360FusionNeRF: Panoramic Neural Radiance Fields with Joint Guidance

Shreyas Kulkarni¹, Peng Yin²,∗, and Sebastian Scherer²

Abstract—Based on the neural radiance fields (NeRF), we present a pipeline for generating novel views from a single 360° panoramic image. Prior research relied on the neighborhood interpolation capability of multi-layer perceptions to complete missing regions caused by occlusion. This resulted in artifacts in their predictions. We propose 360FusionNeRF, a semi-supervised learning framework that employs geometric supervision and semantic consistency to guide the progressive training process. Firstly, the input image is reprojected to 360° images, and depth maps are extracted at different camera positions. In addition to the NeRF color guidance, the depth supervision enhances the geometry of the synthesized views. Furthermore, we include a semantic consistency loss that encourages realistic renderings of novel views. We extract these semantic features using a pre-trained visual encoder CLIP, a Vision Transformer (ViT) trained on hundreds of millions of diverse 2D photographs mined from the web with natural language supervision. Experiments indicate that our proposed method is capable of producing realistic completions of unobserved regions while preserving the features of the scene. 360FusionNeRF consistently delivers state-of-the-art performance when transferring to synthetic Structured3D dataset (PSNR ~ 5%, SSIM ~ 3% LPIPS ~ 13%), real-world Matterport3D dataset (PSNR ~ 3%, SSIM ~ 3% LPIPS ~ 9%) and Replica360 dataset (PSNR ~ 8%, SSIM ~ 2% LPIPS ~ 18%). We provide the source code at https://github.com/MetaSLAM/360FusionNeRF.

Index Terms—Scene representation, View synthesis, Neural Radiance Field, 360° image, 3D deep learning

I. INTRODUCTION

Omnidirectional cameras have become more easily accessible, with a rising number of panoramas shared on media and 360° datasets made available. 360° cameras can capture complete environments in a single shot making 360° imaging desirable for several computer vision applications. In the computer vision community, they are becoming increasingly prevalent. The omnidirectional 360° field-of-view captured by these devices is appealing for tasks such as robust, omnidirectional SLAM [1], [2], scene understanding and layout estimation [3], [4], or VR photography and video [5], [6]. While many strategies are proposed to synthesize novel views by taking the perspective images as the input, prior work rarely considers the panorama image as a single source for modeling and rendering. Although perspective images can be acquired conveniently, constructing a complete scene needs many dense samples. Furthermore, additional camera variables are required for determining relative poses and matching [7].

Synthesizing novel views with parallax provides immersive 3D experiences [8]. Traditional computer vision solutions involve reconstruction techniques like structure from motion [9] and image-based rendering [10] using a set of densely captured images. However, these approaches suffer from the cost of matching and reconstruction computation for both time and capacity. The recent development in this field focuses on deep learning approaches for its strong capability of modeling 3D geometry and rendering novel views [7]. Neural network-based rendering approaches have been rapidly developed in recent years, and the neural radiance field (NeRF) [11] is a promising method for simulating photorealistic views. However, NeRF needs hundreds of images with their known relative poses as input. They should be captured under same shooting settings, which is a time-consuming and labor-intensive process [12]. Accordingly, various attempts have been made to minimize the number of input images [13], [14] or ease the shooting conditions [15], [16].

We attempt to construct a 3D representation of a scene from a single 360° photograph. It is useful to learn NeRF from a single 360° picture since we do not need to align the shooting circumstances across shots. In addition, we do not need to know the relative locations of photos since we simply need a single image that contains a wealth of omnidirectional information [12]. OmniNeRF [7] is a prior study of this approach; however, it depends only on the neighborhood interpolation capacity of the multi-layer perceptron to complete the missing regions caused by occlusion. The single source image does not contain sufficient information to infer the occlusion and the opposite side of objects [12]; thus, the results are deteriorated. [12] attempts to fill in the missing regions of the reprojected images by utilizing a self-supervised generative model. However, the model fails when reprojected images have large missing areas.

Only a few NeRF approaches have been developed to use depth measurements concurrently with color in the volumetric rendering pipeline [17], [18]. In this paper, we investigate depth as an additional, inexpensive form of supervision for guiding the geometry learned by OmniNeRF. We propose concurrently extracting depth information while the input image is reprojected at different camera positions. It was noticed that depth information may significantly enhance geometry compared to color information alone.

OmniNeRF is still estimated per scene and does not benefit

¹S. Kulkarni is with the Department of Mechanical Engineering, Indian Institute of Technology, Madras (shreyas.kulkarni@iitm.ac.in).
²P. Yin and S. Scherer are with Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA (pyin2, basi@andrew.cmu.edu).
from prior knowledge from other images and objects. Prior knowledge is required when a scene reconstruction problem is underdetermined. 3D reconstruction methods struggle when regions of an object are never observed. This is especially challenging when rendering an object at significantly different poses. During training, unobserved regions become visible when rendering a scene with a significant baseline shift. A view synthesis system must produce plausible missing details to fill in the gaps. Due to lack of prior knowledge, even a regularized NeRF is incapable of learning accurate extrapolations to unseen regions.

We also exploit the consistency principle approach employed by DietNeRF [14]: objects share high-level semantic properties across their views. Image recognition models learn to extract several such high-level semantic features, including object identity. We transfer prior knowledge from image encoders that have been trained on extremely varied 2D single-view image data to tackle the view synthesis problem. In the single-view setting, such encoders are typically trained on millions of realistic images such as ImageNet [19]. CLIP is a recent multi-modal encoder trained to match pictures with captions in a massive web scrape containing 400M images [20]. Due to the diversity of its data, CLIP demonstrated promising zero and few-shot transfer performance on image recognition tasks. These models also contain prior knowledge, which is useful for novel view synthesis.

Our contributions in this paper are as follows:

- We propose 360FusionNeRF, a neural scene representation framework based on OmniNeRF that can be estimated from a single RGB-D 360° Panoramic Image and can generate views with unobserved regions.
- In addition to minimizing NeRF’s mean squared error losses at known poses in pixel space, 360FusionNeRF penalizes a geometric loss via Auxiliary depth of the projected images and a self-supervised semantic consistency loss through activations of CLIP’s ViT.
- We demonstrate qualitatively and quantitatively that our proposed method results in a generalizable scene representation and enhances the perceptual quality.

II. RELATED WORK

A. Neural 3D Rendering

Neural Radiance Fields (NeRFs) [11] have shown promising progress in view synthesis by learning an implicit neural scene representation. Since its origin, significant efforts have been made to enhance its quality [21], [22], speed [23], [24], artistic effects [25], [26], and generalization ability [27], [28]. In particular, Mip-NeRF [29] proposes to cast a conical frustum rather than a single ray for anti-aliasing. Map-NeRF 360 [30] further extends this to unbounded scenes through efficient parameterization. DS-NeRF [17] employs extra depth supervision to enhance the quality of reconstruction. RegNeRF [34] proposes a normalizing flow and depth smoothness regularization. DietNeRF [14] uses the CLIP embeddings to impose semantic constraints on unseen views. PixelNeRF [13] employs a ConvNets encoder to extract context information via large-scale pre-training and renders novel views from a single input successfully. However, it can only work on simple objects (e.g., ShapeNet) [31], while the results on complex scenes are uncertain. In addition, the method depends on the availability of the entire reference image for supervision, but the reprojected images are incomplete in a panoramic setting. SinNeRF [31] also proposes a multi-supervision NeRF, but its approaches are quite object-centric. Our study focuses on complex scene reconstruction with panoramic images.

B. 360 Panorama View Synthesis

OmniNeRF [32] generates novel fish-eye projection images by using spherical sampling to enhance the quality of the

---

**Fig. 1**: Visualization of novel view rendering in the proposed method on a sample of Structured3D. A plausible viewpoint image with 3D consistency is synthesized at positions different from the camera position of the input.
results. 360Roam [33] is a scene-level NeRF system that can synthesize images of large-scale indoor scenes in real time and it supports VR roaming. PanoHDR-NeRF [34] proposes a pipeline for predicting the full HDR radiance of an indoor scene without using specialized hardware, meticulous scanning of the scene, or finely calibrated camera configurations. This paper, however, focuses on synthesizing novel views from a single Equiangular Panorama 360° RGB-D image.

OmnNeRF [7] learns a complete scene from a single 360° RGB image, without the requirement to define relative positions or shooting conditions. However, it depends only on the neighborhood interpolation capability of the multi-layer perceptron to fill the missing regions caused by zooming and occlusion which results in artifacts, and the image quality is highly reduced when moving away from the camera position of the input image [12]. An alternative method to NeRF, Pathdreamer [35], generates novel views from a single 360° RGB-D image. However, it has a problem with low 3D consistency in the synthesized views due to its dependence on 2D image-to-image translation.

In [12], a self-supervised trained generative model fills the missing regions of the reprojected images of OmniNeRF, and the completed images are used to train the NeRF. They introduce a technique for training NeRF while dynamically selecting a sparse collection of completed images to lower the discrimination error of the synthesized views with real images. However, it fails to generate plausible views when there are huge missing regions that exceed the image completion capabilities.

III. BACKGROUND

A. Preliminaries

Neural Radiance Fields (NeRFs) [11] synthesize images by sampling from 5D coordinates (location \(\mathbf{x}, \mathbf{y}, \mathbf{z}\)) and viewing direction \(\mathbf{\theta}, \mathbf{\phi}\) along camera rays, and mapping them to color \(\mathbf{r}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{\theta}, \mathbf{\phi})\) and volume density \(\sigma(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{\theta}, \mathbf{\phi})\) [11] first proposed the same using coordinate-based multi-layer perception networks to parameterize this function and use volumetric rendering techniques to compose the values at each location to obtain the final rendered images.

OmnNeRF [7] creates multiple images at virtual camera positions from a single 360° RGB-D image and uses these images to train NeRF. The given RGB-D panorama is used to produce a set of 3D points, which are then reprojected into multiple omnidirectional images corresponding to different virtual camera locations. The resulting omnidirectional images are likely to be flawed, since there may be gaps and cracks between pixels due to occlusion or limited resolution. OmnNeRF addresses this issue by using the pixel-based prediction property of its MLP model, which takes a single pixel as input rather than an entire image.

The ray is represented by \(\mathbf{r}(t) = \mathbf{d} + t\mathbf{d}\) where \(\mathbf{d}\) and \(\mathbf{d}\) denote the ray origin and ray direction. At each discrete sample on the ray, the final RGB values \(C(r)\) are optimized by aggregation of color \(c_i\) and opacity \(\sigma_i\). A positional encoding technique [36] is applied to the rays for capturing high frequency information. The function of the color composition follows the following rule in volume rendering [37]:

\[
\hat{C}(r) = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) c_i, \tag{1}
\]

where \(T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)\) and \(\delta_i = t_{i+1} - t_i\) is the interval between two adjacent samples. The overall volume sampling principles are done hierarchically with a coarse and a refined stage. The coarse and refined networks are identical, with the exception of the process of sampling pixels on a ray. At the coarse stage, \(N_c\) intervals are uniformly sampled along the ray, while at the refined stage, \(N_f\) intervals are decided according to the densities from the coarse stage. These predictions are optimized by the ground-truth color. It optimizes the radiance field by minimizing the mean squared error between the ground truth color and the rendered color,

\[
\mathcal{L}_{\text{Color}} = \sum_{\mathbf{r} \in R_i} ||(C(r) - \hat{C}(r))||^2 \tag{2}
\]

where \(R_i\) is the set of input rays during training.

B. Challenges with OmnNeRF

1) Overfitting to Training Views: Conceptually, OmnNeRF is trained by simulating the image-formation process at observed poses. With several training views, the MLP in OmnNeRF is able to recover the accurate textures and occupancy that allow interpolations to new views. The high-frequency representational capacity enables OmnNeRF to overfit to each input view. Fundamentally, the plenoptic function representation suffers from near-field ambiguity [38] where distant cameras each observe significant regions of space that no other camera observes. In this situation, the ideal scene representation is underdetermined. Additionally, degenerate solutions may exploit the view dependence of the radiance field. [14] pointed out that although a rendered view from a pose near a training image has reasonable textures, it is skewed incorrectly and contains cloudy artifacts from incorrect geometry. As the geometry is not estimated correctly, a distant view has nearly no correct information. High-opacity regions impede the camera. Without supervision from a nearby camera, opacity is susceptible to random initialization.

2) No Generalization to Unseen Views: As OmnNeRF estimates each scene from scratch, it has no prior knowledge of natural objects, such as common symmetries and object parts. NeRF gets no supervisory signal from \(\mathcal{L}_{\text{Color}}\) to the unobserved regions and instead depends on the inductive bias of the MLP for any inpainting. We want to provide prior knowledge that enables NeRF to utilize bilateral symmetry for plausible completions.

IV. PROPOSED METHOD

A. Geometric Supervision

Directly overfitting the reference images results in a corrupted neural radiance field collapsing towards the provided
views. We begin by adopting the depth prior to reconstruct reasonable 3D geometry.

1) Extracting Depth Ground Truths: We adhere to the projection procedure from OmniNeRF [7] to concurrently derive depth information from the view. First, all pixels are projected onto a uniform sphere using their 2D coordinates. For a pixel \((x, y)\) on the panorama, its vertical and horizontal viewing angles are determined by \(\theta = \pi y/H, \phi = 2\pi x/W\), where \(H\) and \(W\) are the height and width of the panorama, respectively. The coordinate center would correspond to the current camera position, namely the ray origin. Similarly, the direction of the ray is a unit vector from the center to the sphere. Consequently, a novel panoramic view and the depth information may be derived by repositioning the camera and analyzing what would be sampled on the new sphere by the emitted rays, using the preceding equations. Not all pixels are supposed to be visible from the new viewpoint. The key of the projection mechanism is to verify which parts of the ground truth are visible to specific ray origin.

2) Volumetric Rendering: Similar to color rendering, the depth can be represented with volume density using:

\[
\hat{D}(r) = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) t_i, \tag{3}
\]

where \(T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)\) and \(\delta_i = t_{i+1} - t_i\) is the interval between two consecutive samples.

3) Optimization: The network parameters are tuned using a collection of RGB-D frames, including a color, depth, and camera pose data. \(L_{\text{Color}}\) in Equation 2 represents the photometric loss. The geometric loss is the absolute difference between predicted and actual depths, normalized by the depth variance to discourage the use of weights with a high degree of uncertainty. The geometric loss is given by:

\[
L_{\text{Geo}} = \sum_{r \in R} \frac{|\hat{D}(r) - D(r)|}{\sqrt{\hat{D}_{\text{var}}(r)}}, \tag{4}
\]

where \(\hat{D}_{\text{var}}(r) = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i))(\hat{D}(r) - t_i)^2\) depth variance of the image.

B. Semantic Consistency

In contrast to the geometry pseudo labels, which ensures consistency in 3D space, pseudo semantic labels are used to regularize the 2D image fidelity. We present a global structure backed by a pre-trained ViT network. This guidance enables SinNeRF to render visually-pleasing results for each view.

Vision transformers (ViT) have been shown to be expressive semantic priors, even amongst misaligned pictures [39]. Similar to [31], we suggest using a pre-trained ViT for global structural guidance that maintains semantic consistency across unseen views. Although there is pixel-wise misalignment between the views, we agree with [31] that the extracted representation of ViTs is robust to this misalignment and offers semantic supervision. The content and style of the two views are similar, and a deep neural network is capable of learning an invariant representation.

Here, we adopt CLIP-ViT [20], a self-supervised vision transformer trained on ImageNet dataset. CLIP produces normalized image embeddings. When the embedding is a vector, the \(L_{\text{SC}}\) simplifies to cosine similarity up to a constant and a scaling factor that may be absorbed into the loss weight \(\lambda\):

\[
L_{\text{SC}}(I_1, I_2) = \lambda \phi(I_1)^T \phi(I_2), \tag{5}
\]
where \( \phi(.) \) is the normalized image embedding and \( I_1, I_2 \) are unseen views. The unseen views are chosen at random.

C. Final Pipeline

For generation of \( L_{SC} \) requires volume rendering, which is computationally intensive. Hence, we calculate the semantic consistency over a smaller resolution of views. The \( L_{SC} \) converges more quickly than \( L_{Color} \) and \( L_{Geo} \). Hence, we minimize \( L_{SC} \) only once for every \( k \) iterations of minimization of \( L_{Color} \) and \( L_{Geo} \). Finally the loss can be represented as:

\[
L = L_{Color} + \lambda_{Geo}L_{Geo} + \lambda_{SC}L_{SC},
\]

where \( \lambda_{SC} \) and \( \lambda_{Geo} \) are scaling factors to decide the importance given to every loss. They can be tuned experimentally.

V. EXPERIMENTS

A. Dataset

Our method is evaluated on both synthetic and real-world datasets. All the panorama images in this work are projected to equirectangular geometry with a resolution of 512 \( \times \) 1024 for Structured3D dataset, 1024 \( \times \) 2048 for Replica360 and Matterport3D datasets. We randomly choose 3 scenes from each dataset and analyze them quantitatively and qualitatively.

1) Structured3D: Structured3D dataset [40] contains 3,500 synthetic departments (scenes) with 185,985 photorealistic panoramic renderings. As the original virtual environment is not publicly accessible, we utilized the rendered panoramas directly.

2) Matterport3D: Matterport3D dataset [41] is a large-scale indoor real-world 360\(^\circ\) dataset captured by Matterport’s Pro 3D camera in 90 furnished houses (scenes). The dataset contains 10,800 RGB-D panorama images, where the RGB-D signals near the polar region are missing.

3) Replica360: Replica360 dataset [42] contains 18 highly photorealistic 3D indoor scene reconstructions at room and building scale.

B. Implementation Details

The Adam optimizer is used throughout the training process. The learning rate is initialized to 5.10\(^{-4}\) and exponentially reduced to 5.10\(^{-5}\). We train the model for 200,000 epochs for each experiment with a batch size of 1,400 on a DGX A100 GPU. We set \( N_c = 64 \) and \( N_f = 128 \) in the coarse and refined networks. The network architecture is identical to that of OmniNeRF [7].

C. Comparison with other methods

The community lacks open-source codes that have been implemented for 360 Panoramas. Similar work like PanoHDR-NeRF [34] and [12] are not open source. There are major differences in the problem statement considered with certain proposed methods. Pathdreamer [35] uses input data containing a trajectory of RGB, Depth and Semantic maps. Our method does not use Semantic maps and performs with only a single 360-degree panorama input, contrary to Pathdreamer. We have made an active effort to compare works in similar settings, and have been motivated by the lack of prior research on novel view synthesis for 360-degree methodologies. In this work, we build upon OmniNeRF and discuss the improvements with OmniNeRF as baseline.

D. Qualitative Evaluation

We qualitatively validate the novel view synthesis using a single 360\(^\circ\) RGB-D image. Figure 3 compares the synthesized novel views by the proposed method to OmniNeRF. Each column refers to a scene in a dataset, and each row contains the results of a method. It is evident that our method preserves the optimal geometry and perceptual quality.

In the Replica360 sample, we can see the vase has been blurred for predictions by OmniNeRF while its shape has been well recovered effectively in our method. Some artifacts are formed on the walls, albeit they have been drastically reduced. Even though the chair is tiny, our method has produced the shape well.

Matterport3D dataset has panoramas blurred at the poles. This makes it challenging for OmniNeRF to replicate an object near the ceilings or the floor. As observed, the backdrop ceiling light has become blurry, yet our approach can still identify it. Even the face of the idol and the cups are considerably more distinct and have not lost their form as compared to OmniNeRF.

Regarding synthesis for the Structured3D sample, the texture of the walls has been lost in the case of OmniNeRF, while our method has preserved it nicely. In the vicinity of the bookcase, the artifacts are prohibited, and even clear things, like as the glass cup, have not collapsed, as was the case with OmniNeRF.

E. Quantitative Evaluation

We quantitatively analyzed each method using the three evaluation criteria detailed below:

1) PSNR: Peak-to-signal-noise ratio expresses the mean-squared error in logarithmic space. This metric evaluates the performance of the input image reconstruction. We compute the PSNR between the reference image and the synthesized image at the position of the input image.

2) SSIM: Structural Similarity Index Measure [43] assesses the loss of image quality in the reconstructed image; the greater the value, the better. However, it often contradicts human assessments of similarity [44].

3) LPIPS: Deep CNN activations resemble characteristics of the human perception. We measure the perceptual image quality using LPIPS [44], which computes MSE between normalized features from all layers of a pre-trained VGG encoder.

We extract three scenes each from the Structure3D, Matterport3D and Replica360 datasets. Table I presents the evaluation results for each image. In almost all scenes, the proposed method outperforms OmniNeRF in terms of the PSNR, SSIM and LPIPS, indicating that the proposed method
Fig. 3: Qualitative comparison of OmniNeRF and the proposed method on Replica360 (R1), Matterport3D (M1), and Structured3D (S1) datasets.

|                     | Replica360 | Matterport3D | Structured3D |
|---------------------|------------|--------------|--------------|
|                     | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| Replica360          |        |        |        |        |        |        |        |        |        |
| OmniNeRF            | 28.83  | 0.9226 | 0.2715 | 30.61  | 0.9374 | 0.3385 | 27.39  | 0.8865 | 0.3701 |
| 360FusionNeRF (ours)| 32.76  | 0.9451 | 0.2116 | 33.23  | 0.9520 | 0.2790 | 27.61  | 0.8951 | 0.3184 |
| Matterport3D        |        |        |        |        |        |        |        |        |        |
| OmniNeRF            | 25.61  | 0.8000 | 0.2994 | 25.01  | 0.8860 | 0.2720 | 19.01  | 0.8481 | 0.2948 |
| 360FusionNeRF (ours)| 25.53  | 0.8336 | 0.2575 | 26.88  | 0.8934 | 0.2573 | 19.13  | 0.8622 | 0.2748 |
| Structured3D        |        |        |        |        |        |        |        |        |        |
| OmniNeRF            | 26.41  | 0.8259 | 0.2860 | 22.29  | 0.8627 | 0.2614 | 29.72  | 0.8903 | 0.2472 |
| 360FusionNeRF (ours)| 28.05  | 0.8734 | 0.2260 | 23.45  | 0.8731 | 0.2570 | 30.20  | 0.9061 | 0.2047 |

TABLE I: Quantitative evaluation of each method on three scenes for the Replica360, Matterport3D and Structured3D datasets on PSNR, SSIM and LPIPS metrics. Higher PSNR and SSIM values with lower LPIPS values are desirable.

can more effectively synthesize plausible views with features similar to the dataset.

The inferior performance of both models in Matterport3D may be attributed to the distortion caused by the blurring of the panoramic images at the poles. In M1, OmniNeRF achieves a higher PSNR score, whereas our method outperforms both SSIM and LPIPS. Because of uncertainty, blurry renderings will outperform sharp but incorrect renderings on average error metrics like MSE and PSNR. Arguably, perceptual quality and sharpness are better metrics than pixel error for graphics applications like photo editing and virtual reality, as plausibility is emphasized.

F. Ablation Study

In this section, we study the effectiveness of geometric and semantic guidance proposed in our model. We evaluate on R2 and present results through Tab. II and Fig. 4. Both guidances partially contribute to the overall enhancement over OmniNeRF. The semantic guidance functions as a supplement to the geometric guidance in order to achieve final outcomes.

VI. CONCLUSIONS

This paper proposes a method for synthesizing novel views by learning the neural radiance field from a single 360° image. The proposed method reprojects the input image onto
TABLE II: Ablation Study on variants of guidance. Here “w/o $\mathcal{L}_{SC}$” refers to the variant without semantic guidance and “w/o $\mathcal{L}_{Geo}$” refers to the variant without geometric guidance. The experiments are performed on R2.

| Methods          | PSNR↑ | SSIM↑ | LPIPS↓ |
|------------------|-------|-------|--------|
| w/o $\mathcal{L}_{SC}$ | 31.56 (-1.67) | 0.9435 (-0.0085) | 0.3025 (+0.0235) |
| w/o $\mathcal{L}_{Geo}$ | 32.54 (-0.69) | 0.9470 (-0.0050) | 0.2936 (+0.0146) |
| Full Model       | 33.23 | 0.9520 | 0.2790 |

360° images at other camera positions and estimates its depth map. A geometric loss and semantic consistency loss were introduced in addition to the Color loss. Experiments demonstrated that the proposed method could synthesize plausible novel views while preserving the features of the scene for both synthetic and real-world scenes. These findings demonstrate the effectiveness of employing geometric and semantic supervision for panoramic novel view synthesis. We intend to expand our work for multi 360 degree views of a scene and apply it to large scale navigation tasks. We plan to also tackle the training efficiency issue with instant neural graphs primitives [45].

VII. ACKNOWLEDGMENT

This research was supported by grants from NVIDIA and utilized NVIDIA SDKs (CUDA Toolkit, TensorRT, and Omniverse).

REFERENCES

[1] C. Won, H. Seok, Z. Cui, M. Pollefeys, and J. Lim, “Omnislam: Omnidirectional localization and dense mapping for wide-baseline multi-camera systems,” in 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 559–566.

[2] S. Sumikura, M. Shibuya, and K. Sakurada, “Openslam: A versatile visual slam framework,” in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2292–2295.

[3] L. Jin, Y. Xu, J. Zheng, I. Zhang, R. Tang, S. Xu, J. Yu, and S. Gao, “Geometric structure based and regularized depth estimation from 360 indoor imagery,” in Proceedings of the IEEE/ECCV Conference on Computer Vision and Pattern Recognition, 2020, pp. 889–898.

[4] C. Sun, M. Sun, and H.-T. Chen, “Hohonet: 360 indoor holistic understanding with latent horizontal features,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2573–2582.

[5] T. Bertel, M. Yuan, R. Lindroos, and C. Richardt, “Omniphotos: Casual 360° vr photography,” ACM Trans. Graph., vol. 39, no. 6, nov 2020. [Online]. Available: https://doi.org/10.1145/3414685.3417770

[6] A. Serrano, I. Kim, Z. Chen, S. DiVerdi, D. Gutierrez, A. Hertzmann, and B. Masia, “Motion parallax for 360 rgbd video,” IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 5, pp. 1817–1827, 2019.

[7] C.-Y. Hsu, C. Sun, and H.-T. Chen, “Moving in a 360 world: Synthesizing panoramic parallaxes from a single panorama,” arXiv preprint arXiv:2106.10859, 2021.

[8] H.-Y. Shum, K.-T. Ng, and S.-C. Chan, “A virtual reality system using the concentric mosaic: construction, rendering, and data compression,” IEEE Transactions on Multimedia, vol. 7, no. 1, pp. 85–95, 2005.

[9] R. Hartley and A. Zisserman, “Camera geometry and single view geometry,” in Multiple View Geometry in Computer Vision. Cambridge Univ. Press, 2004, pp. 151–236.

[10] H.-Y. Shum, S.-C. Chan, and S. B. Kang, “Rendering techniques,” Image-Based Rendering, pp. 71–90, 2007.

[11] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[12] T. Hara and T. Harada, “Enhancement of novel view synthesis using omnidirectional image completion,” arXiv preprint arXiv:2203.09957, vol. 222.

[13] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, “pixelnert: Neural radiance fields from one or few images,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4578–4587.

[14] A. Jain, M. Tancik, and P. Abbeel, “Putting nerf on a diet: Semantically consistent few-shot view synthesis,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5885–5894.

[15] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf in the wild: Neural radiance fields for unconstrained photo collections,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7210–7219.

[16] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Priscaraci, “Nerf++: Neural radiance fields without known camera parameters,” arXiv preprint arXiv:2102.07064, 2021.

[17] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, “Depth-supervised nerf: Fewer views and faster training for free,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12 882–12 891.

[18] T. Neff, P. Stadlbauer, M. Parger, A. Kurz, J. H. Mueller, C. R. A. Chaintay, A. Kaplanian, and M. Steinberger, “Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks,” in Computer Graphics Forum, vol. 40, no. 4. Wiley Online Library, 2021, pp. 45–59.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet:
[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[21] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, and P. P. Srinivasan, “Ref-nerf: Structured view-dependent appearance for neural radiance fields,” *arXiv preprint arXiv:2112.03907*, 2021.

[22] Y.-C. Guo, D. Kang, L. Bao, Y. He, and S.-H. Zhang, “Nerfren: Neural radiance fields with reflections,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 49–59.

[23] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *arXiv preprint arXiv:2201.05989*, 2022.

[24] C. Sun, M. Sun, and H.-T. Chen, “Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5459–5469.

[25] C. Wang, M. Chai, M. He, D. Chen, and J. Liao, “Clip-nerf: Text-and-image driven manipulation of neural radiance fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3835–3844.

[26] Z. Fan, Y. Jiang, P. Wang, X. Gong, D. Xu, and Z. Wang, “Unified implicit neural stylization,” *arXiv preprint arXiv:2204.01943*, 2022.

[27] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, “Ibrnet: Learning multi-view image-based rendering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4690–4699.

[28] Y. Liu, S. Peng, L. Liu, Q. Wang, P. Wang, C. Theobalt, X. Zhou, and W. Wang, “Neural rays for occlusion-aware image-based rendering,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7824–7833.

[29] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, “Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5855–5864.

[30] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, “Mip-nerf 360: Unbounded anti-aliasing neural radiance fields,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5470–5479.

[31] D. Xu, Y. Jiang, P. Wang, Z. Fan, H. Shi, and Z. Wang, “Sinnerf: Training neural radiance fields on complex scenes from a single image,” *arXiv preprint arXiv:2204.00928*, 2022.

[32] K. Gu, T. Maugey, S. Knorr, and C. Guillemot, “Omni-nerf: neural radiance field from 360° image captures,” in *ICME 2022-IEEE international conference on multimedia and expo*, 2022.

[33] H. Huang, Y. Chen, T. Zhang, and S.-K. Yeung, “360roam: Real-time indoor roaming using geometry-aware 360 radiance fields,” *arXiv preprint arXiv:2208.02705*, 2022.

[34] P. Gera, M. R. K. Dastjerdi, C. Renaud, P. Narayanan, and J.-F. Lalonde, “Casual indoor hDr radiance capture from omnidirectional images,” *arXiv preprint arXiv:2208.07903*, 2022.

[35] J. Y. Koh, H. Lee, Y. Yang, J. Baldridge, and P. Anderson, “Pathdreamer: A world model for indoor navigation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Springer, 2021, pp. 14 738–14 748.

[36] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Ragha- van, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, “Fourier features let networks learn high frequency functions in low dimensional domains,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.

[37] N. Max, “Optical models for direct volume rendering,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 2, pp. 99–108, 1995.

[38] R. Zhang, G. Riegler, N. Snavely, and V. Koltun, “Nerf++: Analyzing and improving neural radiance fields,” *arXiv preprint arXiv:2010.07492*, 2020.

[39] N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel, “Splicing vit features for semantic appearance transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 748–10 757.