OmniDreamer (Akimoto, Matsuo, and Aoki 2022) and Bottom - Text2Light(Chen, Wang, and Liu 2022).

In recent years, deep learning methods have been explored to generate photo-realistic 360-degree images. For instance, OmniDreamer (Akimoto, Matsuo, and Aoki 2022) formulates a 360-degree image generation pipeline with a VQ-GAN (Esser, Rombach, and Ommer 2021) by treating NFoV images as incomplete 360-degree images. Conditioned on text guidances, Text2Light (Chen, Wang, and Liu 2022) introduces two VQGANs for a global-to-local modelling strategy in pursuit of generating high-resolution 360-degree images. ImmerseGAN (Dastjerdi et al. 2022) applies domain

*Corresponding author.

adaptation methods on pretrained GANs, which can be conditioned on both NFoV images and text guidances. While these methods show encouraging performance, the challenge remains regarding the usage of given NFoV images and user-provided open vocabulary text guidances individually or jointly for enhanced control in 360-degree image generation. Specifically, the existing methods typically fall short in synthesizing intricate visual details as shown in the top part of Fig. 1, where the details are vague or missing with OmniDreamer compared to our approach, which are highlighted in the red bounding boxes. Moreover, the generated images and user-provided text guidances tend to be inconsistent, especially under an open-vocabulary setting, as depicted in the bottom part of Fig. 1 by comparing Text2Light and our solution.

In this study, we propose a novel autoregressive omni-aware generative network (AOG-Net) for generating 360-degree images conditioned on open vocabulary text guidances and given NFoV images jointly or individually. Overall, the generation is formulated as an autoregressive stochastic process to outpaint an incomplete 360-degree image progressively, in which each step outpaints a local region under its corresponding NFoV view. This autoregressive scheme not only allows for deriving finer-grained and prompt-consistent patterns by dynamically observing and adjusting the generation process but also offers users greater flexibility to modify or introduce new conditions throughout the generation process. Furthermore, a global-local conditioning mechanism is devised to comprehensively formulate the outpainting guidance for each autoregression step. Text prompts, omni-visual cues, NFoV inputs and omni-geometry are encoded and further formulated with cross-attention based transformers into a global stream and a local stream for a conditioned generative backbone model. This study further explores the potential to leverage large-scale models for the conditional encoder and the generative prior, which helps complete the generation using open-vocabulary prompts. Comprehensive experiments on two commonly used 360-degree image datasets for both indoor and outdoor settings demonstrate the state-of-the-art performance of our proposed method.

In summary, the key contributions of this study are:

- A novel autoregressive outpainting approach is proposed to produce photo-realistic 360-degree images by dynamically adjusting the generation process for improved finer-grained details and prompt-consistency.
- A global-local conditioning mechanism is devised to formulate the guidance encompassing open-vocab text guidances, omni-visual cues, NFoV inputs and omni-geometry with cross-attention based transformers.
- Comprehensive experiments were conducted on two commonly used benchmarks, demonstrating the state-of-the-art performance of AOG-Net in both indoor and outdoor settings with as few as 40 training samples.

Related Work

We first review the studies in both the field of 360-degree image generation and the field of image outpainting which are relevant to our study. As our work takes image and text guidances as conditions, we further review the related studies on conditional image generation.

360-Degree Image Generation

Unlike general NFoV images, 360-degree image generation requires to take the omni-directional continuity into account. Early studies, for example, (Sengupta et al. 2019) estimates a coarse 360-degree image from an NFoV image with inverse rendering technique, which ignores such geometrical continuity and generates 360-degree images lack of fine details. To address this, 360IC (Akimoto et al. 2019) and SIG-SS (Hara, Mukuta, and Harada 2021) were proposed to improve geometrical continuity by taking the intrinsic horizontal cyclicity into consideration and encoding it as positional conditions to connect the two ends of 360-degree images in equirectangular representations. EnvMapNet (Somanath and Kurz 2021) improves visual quality of the outpainted 360-degree images by introducing a projection loss and a clustering loss for accurate lighting and shadowing. OmniDreamer (Akimoto, Matsuo, and Aoki 2022) was further developed by leveraging the Taming-Transformer (Esser, Rombach, and Ommer 2021), where a circular inference scheme was introduced to fit the intrinsic horizontal cyclicity for 360-degree image synthesis, conditioned on provided NFoV images, yielding diverse and photo-realistic results. However, OmniDreamer is limited to a single condition where only an initial NFoV image is accepted, while the controllability of the overall synthesis process is limited. ImmenseGAN (Dastjerdi et al. 2022) aims for finer controllability over the outpainting by introducing a text guidances to fine-tune a generative model with a large-scale private text-image pair dataset. Due to the lack of public text-image paired dataset, Text2Light (Chen, Wang, and Liu 2022) introduces a zero-shot text-guided 360-degree image synthesis pipeline without using initial NFoV images, in which a pre-trained CLIP model is adopted (Radford et al. 2021).

However, the existing methods typically fall short in synthesizing intricate visual details and inconsistencies can be observed between generated images and user-provided text guidances, especially under an open-vocabulary setting, which demands further mechanisms to address these issues.

Image Outpainting

Image outpainting, a fundamental task in computer vision, focuses on expanding the unknown regions outside the primary known content. Unlike inpainting, outpainting may not be able to leverage information from pixels adjacent to the unknown area (Sabini and Rusak 2018; Hoorick 2020; Wang et al. 2019), as seen in inpainting methods. In (Wang et al. 2019), the semantic information of incomplete images was utilized to guide a GAN for outpainting. In (Yao et al. 2022), a query-based outpainting method was proposed, where an image is divided into small patches and the patches with unknown pixels are completed by taking the conditions from both distant and neighbour patches into account. In an iterative manner, (Gardias, Arthur, and Sun 2020) extends one side of the a regular image for outpainting step by step, using the context of the past generation as guidance. (Lu, Chang,

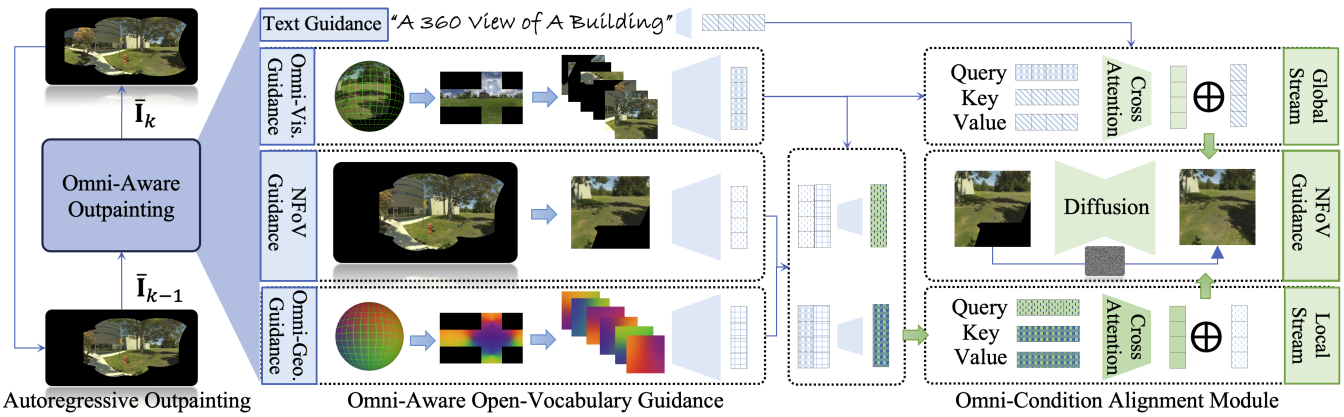

Figure 2: Illustration of the proposed AOG-Net architecture.

and Chiu 2021) delves into the idea of synthesizing unknown regions by exploiting the correlations between distant image patches to establish the global semantics of known pixels. Similarly, in (Esser, Rombach, and Ommer 2021; Chang et al. 2022), image outpainting methods were studied with transformers (Vaswani et al. 2017), which predict the most probable pixel value recursively. However, these conventional outpainting methods do not account for the unique omni-directional continuity inherent to 360-degree images, often leading to discontinuities and artifacts.

Conditional Image Generation

Conditional image generation refers to the synthesis of images based on specific conditions, such as text prompts (Rombach et al. 2022; Kang et al. 2023), semantic maps (Esser, Rombach, and Ommer 2021; Chang et al. 2022) and audio cues (Yariv et al. 2023). For instance, (Isola et al. 2017) achieves conditional image generation using a conditional GAN (Mirza and Osindero 2014) to formulate the joint probability of images and conditions. (Chen et al. 2020; Esser, Rombach, and Ommer 2021) treat an image as a sequence of pixels and therefore generate pixels in an iterative manner. Building on the success of diffusion methods in image generation (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021), various conditional diffusion models have been investigated. For example, (Dhariwal and Nichol 2021) introduces an auxiliary classifier to guide the generation of images within a specific category. (Ho and Salimans 2022) presents a unified framework for conditional generation using diffusion models, introducing a mechanism to control the correlation between the generated image and its input guidance. However, alignment of these conditions with omni-directional geometry is not trivial and further omni-aware alignment strategy is required.

Methodology

As shown in Fig. 2, our proposed AOG-Net for 360-degree image generation follows an autoregressive manner by outpainting a local region progressively. In each step, a global-local conditioning mechanism is introduced to formulate text, omni-visual, NFoV and omni-geometry guidances with cross-attention based transformers into a global stream and a local stream. Such conditions are further adopted a backbone generative prior for the outpainting. The details of these components are discussed in this section.

360-Degree Images & Problem Formulation

Given a 360-degree image, denoted as $\mathbf{I}$, there are three typical representations as shown in Fig. 3 (a) - (c). Each of them can be transformed into the others. Specifically, we have:

- *Spherical representation* $\mathbf{I}(\omega, \phi)$, where ω from $-180°$ to $180°$ denotes the longitude and ϕ indicates the latitude from $-90°$ to $90°$ of a pixel. In practice, cos and sin transforms are adopted for ω and ϕ, respectively, regarding the periodical property for traversal within an image.
- *Cubemap projection* treats $\mathbf{I}$ as a set of general 2D images, which are the faces of a cubic. In detail, we have $\mathbf{I} = \{\mathbf{i}_F, \mathbf{i}_L, \mathbf{i}_B, \mathbf{i}_R, \mathbf{i}_U, \mathbf{i}_D\}$, where each image $\mathbf{i} \in \mathbb{R}^{C \times H_\mathbf{i} \times W_\mathbf{i}}$ can be viewed as a general NFoV image, where $H_\mathbf{i}$ and $W_\mathbf{i}$ denote the height and the width of a face, respectively, and C is the number of channels.
- *Equirectangular projection* maps $\mathbf{I}$ to a general image in $\mathbb{R}^{C \times H_\mathbf{I} \times W_\mathbf{I}}$, where $H_\mathbf{I}$ and $W_\mathbf{I}$ indicate height and width, respectively. Compared to cubemap projection, equirectangualr maps the entire spherical 360-degree image into a single rectangular grid, characterized by its noticeable pixel distortation around the top and bottom regions.

As spherical representation inherently conforms to the 360-degree geometry coordinates, we project these geometry information to the cubemap form as shown in Fig. 3 (d) - (e). We denote such geometry information as $\mathbf{\Gamma} = \{\gamma_F, \gamma_L, \gamma_B, \gamma_R, \gamma_U, \gamma_D\}$, where $\gamma_\cdot \in \mathbb{R}^{2 \times H_\mathbf{i} \times W_\mathbf{i}}$ contains the geometry information of a cubic face.

Given an NFoV image $\mathbf{X} \in \mathbb{R}^{C \times H_\mathbf{X} \times W_\mathbf{X}}$, such as a 2D image taken by smartphones, where $H_\mathbf{X}$ and $W_\mathbf{X}$ are its height and width, respectively; and a text guidance with its embedding $\mathbf{T} \in \mathbb{R}^{C_T \times L}$, where C_T is the dimension of textual feature and L is the length of text guidance, our method aims to synthesize a 360-degree image $\hat{\mathbf{I}}$ by given $\mathbf{X}$ and $\mathbf{T}$.

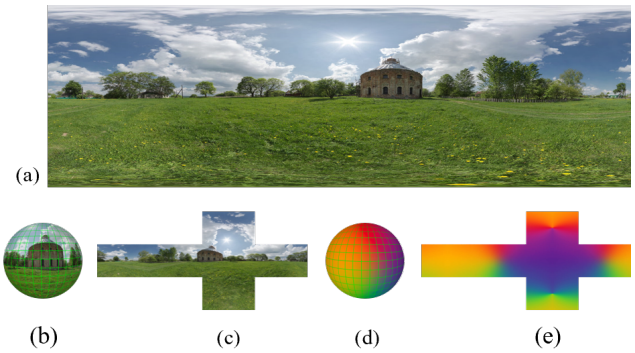

Figure 3: Different representations of a 360-degree image. (a) Equirectangular projection. (b) Spherical representation. (c) Cubemap projection. (d) A spherical representation with geometry coordinates. (e) Geometry projection on cubemap.

Autoregressive Omni-Traversal for Outpainting

The autoregressive process outpaints the given NFoV image $\mathbf{X}$ progressively to a complete 360-degree image $\hat{\mathbf{I}}$ under the guidance of the text $\mathbf{T}$. Specifically, each step completes a local NFoV view, which is extracted from the incomplete 360-degree image with an unknown region that is neighbouring to a known region.

Local View Projection & Backprojection. To leverage a wide range of NFoV domain knowledge such as the weights from large-scale pretrained weights on NFoV image datasets, we correspondingly retrieve a local view from a 360-degree image $\mathbf{I}$ centred at the location ω and ϕ as a forward projection. In detail, we project the local view in $\mathbf{I}$ to an NFoV image $\mathbf{X}$ via a gnomonic projection, denoted as $\mathbf{X} = O(\mathbf{I}, \omega_{\mathbf{X}}, \phi_{\mathbf{X}})$, where $\omega_{\mathbf{X}}$ and $\phi_{\mathbf{X}}$ are the centroid longitude and latitude of the local view. Similarly, we have a backprojection - an inverse gnomonic projection maps the pixels in $\mathbf{X}$ back to $\mathbf{I}$ partially within the scope of the corresponding NFoV view, denoted as $\tilde{\mathbf{I}} = O^{-1}(\mathbf{X}, \omega_{\mathbf{X}}, \phi_{\mathbf{X}})$. Note that the pixel value out of the scope of $\mathbf{X}$ in $\tilde{\mathbf{I}}$ is defined as $-inf$. Furthermore, we define an operator $\oplus$ for two 360-degree inputs: $\mathbf{I}_\alpha$ and $\mathbf{I}_\beta$ as:

$$\mathbf{I}_\alpha \oplus \mathbf{I}_\beta(\omega, \phi) = \begin{cases} \mathbf{I}_\alpha(\omega,\ \phi) \text{ if } \mathbf{I}_\alpha(\omega,\ \phi) \neq -inf, \\ \mathbf{I}_\beta(\omega,\ \phi) \text{ otherwise,} \end{cases} \tag{1}$$

which is used for attaching a newly generated partial 360-degree image to the current incomplete image.

Single-Step Outpainting. In a single-step outpainting, without loss of generality, for the k^{th} step, an incomplete NFoV image $\bar{\mathbf{X}}_k = O(\bar{\mathbf{I}}_k, \omega_{\bar{\mathbf{X}}_k}, \phi_{\bar{\mathbf{X}}_k})$ is retrieved from an incomplete 360-degree image $\bar{\mathbf{I}}_k$. Particularly, we denote a conditioned outpainting model $F_{\mathbf{\Theta}}$, where $\mathbf{\Theta}$ are learn-able parameters. $F_{\mathbf{\Theta}}$ estimates the unknown pixels in $\bar{\mathbf{X}}_k$, where the outpainted results is denoted as $\hat{\mathbf{X}}_k = F_{\mathbf{\Theta}}(\bar{\mathbf{X}}_k, \bar{\mathbf{I}}_k, \mathbf{T})$. The estimation $\hat{\mathbf{X}}_k$ is then backprojected to 360-degree view and a 360-degree outpainted estimation can be obtained as $\hat{\mathbf{I}}_k = O^{-1}(\hat{\mathbf{X}}_k, \omega_{\hat{\mathbf{X}}_k}, \phi_{\hat{\mathbf{X}}_k}) \oplus \bar{\mathbf{I}}_k$. Note that $\omega_{\hat{\mathbf{X}}_k} = \omega_{\bar{\mathbf{X}}_k}$ and $\phi_{\hat{\mathbf{X}}_k} = \phi_{\bar{\mathbf{X}}_k}$ as $\hat{\mathbf{X}}_k$ retains its omni-geometry location. More details about $F_{\mathbf{\Theta}}$ can be found in the subsequent discussions.

Generally, $\bar{\mathbf{I}}_1 = O^{-1}(\mathbf{X}, \omega_{\mathbf{X}}, \phi_{\mathbf{X}})$ is initialized with the input NFoV image $\mathbf{X}$. To optimize $F_{\mathbf{\Theta}}$, the known pixels in $\bar{\mathbf{X}}_k$ and $\bar{\mathbf{I}}_k$ can be extracted from the ground truth $\mathbf{I}$; we denote $\mathbf{X}_k$ and $\mathbf{I}_k$ as the ground truth for the k^{th} step. For inference, the known pixels in $\bar{\mathbf{X}}_k$ and $\bar{\mathbf{I}}_k$ can be based on the accumulated estimations $\hat{\mathbf{X}}_{k-1}$ and $\hat{\mathbf{I}}_{k-1}$, respectively.

Autoregressive Outpainting. Following an autoregressive stochastic process, a 360-degree image can be produced progressively:

$$p(\mathbf{I}|\mathbf{T}) = \prod_k p(\mathbf{I}_k|\mathbf{I}_{<k}, \mathbf{T}), \tag{2}$$

where $\mathbf{I}_{<k}$ indicates $\mathbf{I}_1$,..., $\mathbf{I}_{k-1}$, which are incomplete 360-degree images. As our proposed method mainly outpaints a small portion of unknown pixel in an incomplete 360-degree image, Eq. (2) can be written with the Markov property:

$$p(\mathbf{I}|\mathbf{T}) = \prod_k p(\mathbf{I}_k|\bar{\mathbf{I}}_k, \mathbf{T}) = \prod_k p(\mathbf{X}_k|\,\bar{\mathbf{I}}_k, \mathbf{T}). \tag{3}$$

In line with the single-step outpainting, $F_{\mathbf{\Theta}}$ is used to compute the conditioned probability terms as:

$$p(\mathbf{I}|\mathbf{T}) \approx \prod_k F_{\mathbf{\Theta}}(\bar{\mathbf{X}}_k, \bar{\mathbf{I}}_k, \mathbf{T}). \tag{4}$$

To this end, an autoregressive stochastic process has been formulated for 360-degree image generation. Note that we use an incremental pathway to identify $\bar{\mathbf{X}}_k$ that prioritizes the generation process along the longitude direction.

Global-Local Conditioning by Omni-Aware Open-Vocabulary Guidance

AOG-Net incorporates multiple conditions to ensure its alignment to user text guidances and known NFoV views regarding the omni-geometry. Specifically, in each autogressive step, a global-local conditioning mechanism is devised to thoroughly capture the following conditions:

- Text guidance $\mathbf{c}_{\text{text}}$: a text encoder $\mathcal{E}_{\text{text}}$ encodes user text description $\mathbf{T}$, which is based on the CLIP pre-trained textual model and enables an open-vocabulary paradigm to align the text features within a latent space shared with visual patterns. Note that this text guidance remains constant for each autoregressive step k and acts as a global context. However, it can be modified according to user preferences to adjust during the generation process.
- Omni-visual guidance $\mathbf{c}_{360,k}$: a visual encoder $\mathcal{E}_{360}$, which leverages the CLIP pre-trained visual model, transforms a 360-degree image into the latent space that is shared with $\mathbf{c}_{\text{text}}$. Specifically, we encode each face in the cubemap representation of $\bar{\mathbf{I}}_k$ and denote the results as $\mathbf{c}_{360,k} = \{\mathbf{c}_{F,k}, \mathbf{c}_{L,k}, \mathbf{c}_{B,k}, \mathbf{c}_{R,k}, \mathbf{c}_{U,k}, \mathbf{c}_{D,k}\}$.
- NFoV guidance $\mathbf{c}_{\text{NFoV},k}$: a visual encoder $\mathcal{E}_{\text{NFoV}}$ encodes the incomplete NFoV image $\bar{\mathbf{X}}_k$ jointly with the 360-degree image $\bar{\mathbf{I}}_k$ in its cubemap form aiming for a omni-visual local latent representation.
- Omni-geometry guidance $\mathbf{c}_{\text{geometry},k}$: an omni-geometry encoder $\mathcal{E}_{\text{geometry}}$ formulates the geometry $\bar{\gamma}_k$ of an incompleted local NFoV image $\bar{\mathbf{X}}_k$, jointly with $\mathbf{\Gamma}$, to introduce the omni-geometry information for outpainting.

Global-Local Conditioning. This module aligns the derived conditions for 360-degree outpainting through both a global and a local stream, leveraging cross-attention mechanisms. Globally, the incomplete 360-degree visual guidance $\mathbf{c}_{360,k}$ is cross-referenced with the text guidance $\mathbf{c}_{\text{text}}$ to guarantee alignment between the content that already presents in $\mathbf{c}_{360,k}$ and the content awaiting generation. Intuitively, we adopt a cross-attention based transformer for this purpose by treating the query as visual conditions $\mathbf{c}_{360,k}$, while the value and key are as text conditions $\mathbf{c}_{\text{text}}$. We denote the results as a global condition $\mathbf{c}_{\text{global},k}$.

Likewise, the local stream incorporates the NFoV guidance and the omni-geometry guidance using a transformer grounded in cross-attention. This integration facilitates the local fine-grained details during the generation process. Specifically, the query adopts the NFoV condition $\mathbf{c}_{\text{NFoV},k}$ supplemented by $\mathbf{c}_{\text{geometry},k}$, while the value and the key are with the 360-degree visual guidance $\mathbf{c}_{360,k}$ supplemented by $\mathbf{c}_{\text{geometry},k}$. The resultant local condition is denoted as $\mathbf{c}_{\text{local},k}$.

Omni-Aware Diffusion for Outpainting

Leveraging the recent success of the diffusion approach for NFoV content generation, in each autoregressive step, $F_{\mathbf{\Theta}}$ employs a stable diffusion backbone (Rombach et al. 2022), incorporating the conditions $\mathbf{c}_{\text{global},k}$ and $\mathbf{c}_{\text{local},k}$. For the k^{th} atuoregressive step, in a diffusion, we further denote t as the diffusion temporal index and $\epsilon_{\mathbf{\Theta}}(\mathbf{z}_t, t)$ as the predicted noise introduced in the t^{th} step, where $\epsilon_{\mathbf{\Theta}}$ is a U-Net. To optimize $\epsilon_{\mathbf{\Theta}}$, we minimize the following loss function:

$$\mathcal{L} := \mathbb{E}_{\epsilon_t \sim \mathcal{N}(0,1),t}\left[\left\|\epsilon_t - \epsilon_\theta\left(\mathbf{z}_t, t, \tau_{\mathbf{\Theta}}(\mathbf{c}_{\text{global},k}, \mathbf{c}_{\text{local},k})\right)\right\|_2^2\right], \tag{5}$$

where $\tau_{\mathbf{\Theta}}$ maps the conditions to guide the denoising process in the latent space via a cross-attention mechanism (Vaswani et al. 2017).

Experiments & Discussions

Datasets

360-Degree Images. Following the existing studies (Akimoto, Matsuo, and Aoki 2022; Somanath and Kurz 2021), we evaluate our proposed method with the LAVAL indoor HDR dataset (Gardner et al. 2017) for the 360-degree indoor image generation setting, which contains 2,233 360-degree images for extensive indoor scenes with a resolution of $7,768 \times 3,884$. For a fair comparison, we used the official training and testing split in our experiments, in which we have 1,921 training samples and 312 testing samples.

For the outdoor setting, we utilize the LAVAL outdoor HDR dataset (Zhang and Lalonde 2017), which contains 210 360-degree images with a resolution of $7,768 \times 3,884$. In this setting, we randomly sample 170 images as the training split and 40 images for testing purpose. In the training of both settings, the resolution of 360-degree image is downsampled to $4,096 \times 2048$ for computation efficiency.

Text Captioning. As the lack of text captions in both datasets, we adopted a large-scale captioning model BLIP2 (Li et al. 2023) to generate captions for 360-degree images. We first caption an image in its equirectangular form to obtain an overall text guidance with an average of 5-6 words. Next, we caption the horizontal faces individually of its cubemap to obtain additional text guidances.

Data Augmentation. To increase the diversity of the 360-degree images generated, we augmented the training 360-degree image samples by adopting random clockwise rotation based on the intrinsic horizontal cyclicity. To improve the diversity of text guidance, besides randomly swapping words with TextAttack (Morris et al. 2020), we randomly combines the overall text guidance and one randomly selected text guidance associated with a face of the cubemap during training.

Implementation Details

Pre-Trained Models & Network Architecture. In our experiment, we adopted the pretrained Stable Diffusion generative prior for each autoregressive generation step. In addition, We utilized the visual encoder and the text encoder of OpenCLIP (Cherti et al. 2023) for $\mathcal{E}_{360}$ and $\mathcal{E}_{\text{text}}$, respectively. We utilized T2I-Adapter (Mou et al. 2023) as the architecture for NFoV guidance encoder $\mathcal{E}_{\text{NFoV}}$ and omni-geometry guidance encoder $\mathcal{E}_{\text{geometry}}$. In both local stream and global stream, we utilized a 16-layer cross-attion base transformer to compute $\mathbf{c}_{\text{local},k}$ and $\mathbf{c}_{\text{global},k}$ respectively.

Training and Inference. AOG-Net was trained using an AdamW optimizer (Loshchilov and Hutter 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. It was trained for 240 epochs, with learning rate 1×10^{-4} and batch size 1. For inference, we leveraged DPM-Solver++ (Lu et al. 2023) as sampler with a step set to 25 and classifier-free-guidance (Ho and Salimans 2022) scale set to 2.5. All experiments were conducted on an Nvidia RTX 3090.

Comparison with State of the Art

Baselines. Our method is compared with the recent state-of-the-art 360-degree image outpainting methods from three perspectives. 1) NFoV image guided generation methods without text guidance: ImmerseGAN (Dastjerdi et al. 2022), OmniDreamer(Akimoto, Matsuo, and Aoki 2022) and EnvMapNet (Somanath and Kurz 2021). For a fair comparison, the text guidance in our method is set as a blank prompt. 2) Text-guided generation method without NFoV guidance - Text2Light (Chen, Wang, and Liu 2022). In this case, we generated our initial input NFoV image using Stable Diffusion Outpainting model for our method. 3) NFoV image and text guided generation method (Dastjerdi et al. 2022).

Evalution Metrics. To quantitatively evaluate our AOG-Net, we adopted LPIPS (Zhang et al. 2018) and Fréchet Inception Distance (FID) (Heusel et al. 2017) as the evaluation metrics to measure the similarity of latent representations between the generated 360-degree images and the ground truth. To evaluate the semantic consistency (SC) between the generated 360-degree image and the input text guidance, we compared the similarity between input text guidance and the captioning texts obtained from the generated image. Specifically, we leveraged a large-scale captioning model BLIP2 (Li et al. 2023) and computed the similarity with sentence embeddings (Reimers and Gurevych 2019) for this purpose. In addition, we leverage Inception Score (IS) (Salimans et al.

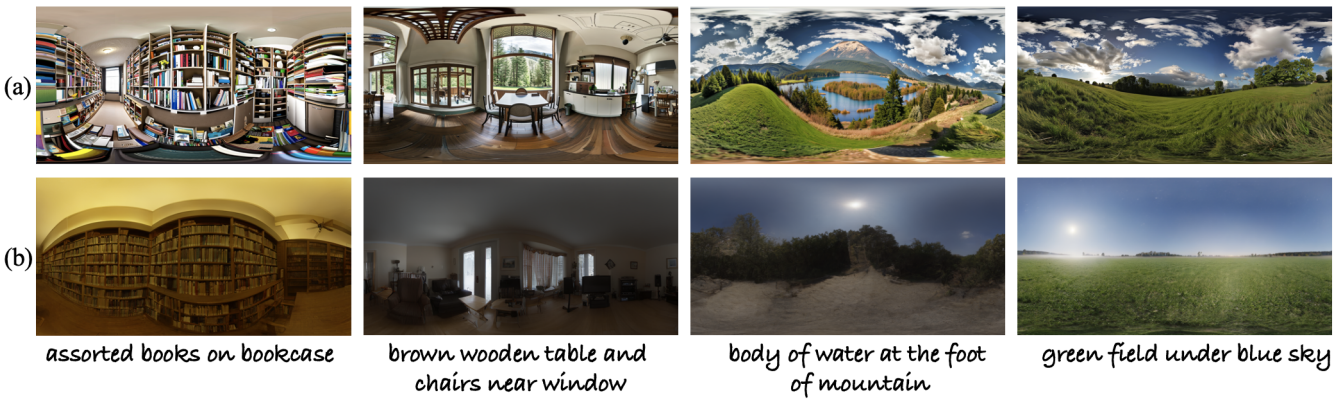

Figure 4: Comparison between Text2Light for indoor and outdoor settings. (a) Ours. (b) Text2Light.

2016) to measure the quality of the generated images as Text2Light does not involve ground truth images.

Method	FID ↓	LPIPS↓
Indoor setting		
SIG-SS (2017)	197.4	-
EnvMapNet (2021)	52.70	-
OmniDreamer (2022)	46.15	0.45
ImmenseGAN(2022)	42.78	-
AOG-Net (Ours)	**38.60**	**0.37**
Outdoor setting		
OmniDreamer (2022)	24.5	0.41
AOG-Net (Ours)	**18.4**	**0.36**

Table 1: Comparison with the state-of-the-art methods using NFoV image guidance.

Method	SC ↑	IS↑
Outdoor setting		
Text2Light (2022)	0.45	3.9
AOG-Net (Ours)	**0.53**	**4.2**
Indoor setting		
Text2Light (2022)	0.33	4.5
AOG-Net (Ours)	**0.36**	**5.1**

Table 2: Comparison with the state-of-the-art methods using text guidance.

Overall Performance. For the methods requiring an initial NFoV image as guidance, our method achieve the best performance as shown in Table 1. It achieves an FID score 35.6 and an LPIPD value 0.37 under an indoor setting and an FID score 18.4 and an LPIPS value 0.36 under an outdoor setting. Note that only OmniDreamer conducted evaluation for outdoor setting in the literature. As shown in the first example (in the first column) in Fig. 5, our method outpaints the house and the garden smoothly, while OmniDreamer make the neighbouring region of the house smudged with sudden color changes in the garden. For the third example (in the third column), our method is able to deliver detailed outpainting regarding objects compared to Omnidremaer.

Regarding the comparison with text-conditioned method - Text2Light with open vocabulary text guidances, the performance metrics are listed in Table 2. Our method outperform Text2Light with an SC score 0.53 and an IS value 4.2 for an outdoor setting and an SC score 0.36 and an IS value 5.1 for an indoor setting. Due to the complexity of the indoor setting and the lack of in-depth text description, the semantic consistence of both methods drop. However, our method still provide overall better semantically consistent images with higher image quality. As depicted in Figure 4, our method is able to produce visually appealing images, while the images of Text2Light (Chen, Wang, and Liu 2022) are much dimmer, leading to degenerated visual quality and lack of details. Additionally, under the outdoor setting, our method generates 360-degree image with fine-grained details (such as trees, grasses), while the Text2Light produces smudged-out patterns in images (in the third column of Fig. 4).

For ImmenseGAN, which leverages both NFoV and text guidances, our method peforms superior under the indoor setting according to the metrics the authors reported in their work. ImmenseGAN was trained with a private large-scale dataset and the authors did not evaluate their method under an outdoor setting. Note that our method leverages the pretrained diffusion models and only requires 40 randomly selected training samples to achieve its current performance.

Ablation Study

Ablation studies are conducted to demonstrate the effectiveness of individual components in AOG-Net. The results are listed in Table 3 and an example is shown in Fig. 6.

Global Guidance. The global guidance $\mathbf{c}_{global}$ and its related components are removed from the pipeline. The backbone model is only guided by CLIP text guidance embedding and $\mathbf{c}_{local}$. Although the presence of the text guidance provides a coarse global condition, the overall semantic consistency is decreased. As shown in Fig.6 (c), the model fails to deliver a consistent floor texture without global guidance.

Local Guidance. The local guidance $\mathbf{c}_{local}$ and its related

Figure 5: Qualitative examples with input and ground truth. (a) Input, 90° in both longitude and latitude direction. (b) grount truth. (c) OmniDreamer. (d) AOG-Net (Ours).

components are excluded from our method. While the semantic consistency between the outputs and the text prompts is only slightly affected, the deterioration in the quality of the outpainted images is significant. In Fig. 6 (d), there are various artifacts such as black patches and human hands.

Geometry Guidance. In this setting, we removes all 360-degree geometry information $\mathbf{c}_{\text{geo}}$ in computing $\mathbf{c}_{local}$, which would only rely on pixel-wise semantics to connect distant patches. The results reveal minimal effects on SC, but there is a notable decrease in image quality. As illustrated in Fig. 6 (e), black patterns appear on the floor and the ceiling's color lacks consistency with distant regions.

Backbone Only. In this setting, only the pre-trained Stable Diffusion backbone is employed in a traditional manner, integrating the NFoV input image and text guidance. This settings produces the poorest SC values and FID scores, suggesting that the outpainted 360-degree images are of low quality and misaligned with the text guidance. Referring to the generation example shown in Fig. 6 (f), the model struggles to produce a text-consistent and sharp 360-degree image outpainting, with evident localized artifacts.

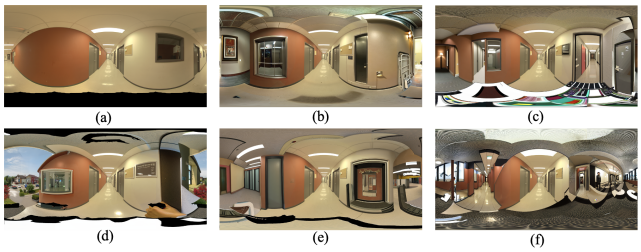

Figure 6: Ablation study. (a) Ground truth. (b) AOG-Net. (c) w/o global condition. (d) w/o local condition. (e) w/o geometry condition. (f) Autoregressive w/ backbone only.

Generalization

We further explore an open-image conditioned task. Our method is required to outpaint unseen oil painting artworks to 360-degree images with text guidances. As shown in Fig. 7, the generated images are consistent with the style of input NFoV artworks, demonstrating the potential of our method in accepting out-of-domain NFoV image as conditions.

Method	LPIPS↓	FID↓	SC↑
AOG-Net (Ours)	<u>0.37</u>	**35.6**	**0.72**
w/o global condition	0.38	40.08	0.70
w/o local condition	0.40	47.46	0.71
w/o geometry condition	**0.36**	<u>37.2</u>	**0.72**
Autoregressive w/ backbone only	0.43	67.4	0.61

Table 3: Ablation study on the Laval Indoor HDR dataset.

Figure 7: Open-image conditioned generation results, with the prompt "a 360 image of a city, oil painting , ultracolorful, impressionist style, Van Gogh style". (a) Input. (b) Output.

Limitation and Future Work

AOG-Net relies on a pre-trained backbone model, which introduces two primary limitations. Firstly, AOG-Net is somewhat constrained by the data on which the backbone model was pre-trained, potentially, this method may suffer from the internal biases introduced by the backbone diffusion model. Secondly, the diffusion model's prolonged inference time affects its utility in applications that require real-time performance. Future endeavors might focus on developing a backbone or the exploration of conditional 360-degree image generation similar to (Stan et al. 2023). Finally, by capitalizing on the autoregressive characteristics, our method has the potential to be extended to facilitate text-guided 360-degree video generation.

Conclusion

In this work, we present a novel deep learning method AOG-Net for 360-degree image generation with an autoregressive scheme guided by NFoV images and open vocabulary text prompts. A global-local conditioning mechanism is devised to adaptively encode guidances considering omni-directional properties. With these design, AOG-Net is able to generate realistic 360-degree image with fine details while aligning with the text guidance. Comprehensive experiments demonstrate the effectiveness of AOG-Net.

Acknowledgments

This study was partially supported by Australian Research Council (ARC) grant #DP210102674.

References

Akimoto, N.; Kasai, S.; Hayashi, M.; and Aoki, Y. 2019. 360-Degree Image Completion by Two-Stage Conditional Gans. In *IEEE International Conference on Image Processing*.

Akimoto, N.; Matsuo, Y.; and Aoki, Y. 2022. Diverse Plausible 360-Degree Image Outpainting for Efficient 3DCG Background Creation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. MaskGIT: Masked Generative Image Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020. Generative Pretraining from Pixels. In *International Conference on Machine Learning*.

Chen, Z.; Wang, G.; and Liu, Z. 2022. Text2Light: Zero-Shot Text-Driven HDR Panorama Generation. *ACM Transactions on Graphics*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Dastjerdi, M. R. K.; Hold-Geoffroy, Y.; Eisenmann, J.; Khodadadeh, S.; and Lalonde, J.-F. 2022. Guided Co-Modulated GAN for 360° Field of View Extrapolation. In *International Conference on 3D Vision*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gardias, P.; Arthur, E.; and Sun, H. 2020. Enhanced Residual Networks for Context-based Image Outpainting. arXiv:2005.06723.

Gardner, M.-A.; Sunkavalli, K.; Yumer, E.; Shen, X.; Gambaretto, E.; Gagné, C.; and Lalonde, J.-F. 2017. Learning to Predict Indoor Illumination from a Single Image. arXiv:1704.00090.

Hara, T.; Mukuta, Y.; and Harada, T. 2021. Spherical Image Generation from a Single Image by Considering Scene Symmetry. *AAAI Conference on Artificial Intelligence*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *2017 Advances in Neural Information Processing Systems*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598.

Hoorick, B. V. 2020. Image Outpainting and Harmonization using Generative Adversarial Networks. arXiv:1912.10960.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2023. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. arXiv:2211.01095.

Lu, C.-N.; Chang, Y.-C.; and Chiu, W.-C. 2021. Bridging the visual gap: Wide-range image blending. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. arXiv:2302.08453.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sabini, M.; and Rusak, G. 2018. Painting Outside the Box: Image Outpainting with GANs. arXiv:1808.08483.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in Neural Information Processing Systems*.

Sengupta, S.; Gu, J.; Kim, K.; Liu, G.; Jacobs, D. W.; and Kautz, J. 2019. Neural inverse rendering of an indoor scene from a single image. In *IEEE/CVF International Conference on Computer Vision*.

Somanath, G.; and Kurz, D. 2021. HDR environment map estimation for real-time augmented reality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Stan, G. B. M.; Wofk, D.; Fox, S.; Redden, A.; Saxton, W.; Yu, J.; Aflalo, E.; Tseng, S.-Y.; Nonato, F.; Muller, M.; et al. 2023. LDM3D: Latent Diffusion Model for 3D. *arXiv preprint arXiv:2305.10853*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*.

Wang, Y.; Tao, X.; Shen, X.; and Jia, J. 2019. Wide-Context Semantic Image Extrapolation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yao, K.; Gao, P.; Yang, X.; Sun, J.; Zhang, R.; and Huang, K. 2022. Outpainting by Queries. In *European Conference on Computer Vision*.

Yariv, G.; Gat, I.; Wolf, L.; Adi, Y.; and Schwartz, I. 2023. AudioToken: Adaptation of Text-Conditioned Diffusion Models for Audio-to-Image Generation. arXiv:2305.13050.

Zhang, J.; and Lalonde, J.-F. 2017. Learning High Dynamic Range from Outdoor Panoramas. In *2017 IEEE International Conference on Computer Vision*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Cision and Pattern Recognition*.