CONSISTENT VIEW SYNTHESIS WITH BIDIRECTIONAL EPIPOLAR ATTENTION AND RECONSTRUCTION

I-Chung Chiu[1] *Jun-Cheng Chen*[2] *I-Hong Jhuo*[3] *Yen-Yu Lin*[1]

[1]National Yang Ming Chiao Tung University [2]Academia Sinica [3]Microsoft AI

ABSTRACT

Novel view synthesis from a single image aims to generate novel scene views given a reference image and a sequence of camera poses. Its primary difficulty lies in effectively leveraging a generative model to achieve high-quality image generation while simultaneously ensuring consistency and faithfulness across synthesized views. In this paper, we propose a novel approach to address the *consistency* and *faithfulness* issues in view synthesis. Specifically, we develop a new attention layer, termed bidirectional epipolar attention, which utilizes a pair of complementary epipolar lines to guide the associations between features from different viewpoints. Each bidirectional epipolar layer calculates forward and backward epipolar lines, enabling geometrically constrained attention that improves cross-view consistency. To ensure faithful synthesis, we introduce an epipolar-aware reconstruction module that prevents creating novel content in regions where the newly generated image overlaps with existing ones. Extensive experimental results demonstrate that our method outperforms previous approaches to novel view synthesis, achieving superior performance in both image quality and consistency. The source code is available at https://github.com/fallantbell/Bidirectional-Epipolar-Synthesis.

***Index Terms*—** Novel view synthesis, epipolar, transformers

1. INTRODUCTION

Novel view synthesis from a single image has gained widespread attention. It involves generating new perspectives of a scene based solely on a single image and a series of camera poses. Its ability to create immersive content makes it highly valuable in fields such as virtual reality [1], film production [2], and gaming [3]. However, its primary challenge lies in ensuring consistency between the generated images and the previous ones. Specifically, in regions where synthesized views overlap, the newly generated images should accurately maintain the geometric relationships in prior views, while preventing the introduction of any content unseen previously.

Several prior works have explored novel view synthesis from a single image. One prominent line of research [4, 5, 6, 7] employs a *render-refine-repeat* strategy to accomplish this task. These approaches use a monocular depth estimation model, e.g., [8], to predict the depth map of the input image, warp the image based on the depth information, and apply generative models to perform inpainting and outpainting in invisible regions. While these methods effectively generate new views, they remain limited by the accuracy of the underlying depth estimation models. More recent studies [9, 10, 11, 12, 13] have adopted implicit methods, employing transformers or diffusion models to learn correspondences between the input and output images. Despite their ability to generate plausible content, they still struggle to ensure geometric consistency and introduce artifacts or content not present in prior views.

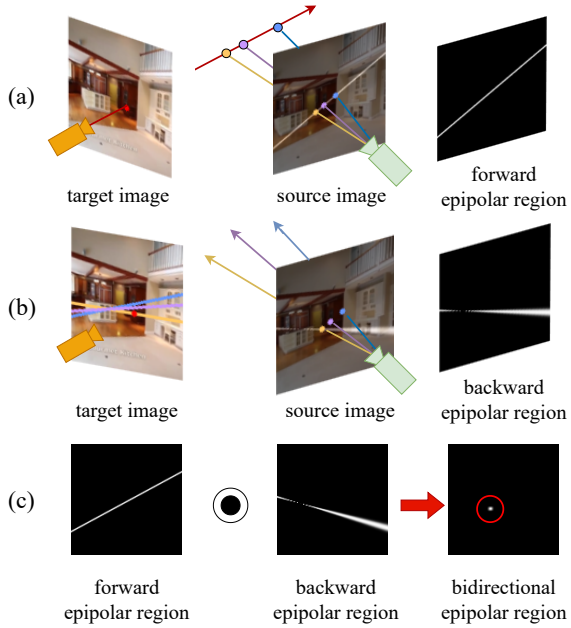

Fig. 1: **Bidirectional epipolar attention.** To determine the relationship between a specific point (red point here) in the target image and the source image: (a) **Forward epipolar** identifies the epipolar line in the source image corresponding to the red point, (b) **Backward epipolar** traces back each point in the source image, finding those whose backward epipolar lines intersect the red point in the target image, and (c) **Bidirectional epipolar attention** computes the intersection of the forward and backward epipolar regions.

To address the issue of consistency, pose-guided diffusion models [11] further incorporate epipolar constraints to relate features across views. They apply epipolar lines to guide attention mechanisms, restricting each pixel in the target (synthesized) image to attend solely to regions along its corresponding epipolar line in the source image. This approach improves cross-view consistency but still exhibits ambiguity, because each pixel attends to an entire line rather than the corresponding point, as shown in Fig. 1a.

To reduce ambiguous attention, we propose the *bidirectional epipolar attention*. Specifically, we also employ the *forward* epipolar line for each pixel in the target image to identify its potential correspondences within the source image. Subsequently, we compute the *backward* epipolar lines for pixels in the source image, as shown in Fig. 1b. Pixels in the source image, whose backward epipolar lines are close to the target pixel, are retained as candidate correspondences for that target pixel. For each target pixel, we take the intersection of the two sets of candidate correspondences, significantly narrowing the attention region and reducing ambiguity as illustrated in Fig. 1c.

Despite the improved view-to-view consistency provided by epipolar attention mechanisms, synthesized views may still exhibit novel content in the overlapping regions with source images, which is caused by the expressive power of a generative model. To alleviate this issue, we introduce an epipolar-based reconstruction method. By referring to the attention map and epipolar lines jointly, we identify pixels that attend to geometrically incorrect regions, which are prone to generating novel content. We apply reconstruction to these pinpointed pixels by utilizing information only from the source image to ensure faithful synthesis.

We evaluate our approach on the RealEstate10K dataset, which consists of 3D indoor scenes. The experimental results demonstrate that our method produces more consistent and high-quality synthesized images. Specifically, the proposed bidirectional epipolar attention module can greatly enhance the consistency across views, leading to better synthesis qualities. Additionally, the introduced epipolar-based reconstruction method effectively alleviates artifacts and the creation of novel content in the overlapping areas, resulting in more faithful synthesis. Our method achieves state-of-the-art performance in multiple metrics, including PSNR, SSIM [14], and LPIPS [15], surpassing previous approaches.

2. RELATED WORK

Novel view synthesis (NVS) has beena prominent research area within computer vision [16, 17, 18]. Its objective is to generate synthetic views of a scene from a limited set of images captured from different viewpoints. Prior work [12] categorizes NVS methods into two primary groups: view interpolation and view extrapolation. View interpolation aims to reconstruct novel views within small camera motion ranges by referring to multiple input images of the same scene. In contrast, view extrapolation begins with a single input image and generates novel views with substantially larger camera motion, often producing unseen portions of the scene.

For view interpolation tasks, Neural Radiance Fields (NeRFs) [19, 20, 21] have demonstrated exceptional capabilities. NeRFs derive an implicit 3D representation of the scene using neural networks, enabling the synthesis of high-quality novel views from sparse input images. Recently, Gaussian Splatting [22, 23] has achieved remarkable results by representing the scene with 3D Gaussians and optimizing them for faithful reconstruction, enabling high-quality synthesis with significantly improved rendering speed.

In contrast to view interpolation focusing on novel view synthesis within small camera motions, view extrapolation aims to generate new content for larger camera movements, extending beyond the original image boundaries. Methods such as Inf [5] and Inf-Zero [4] employ a "render-refine-repeat" strategy, where images are warped based on estimated depth and refined using GAN-based generative models. Scenescape [24] and Text2Room [25] construct 3D meshes based on estimated depth to maintain 3D consistency, while using stable diffusion models [26] to generate unseen regions. In contrast, GeoGPT [10] employs an autoregressive transformer to generate novel scenes without relying on monocular depth estimation. LoR [9] extends GeoGPT [10] by conditioning the generation process on the two preceding images to enhance temporal consistency. More recently, PhotoNVS [12] adopts a two-stream UNet within a diffusion model framework, while Pose-Guided Diffusion [11] employs epipolar attention to enhance consistency. Despite these advancements, producing high-quality, faithful, and cross-view consistent images remains challenging. To address this, we propose bidirectional epipolar attention and reconstruction to fulfill these requirements and achieve more realistic results.

3. METHOD

3.1. Overview

In the task of novel view synthesis from a single image, the input consists of an initial image X_1 and a sequence of camera poses $\{C_i\}_{i=2}^{n}$. The goal is to generate a sequence of images $\{X_i\}_{i=2}^{n}$ corresponding to the given camera poses. As illustrated in Fig. 2, for a given image at the i-th viewpoint $X_i \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the image's height and width, respectively, and the next viewpoint's camera pose C_j, typically $j = i + 1$, we first encode them into the latent space using an image encoder and a camera encoder. Subsequently, a transformer is employed to predict the next image $\hat{X}_j$. To ensure faithfulness to X_i, $\hat{X}_j$ undergoes a refinement process via reconstruction to generate the final output $\tilde{X}_j$.

Image Encoding. For image encoding, we follow previous work [9, 10] by using a pre-trained VQGAN [27]. Given a pre-trained encoder E and a codebook $O = \{o_y \in \mathbb{R}^{d_o}\}_{y=1}^{|O|}$ with $|O|$ entries, the source image X_i is encoded as:

$$z_i = E(X_i) \in \mathbb{R}^{hw\times d_o}, \tag{1}$$

where hw represents the number of tokens obtained by dividing the image into $h \times w$ patch tokens. Each token $z_{i,k}$, which is a latent representation in the token space, is then quantized by finding the closest codebook entry. Specifically, for token $z_{i,k}$, we identify the nearest codebook index y by minimizing the squared distance between the token and the codebook entry. This gives the index $I_{i,k}$:

$$I_{i,k} = \arg\min_{y} \|z_{i,k} - o_y\|^2 \quad \text{and} \quad I_i \in \mathbb{R}^{hw}, \tag{2}$$

where the quantized tokens from the learned codebook O are indexed by the resulting sequence of indices $I_i = \{I_{i,k}\}_{k=1}^{hw}$.

The indices I_i are then mapped to image embeddings $H_i^I \in \mathbb{R}^{hw\times d}$ using an embedding layer $\lambda(\cdot)$, which maps integer indices to embedding vectors of a fixed-dimension d, namely

$$H_i^I = \lambda(I_i). \tag{3}$$

Camera Encoding. For camera encoding, the camera pose $C_j = (\mathbf{K}, \mathbf{R}_{i\to j}, \mathbf{t}_{i\to j})$, representing the intrinsic parameters, rotation, and translation from view i to view j respectively, are first flattened and then concatenated into a vector $P_j \in \mathbb{R}^N$, where N denotes the number of parameters. A linear transformation $E_c(\cdot)$ is then applied to map P_j into the camera embeddings $H_j^C \in \mathbb{R}^{N\times d}$, i.e.,

$$H_j^C = E_c(P_j). \tag{4}$$

Transformer. As shown in Fig. 2, our transformer model comprises alternating layers of conventional self-attention and the proposed bidirectional epipolar attention, with an additional MLP layer at the end to predict the output. During training, we follow the previous approach [9] using a GPT-like architecture [28]. Specifically, $Y = [H_i^I, H_j^C, H_j^I]$ serves as the input tokens, where $M = 2hw + N$. The output tokens corresponding to synthesized image $\hat{X}_j$ are expressed by $\{U_{j,k} \in \mathbb{R}^{|O|}\}$, which represents the probability distribution of the k-th token over the entries in the codebook.

The training objective is defined via the cross-entropy loss, formulated as

$$\mathcal{L} = \sum_{k=1}^{hw} \texttt{CE}(U_{j,k}, I_{j,k}), \tag{5}$$

where CE denotes the cross-entropy loss, and $I_{j,k}$ is the ground-truth codebook index of the k-th token of image X_j.

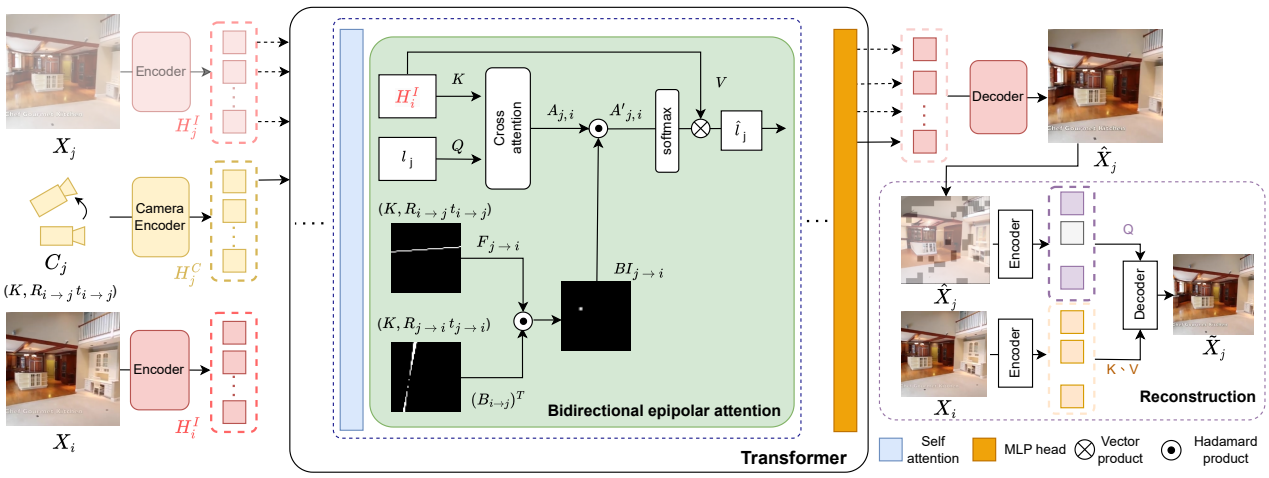

Fig. 2: **Pipeline of our method.** The images and camera pose are expressed as tokens via the image and camera encoders, which are fed into a transformer for autoregressive target view synthesis. In the transformer, a bidirectional epipolar attention layer is included after each self-attention layer, leveraging forward and backward epipolar maps to associate the source and target views. After the image is synthesized by the decoder, epipolar-based reconstruction identifies its poorly generated regions and performs reconstruction through masking and recovery.

Image Decoding. To decode image $\hat{X}_j$ from its predicted tokens, given $\hat{I}_j \in \mathbb{R}^{hw}$, we replace the indices with the corresponding entries from the codebook, namely

$$b_{j,k} = O[\hat{I}_{j,k}] \in \mathbb{R}^{1\times d_o} \quad \text{and} \quad b_j \in \mathbb{R}^{hw\times d_o}. \tag{6}$$

A pre-trained decoder D is employed to decode the image

$$\hat{X}_j = D(b_j). \tag{7}$$

Reconstruction. After generating $\hat{X}_j$, we perform further refinement on this image by applying a masking process to detect its poorly generated regions followed by reconstructing them.

3.2. Bidirectional Epipolar Attention

To improve image consistency, we introduce a Bidirectional Epipolar Attention layer, interleaved with self-attention layers in the transformer. This layer leverages bidirectional epipolar constraints to narrow the attention search space, guiding the target image to focus on corresponding regions in the source image. The process involves three steps: forward epipolar mapping, backward epipolar mapping, and their combination to yield the bidirectional epipolar map.

Forward Epipolar Mapping. Given pose $C_j = (\mathbf{K}, \mathbf{R}_{i\to j}, \mathbf{t}_{i\to j})$, the epipolar line on the source image X_i is computed for each pixel in the target image X_j. For every epipolar line, we generate a weighted map of size $h \times w$, where pixels closer to the epipolar line receive higher values. This computation results in the forward epipolar map from target image X_j to source image X_i:

$$F_{j\to i} \in \mathbb{R}^{hw\times hw}. \tag{8}$$

Backward Epipolar Mapping. Using the inverse camera pose $C'_j = (\mathbf{K}, \mathbf{R}_{j\to i}, \mathbf{t}_{j\to i})$, the epipolar line on the target image for each pixel in the source image can be computed, yielding a backward epipolar map:

$$B_{i\to j} \in \mathbb{R}^{hw\times hw}. \tag{9}$$

$B_{i\to j}$ represents the correspondence for each pixel in the source image, showing which pixels in the target image are most related. Since correspondences are bidirectional, that is, if a pixel p in the source image corresponds well to a pixel q in the target image, q should also correspond well to p. Thereby, we transpose $B_{i\to j}$ to obtain $B_{j\to i}$, aligning it with the representation of the forward epipolar map:

$$B_{j\to i} = (B_{i\to j})^T. \tag{10}$$

Bidirectional Epipolar Map. The bidirectional epipolar map is calculated by combining the forward and backward maps using a Hadamard product, i.e.,

$$BI_{j\to i} = F_{j\to i} \odot B_{j\to i}. \tag{11}$$

This operation highlights pixels with high correspondence in both $F_{j\to i}$ and $B_{j\to i}$, ensuring that their correspondence in $BI_{j\to i}$ remains high. Conversely, if the correspondence of a pixel is low in either $F_{j\to i}$ or $B_{j\to i}$, the resulting correspondence in $BI_{j\to i}$ is suppressed due to the element-wise multiplication.

Cross-Attention Update. Given the output token features $l_j \in \mathbb{R}^{M\times d}$ from the previous self-attention layer and the source image embedding $H_i^I \in \mathbb{R}^{hw\times d}$, l_j serves as the queries, while H_i^I acts as the key-value pairs. Cross-attention is used to compute the affinity matrix $A_{j,i} \in \mathbb{R}^{M\times hw}$. The regions in the affinity matrix that correspond to the target and source images are re-weighted using the bidirectional epipolar map:

$$A'_{j,i}[hw+N:,:] = A_{j,i}[hw+N:,:] \odot BI_{j\to i}[:M-hw-N,:]. \tag{12}$$

By re-weighting with the bidirectional epipolar map, the correspondences within the bidirectional epipolar regions are strengthened, while those outside the regions are suppressed. This helps the target image focus more on the corresponding regions in the source image, effectively improving the alignment between the two images. The updated output $\hat{l}_j$ is then computed as:

$$\hat{l}_j = \text{softmax}(A'_{j,i}) \cdot V, \tag{13}$$

where V is the value matrix computed from the H_i^I. The updated l'_j is subsequently passed as input to the next self-attention layer.

Notably, although the epipolar line guides attention to relevant regions, we still compute attention weights across the entire spatial domain to obtain the epipolar maps.

Source Ground Truth GeoGPT [10] LoR [9] PhotoNVS [12] Ours

Fig. 3: **Qualitative results.** We choose the fifth image from the input video as the ground-truth target for comparison between our method and previous state-of-the-art approaches to NVS on the RealEstate10K dataset.

3.3. Epipolar-based Reconstruction

Although the bidirectional epipolar attention effectively improves the consistency of the generated images, we observe that some unexpected artifacts still appear or novel content is created in the overlapping regions between the two images. To address these issues, we propose epipolar-based reconstruction to perform post-processing on the generated images to restore the problematic regions.

We adopt the reconstruction strategy of Siamese Masked Autoencoders (SMAE) [29]. Given the source image X_i and the predicted image $\hat{X}_j$, both are first encoded into tokens in the latent space using a siamese encoder[30]. For $\hat{X}_j$, we mask a certain proportion of its tokens and leverage the information from the source image X_i to assist in reconstructing a refined image $\tilde{X}_j$. Unlike SMAE, which applies random masking to the target image, we strategically select specific tokens for masking based on the results of the epipolar attention.

The main intuition is that if a particular token attends correctly to the bidirectional epipolar regions during cross-attention, it is more likely to have captured accurate information, leading to better results. In contrast, tokens that fail to attend correctly to the epipolar regions are more likely to have aggregated incorrect information, which could result in errors. These tokens are the ones we need to mask and reconstruct.

Specifically, for a given token k, we evaluate its affinity matrix $A^k_{j,i} \in \mathbb{R}^{h\times w}$ and its corresponding bidirectional epipolar map $BI^k_{j\to i} \in \mathbb{R}^{h\times w}$. We define a mask $M^k \in \mathbb{R}^{h\times w}$ based on a threshold T, a hyperparameter,

$$M^k(\hat{j},\hat{i}) = \begin{cases} 1, & \text{if } BI^k_{j\to i}(\hat{j},\hat{i}) > T, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Here, M^k indicates the regions covered by the bidirectional epipolar constraints.

Next, we compute the average attention within the bidirectional epipolar region, defined as:

$$\text{MEAN}^k_{BI} = \sum (A^k_{j,i} \odot M^k) / \sum M^k, \tag{15}$$

and compare it to the average attention across the attention map:

$$\text{MEAN}^k_{total} = \sum A^k_{j,i} / (h \times w). \tag{16}$$

From these results, we calculate a ratio for token k:

$$\text{Ratio}^k = \text{MEAN}^k_{BI} / \text{MEAN}^k_{total}, \tag{17}$$

which represents the proportion of attention focused on the correct regions for that token.

Using the computed ratios, we rank the tokens in ascending order and select the bottom $s\%$ of tokens for masking and reconstruction, where s is a hyperparameter. It ensures that we can target the tokens most likely to produce suboptimal results, reconstructing them to restore the correct outputs. It is worth noting that, we do not explicitly exclude regions without any correspondences in the input image; such regions may still be selected for reconstruction.

Implementation Details. Our transformer model consists of 32 attention layers. We follow the training details of LoR, including batch size, learning rate, epochs, and optimizer. The input image size is 256×256, encoded into 16×16 tokens. The pre-trained codebook contains 16,384 entries, and the number of camera parameters is $N = 30$. During reconstruction, we set $T = 0.1$ and $s = 80$.

4. EXPERIMENTS

4.1. Experimental Settings

Dataset. We evaluate our method on the RealEstate10K [31] dataset. The original camera motion in its videos is minimal, making it challenging to observe sufficient variations between images. To address this, we follow previous work [32] by sampling the videos at 4 fps to ensure adequate camera motion. We also perform center cropping and resizing of the videos to a spatial resolution of 256×256, consistent with prior approaches [9]. In our experiments, we use 10,000 videos for training and 250 videos for evaluation.

Competing Methods. We compare our approach with three state-of-the-art novel view synthesis methods: GeoGPT [10], LoR [9], and PhotoNVS [12]. GeoGPT provides both explicit and implicit ways for synthesizing novel images. To align with the experimental setup of our study, we utilize their implicit version without utilizing depth information. Our approach shares some similarities with LoR, as both employ a transformer-based GPT-like architecture [28] for novel image synthesis. Unlike LoR, our method leverages bidirectional epipolar attention to more effectively ensure consistency between images. PhotoNVS is a conditional diffusion-based model.

Method	PSNR↑	SSIM↑	LPIPS↓
GeoGPT [10]	15.06	0.49	2.58
LoR [9]	15.53	0.49	2.77
PhotoNVS [12]	17.61	0.59	2.37
Ours	**17.88**	**0.59**	**2.32**

Table 1: **Quantitative evaluation on novel view synthesis.** We report the performnace in PSNR, SSIM, and LPIPS computed between the first five generated and ground-truth frames in the videos.

Method	PSNR↑	SSIM↑	LPIPS↓
Forward Epipolar Attention	16.05	0.52	2.52
Bidirectional	16.18	0.52	2.47
Bidirectional + Reconstruction	17.88	0.59	2.32

Table 2: **Ablation quantitative results.** We present the results in the three metrics with or without using the proposed "bidirectional" epipolar attention and epipolar-based "reconstruction."

Evaluation Metrics. Following previous methods [9, 10, 12], we adopt three common metrics to evaluate the generated images, including 1) Peak Signal-to-Noise Ratio (PSNR), which measures the ratio of the maximum signal to noise between the generated and original images; 2) Structural Similarity Index (SSIM) [14], which evaluates the similarity of structure, luminance, and contrast between images; and 3) Learned Perceptual Image Patch Similarity (LPIPS) [15], a deep learning-based perceptual similarity measure that assesses the visual similarity between generated images and the ground truth. These metrics provide a comprehensive evaluation of the generated image qualities.

4.2. Comparison with SoTA Methods

Quantitative Results. For the quantitative evaluation, we randomly sample testing videos for assessment. Since image extrapolation continuously generates new scenes that are likely to differ from the ground truth, PSNR, SSIM, and LPIPS may not be suitable for long-distance frames. Therefore, for each sampled testing video, we compute these metrics only for the five frames following the input image. As shown in Table 1, we compare our method with three competing methods on the RealEstate10K dataset, and the results demonstrate that our method, equipped with bidirectional epipolar attention and epipolar-based reconstruction, achieves better performance across three different metrics.

Qualitative Results. We present a comparison of our method with other SoTA qualitative results. As shown in Fig. 3, within the blue box highlighting the pillow, our method achieves results that are closest to the ground truth, maintaining a high level of consistency. In contrast, other methods generate results that deviate from the ground truth, failing to produce consistent outputs. In the region marked by the red box, GeoGPT generates some green content that does not exist in the original image. LoR and PhotoNVS produce outputs that are inconsistent with the ground truth in this area. In comparison, our method demonstrates significantly better performance, with more accurate and consistent results.

4.3. Ablation Study

To evaluate the impact of bidirectional epipolar attention and epipolar-based reconstruction, we conduct an ablation study, with quantitative results presented in Table 2 and qualitative results shown in Fig. 4. First, we compare two approaches: one using forward epipolar attention for correspondence calculation and the other utilizing bidirectional epipolar attention. As highlighted by the red boxes in the "forward" and "bidirectional" columns of Fig. 4, forward epipolar attention introduces geometric distortions, while bidirectional epipolar attention generates results that are more geometrically consistent. This demonstrates the advantage of bidirectional epipolar attention in maintaining geometric integrity.

We integrate epipolar-based reconstruction into the bidirectional epipolar framework. As shown in the red-highlighted regions of the "bidirectional" and "bidirectional+Recon" columns of Fig. 4, while bidirectional epipolar attention mitigates geometric distortions, it still introduces some unexpected textures different from the ground truth. With epipolar-based reconstruction, the final results achieve greater alignment with the ground truth, highlighting its effectiveness in enhancing texture accuracy and overall visual fidelity.

Additionally, we measure the inference time for generating 256×256 images using an NVIDIA RTX 3090 GPU and analyze how different components affect the overall generation time. Detailed results are shown in Table 3.

Method	Execution Time(s)	Change(s)
Baseline	13.64	-
+ Forward epipolar	15.68	2.04
+ Bidirectional epipolar	16.62	0.94
+ Reconstruction	18.63	2.01

Table 3: **Time Analysis.** As different components are added, the image generation time increases accordingly.

Fig. 4: **Ablation qualitative results.** We display the synthesized results by our method with or without using the proposed "bidirectional" epipolar attention and epipolar-based "reconstruction."

5. CONCLUSION

In this work, we address the challenge of novel view synthesis from a single image. We propose a novel bidirectional epipolar attention framework coupled with an epipolar-based reconstruction method to ensure consistency in the generated images. Additionally, unfavorable artifacts and novel content in the overlapping regions between images are substantially alleviated. Both quantitative and qualitative results demonstrate that our approach surpasses existing state-of-the-art methods, validating the effectiveness of our proposed method.

Acknowledgment. This work was supported in part by the National Science and Technology Council (NSTC) under grants 112-2221-E-A49-090-MY3, 111-2628-E-A49-025-MY3, and 113-2634-F-002-003, and Academia Sinica under the grant number of AS-CDA-110-M09.

6. REFERENCES

[1] Namrata Singh and Sarvpal Singh, "Virtual reality: A brief survey," in *ICICES*, 2017. 1

[2] Hardeep Singh, Kamaljeet Kaur, and Preet Pinder Singh, "Artificial intelligence as a facilitator for film production process," in *AISC*, 2023. 1

[3] W.T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, "Computer vision for computer games," in *FG*, 1996. 1

[4] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa, "Infinitenature-zero: Learning perpetual view generation of natural scenes from single images," in *ECCV*, 2022. 1, 2

[5] 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 *ICCV*, 2021. 1, 2

[6] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al., "Wonderjourney: Going from anywhere to everywhere," in *CVPR*, 2024. 1

[7] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu, "Wonderworld: Interactive 3d scene generation from a single image," *arXiv preprint arXiv:2406.09394*, 2024. 1

[8] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *TPAMI*, 2022. 1

[9] Xuanchi Ren and Xiaolong Wang, "Look outside the room: Synthesizing a consistent long-term 3D scene video from a single image," in *CVPR*, 2022. 1, 2, 4, 5

[10] Robin Rombach, Patrick Esser, and Björn Ommer, "Geometry-free view synthesis: Transformers and no 3D priors," in *ICCV*, 2021. 1, 2, 4, 5

[11] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf, "Consistent view synthesis with pose-guided diffusion models," in *CVPR*, 2023. 1, 2

[12] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker, "Long-term photometric consistent novel view synthesis with diffusion models," in *ICCV*, 2023. 1, 2, 4, 5

[13] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al., "Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion," in *CVPR*, 2024. 1

[14] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, 2004. 2, 5

[15] 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. 2, 5

[16] S. Avidan and A. Shashua, "Novel view synthesis in tensor space," in *CVPR*, 1997. 2

[17] S. Laveau and O.D. Faugeras, "3-d scene representation as a collection of images," in *ICPR*, 1994. 2

[18] Shenchang Eric Chen and Lance Williams, "View interpolation for image synthesis," in *SIGGRAPH*, 1993. 2

[19] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *CVPR*, 2022. 2

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *CACM*, 2021. 2

[21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *TOG*, 2022. 2

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3D Gaussian splatting for real-time radiance field rendering.," *TOG*, 2023. 2

[23] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann, "pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction," in *CVPR*, 2024. 2

[24] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel, "Scenescape: Text-driven consistent scene generation," *NeurIPS*, 2024. 2

[25] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner, "Text2room: Extracting textured 3D meshes from 2D text-to-image models," in *ICCV*, 2023. 2

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022. 2

[27] Patrick Esser, Robin Rombach, and Bjorn Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021. 2

[28] Alec Radford and Karthik Narasimhan, "Improving language understanding by generative pre-training," 2018. 2, 4

[29] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li, "Siamese masked autoencoders," *NeurIPS*, 2023. 4

[30] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, "Signature verification using a "siamese" time delay neural network," in *NeurIPS*, 1993. 4

[31] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *TOG*, 2018. 4

[32] Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang, "Video autoencoder: self-supervised disentanglement of 3D structure and motion," in *ICCV*, 2021. 4