Zhouchen Lin · Ming-Ming Cheng ·
Ran He · Kurban Ubul ·
Wushouer Silamu · Hongbin Zha ·
Jie Zhou · Cheng-Lin Liu (Eds.)

LNCS 15036

Pattern Recognition and Computer Vision

7th Chinese Conference, PRCV 2024
Urumqi, China, October 18–20, 2024
Proceedings, Part VI

6 Part VI

Springer

MOREMEDIA

AG-NeRF: Attention-Guided Neural Radiance Fields for Multi-height Large-Scale Outdoor Scene Rendering

Jingfeng Guo, Xiaohan Zhang, Baozhu Zhao, and Qi Liu(✉)

South China University of Technology, Guangzhou, China
{ftdrgjf,drliuqi}@mail.scut.edu.cn

Abstract. Existing neural radiance fields (NeRF)-based novel view synthesis methods for large-scale outdoor scenes are mainly built on a single altitude. Moreover, they often require *a priori* camera shooting height and scene scope, leading to inefficient and impractical applications when camera altitude changes. In this work, we propose an end-to-end framework, termed AG-NeRF, and seek to reduce the training cost of building good reconstructions by synthesizing free-viewpoint images based on varying altitudes of scenes. Specifically, to tackle the detail variation problem from low altitude (drone-level) to high altitude (satellite-level), a source image selection method and an attention-based feature fusion approach are developed to extract and fuse the most relevant features of target view from multi-height images for high-fidelity rendering. Extensive experiments demonstrate that AG-NeRF achieves SOTA performance on 56 Leonard and Transamerica benchmarks and only requires a half hour of training time to reach the competitive PSNR as compared to the latest BungeeNeRF.

Keywords: Novel view synthesis · NeRF · Large-scale outdoor scene rendering

1 Introduction

Large-scale outdoor scene reconstruction has an important application prospect to digitize a smart city in virtual reality and augmented reality. With the advance of neural radiance fields (NeRF) [1], the success spurs numerous researchers to study NeRF with high-frequency positional encoding for single object scene reconstruction and novel view synthesis, and achieves impressive results. Nevertheless, due to the limited model capacity, NeRF-based variants can only represent scenes reasonably at a macro scale yet exhibit excessively blurry artifacts and incomplete reconstruction when navigating closer to inspect micro details for large-scale outdoor scenes, as shown on the left of Fig. 1. To address that, several approaches [2–4] geographically decomposed the scene into several cells and trained a sub-NeRF for each cell before merging them, while others applied

Z. Lin et al. (Eds.): PRCV 2024, LNCS 15036, pp. 110–122, 2025.
https://doi.org/10.1007/978-981-97-8508-7_8

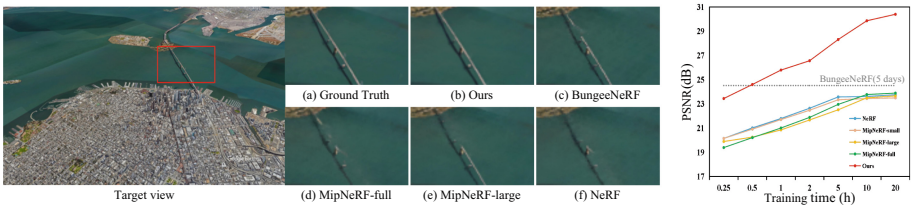

Fig. 1. Performance comparisons on two benchmark datasets. **Left**: visualization on Transamerica dataset. The visual results show that the proposal outperforms other competitors and can reconstruct the bridge completely. **Right**: PSNR versus training time on 56 Leonard dataset. Compared with others, we observe that ours gets 6 $\sim$ 7 dB improvement at PSNR. Moreover, it is worth noting that the proposed method only requires a half hour of training on a single RTX 4090 GPU to achieve competitive performance as the latest BungeeNeRF [7] (training over five days).

plane and grid features in parallel with positional encoding to achieve efficient modeling [5,6]. *However, all of them reconstruct large-sSScale scenes at the basis of identical low-altitude drone photos, rendering them for dealing with images of the same scene captured at varying heights in vain. This limitation arises from the significant detail variation during camera altitude changes. For the same scene, images captured at high-altitude primarily consist of low-frequency details, whereas images taken at low-altitude include more high-frequency details.* The pioneer to address large-scale outdoor scene reconstruction at varying heights was BungeeNeRF [7], by applying a progressive growth network with residual block and a multi-stage training paradigm to learn a hierarchy of scene representations. Nonetheless, to activate high-frequency channels in NeRF's positional encoding, BungeeNeRF [7] has to take use of the camera height to accurately partition the training dataset in a hand-crafted way. Furthermore, as the height increases, it inevitably grapples with an extensive model capacity, which takes up expensive training time non-amicable to limited GPU computing resources.

We seek to reduce the training cost of building good reconstructions by synthesizing novel views of large-scale scenes captured at different levels. To that end, we design the source image selection and attention-based feature fusion approaches to extract potential features from images at different heights as scene priors for NeRF processing. The proposed AG-NeRF significantly outperforms the baseline BungeeNeRF [7] and, as corroborated by indicative empirical results in Fig. 1, competitive to the state-of-the-art in terms of accuracy and speed. Our work makes notable contributions summarized as follows:

- We propose an end-to-end novel view synthesis framework, called AG-NeRF, for large-scale outdoor scene reconstruction. Different from the existing novel view synthesis approaches using drone photos of the same height, the proposed AG-NeRF is not affected by this limitation and is applicable to images captured at different levels. Moreover, the camera height is not prior to know.
- The proposed framework is helpful in providing the most relevant features for synthesizing the target view, and enabling high-quality image rendering

across all heights. This has been demonstrated by the comparison with the SOTA BungeeNeRF [7], where our approach achieves almost 6 dB increase in PSNR.
– Compared to the BungeeNeRF [7], ours takes up only a half hour of training on a single RTX 4090 GPU to achieve remarkable performance, while the competitor requires 5 days for training.

2 Related Work

2.1 NeRF and Its Extension

NeRF [1] has been extensively used in 3D reconstruction and novel view synthesis owing to its detailed scene geometry representation with complex occlusions, where views are synthesized by querying 3D point coordinates along camera rays and volume rendering is utilized to project the output colors and densities into an image. Recent works have been proposed to extend NeRF to unbounded scenes [8–10], dynamic scenes [11–14], few-shot setting [15–19], and large-scale outdoor scenes [2–6]. Another line of works uses display representations [20–25] to accelerate NeRF convergence and inference processes, achieving a huge speed increase.

2.2 Large-Scale Outdoor Scene Reconstruction and Rendering

Much effort has been devoted to extending NeRF to address large-scale outdoor scene reconstruction and rendering. Approaches like Block-NeRF [2] partition streets into discrete blocks and train a sub-NeRF for each block. Similarly, Mega-NeRF [3] divides drone-captured scenes into separate cells and trains a sub-NeRF for each cell. Switch-NeRF [4] points out that manually crafted scene decomposition relies on prior knowledge of the target scene, which limits the generalized use of these models, so a learnable gated network is designed to dynamically allocate 3D points to different sub-NeRF networks. More recently, Grid-NeRF [5] combines NeRF-based methods and grid-based approaches, jointly optimizing two branches for scene reconstruction. GP-NeRF [6] integrates orthogonal 2D high-resolution tri-plane feature and 3D hash-grid feature to achieve efficient scene representation. These methods achieve a fantastic novel view synthesis performance at the same height as the drone view. However, real-world scenes often involve images captured at varying heights even with substantial height disparities, which makes the rendering quality poor. BungeeNeRF [7] was proposed to employ a progressive training paradigm with residual block, dynamically expanding the neural network and synchronously adjusting the training images as the camera height increases. However, the BungeeNeRF requires an explicit split of scales among the input images, which has to proceed with manual adjustment. Moreover, the network becomes wider or deeper along with the growth of parameters as the camera height increases, leading to multi-stage training for several days. In contrast, the proposal is fully differentiable and can therefore be

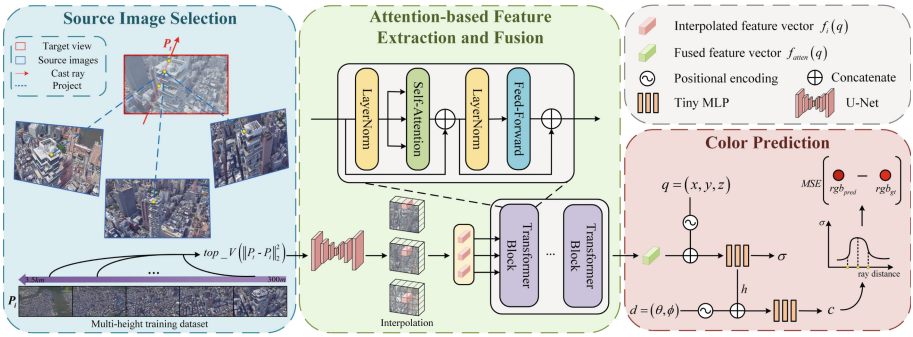

Fig. 2. Our pipline. First, according to the camera's external matrix, we select source images that are most similar to the target view from different heights. Next, a trainable U-Net-like network extracts feature maps from these source images. The 3D sample points along the rays are then projected back onto the image planes and interpolated for the corresponding feature vectors. Subsequently, these feature vectors interact with each other through an attention-based feature fusion approach and are fed into MLPs along with positional encoding. Finally, pixel color is calculated by volume rendering.

trained end-to-end. Different with BungeeNeRF, we directly extract more favorable scene priors from images captured at different heights. This enables us to utilize a tiny MLP for color production. Consequently, no manual intervention operation is required and only a half hour of training time is taken up to achieve competitive performance as BungeeNeRF.

2.3 Attention-Based NeRF Rendering and Optimization

Few works incorporated transformers into NeRF. IBRNet [26] innovatively employs a ray transformer to predict accurate densities for new scenes, expanding the applicability of NeRF from single scenes to multiple scenes. Recent work [27,28] relies on transformer to directly revert pixel color without volume rendering. Additionally, GeoNeRF [29] and LIRF [30] use transformer as a feature extractor in their pipelines. However, transformer have not been applied to large-scale scenes. To our knowledge, we are the first to leverage attention mechanisms, integrating feature vectors from different view images, to tackle the challenge of synthesizing novel views in large-scale outdoor scenes captured at varying heights.

3 Method

Overview. Given a set of multi-height training data, our goal is to synthesize images from any view at any altitude. We obtain camera intrinsic and extrinsic matrices through COLMAP [31] or Google Earth Studio. As shown in Fig. 2, for each target view, we select V source images via our source image selection

method (Sect. 3.1). For each sampled point along casted rays, we project them back to the source images and obtain fused features through our attention-based feature extraction and fusion approach (Sect. 3.2). Finally, colors and densities along each ray through volume rendering are composited to produce a synthesized image (Sect. 3.3).

3.1 Source Image Selection

Current NeRF-based large-scale outdoor reconstructors encode the entire scene into MLP layers to enhance the scene representations, which comes at the cost of high computational complexity. Different with them, we select few images from different heights by scoring to obtain scene priors. This is because the regions captured by close cameras constitute a subspace of those captured by remote cameras. Furthermore, these priors act as a guide to activate the high-frequency channels of positional encoding in close views, or to activate low-frequency channels in distant views. This adaptive mechanism effectively addresses the challenge posed by detail variation caused by altitude changes.

Given a target view with camera extrinsic matrix $P_t \in \mathbb{R}^{3\times4}$, we leverage the extrinsic matrix distance between two views as a metric to measure the similarity of two views, and select V images that are the closest to the target view at different heights as the source images $\{S_i \,|i = 1, \ldots, V\}$. By doing so, the selected source images overlap highly with the target view. This process can be written as:

$$source\_imgs = top\_V\left(\|P_t - P_i\|_2^2 \,|i = 1, \ldots, M\right) \tag{1}$$

where P_i represents the camera's external matrix for the i-th image in the training dataset, M is the number of images in the training dataset, top_V means to find the first V minimum values, and $\|\|_2$ denotes the l_2 norm.

3.2 Attention-Based Feature Extraction and Fusion

Feature Extraction. We apply a trainable U-Net-like network to extract feature maps $\{F_i \,|i = 1, \ldots, V\}$ from the source images. Subsequently, for a sampled point $q \in \mathbb{R}^3$ along a ray, we project q onto the image planes using *a priori* camera intrinsic matrices, then bilinear interpolation is applied between pixel-wise features to extract feature vectors. We also sample RGB colors on the images and concatenate them to the extracted feature vectors to form the final feature vectors $\{f_i\,(q) \,|i = 1, \ldots, V\}$. The feature extraction step can be mathematically expressed as:

$$f_i\,(q) = Inter(S_i, \pi_i\,(q)) \oplus Inter(F_i, \pi_i\,(q)) \tag{2}$$

where $\oplus$ denotes concatenation in channel dimension, $\pi_i\,(q)$ represents the coordinates on the image plane S_i obtained by projecting 3D point q.

Feature Fusion. While these source images with high overlap or panoramic do provide priors for the target view, they also introduce irrelevant information that is absent in the target view. Hence, an attention-based feature fusion

approach is developed, which combines all the feature vectors, to maximize the relevance between the fused feature and the target pixel. It is formalized as:

$$f_{atten}(q) = Transformer(f_1(q), ..., f_V(q)) \tag{3}$$

It is worth noting that the used transformer costs less computation as we only pay attention to the feature vectors extracted by interpolation.

3.3 Rendering and Training

Point Sampling. Similar with the NeRF [1], a hierarchical sampling approach is applied. A coarse and a fine networks are simultaneously optimized. We first uniformly sample N_c points to obtain the output of the coarse network, and then produce a more informed samples along each ray where samples are biased towards the surface of the object. Similarly, a set of N_f points is sampled and all $N_c + N_f$ points are applied to render fine results.

Training Objective. Positional encoding is employed to map 3D point coordinates $q_k = (x, y, z)$ and directions $d_k = (\theta, \phi)$ into high-dimensional space, making our MLP easier to approximate higher-frequency functions. We concatenate the positional encoding $(\gamma(q_k), \gamma(d_k))$ and the fused features $f_{atten}(q_k)$ as input to the MLPs to obtain the color c_k and density σ_k of the k-th sample on the ray. That is:

$$\begin{aligned} \sigma_k, h_k &= MLP_1(\gamma(q_k), f_{atten}(q_k)) \\ c_k &= MLP_2(\gamma(d_k), h_k) \end{aligned} \tag{4}$$

The volume rendering is then used to get the predicted color $\widehat{C}(r)$ of each pixel:

$$\widehat{C}(r) = \sum_{k=1}^{N} T_k (1 - \exp(-\sigma_k \delta_k)) c_k \tag{5}$$

where δ_k is the distance between adjacent sample points, $T_k = \exp\left(-\sum_{j=1}^{k-1} \sigma_k \delta_k\right)$ is the accumulated transmittance, and N is the number of sample points along a cast ray r.

We render the color of each ray using both the coarse and fine set of samples, and minimize the total squared error between the rendered colors and true pixel colors for training:

$$\mathcal{L} = \sum_{r \in R} \left[\left\| C(r) - \widehat{C}_c(r) \right\|_2^2 + \left\| C(r) - \widehat{C}_f(r) \right\|_2^2 \right] \tag{6}$$

where R is a set of rays in each batch, $C(r)$, $\widehat{C}_c(r)$ and $\widehat{C}_f(r)$ are the ground truth, coarse predicted, and fine predicted colors for ray r, respectively.

4 Experiments

4.1 Implementation Details

Datasets. Following BungeeNeRF [7], we evaluate our AG-NeRF on two large-scale outdoor scene datasets: 56 Leonard and Transamerica. These datasets comprehensively depict the real-world landscapes of New York and San Francisco, spanning a diverse range of altitudes from drone-level (about 300 m) to satellite-level (about 3.5 km). These images were captured during a circular ascent, evenly distributed across various altitudes. The images and camera parameter matrices for both datasets are obtained from Google Earth Studio by BungeeNeRF [7].

Evaluation Metrics. Similar with the NeRF-based approaches, we test our method against competitors in quantitative metrics as PSNR, SSIM [32], and LPIPS [33] implemented with VGG.

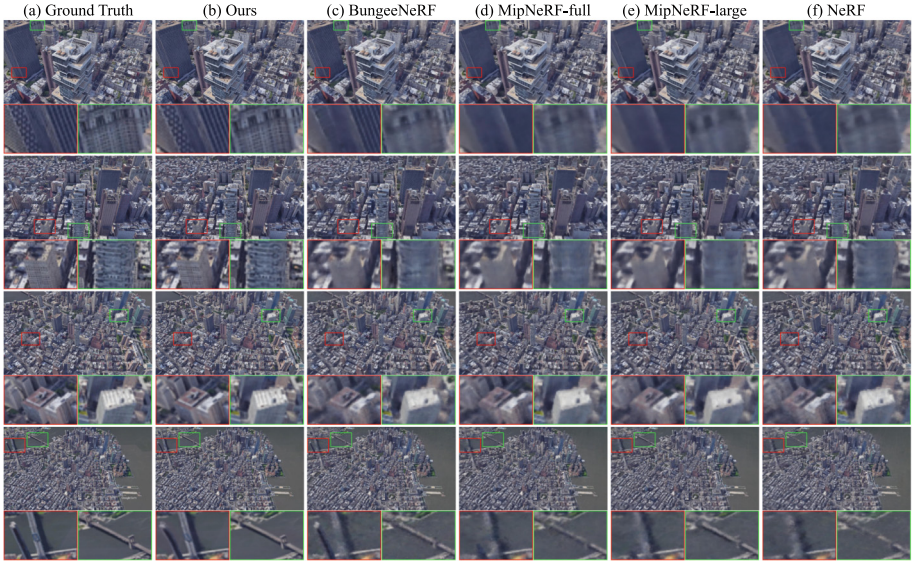

Fig. 3. Qualitative comparisons on 56 Leonard dataset.

Training Details. Both coarse and fine points are set as $N_c = 64$ and $N_f = 64$ per ray. The image feature extraction network using a U-Net-like architecture with ResNet34 [34] truncated after layer3 as the encoder, and two additional up-sampling layers with convolutions and skip-connections as the decoder. The dimension of the feature vector $f_i(q)$ is 35. The number of source images is empirically set as $V = 10$. The Feature fusion module employs 2 layers transformer block. The MLP_1 has 4 layers with 64 feature dimensions. The MLP_2 has only 1 layer with 64 feature dimensions. We apply the Adam as optimizer, with initial learning rates of $10e^{-3}$ for the feature extraction network and $5 \times 10e^{-4}$

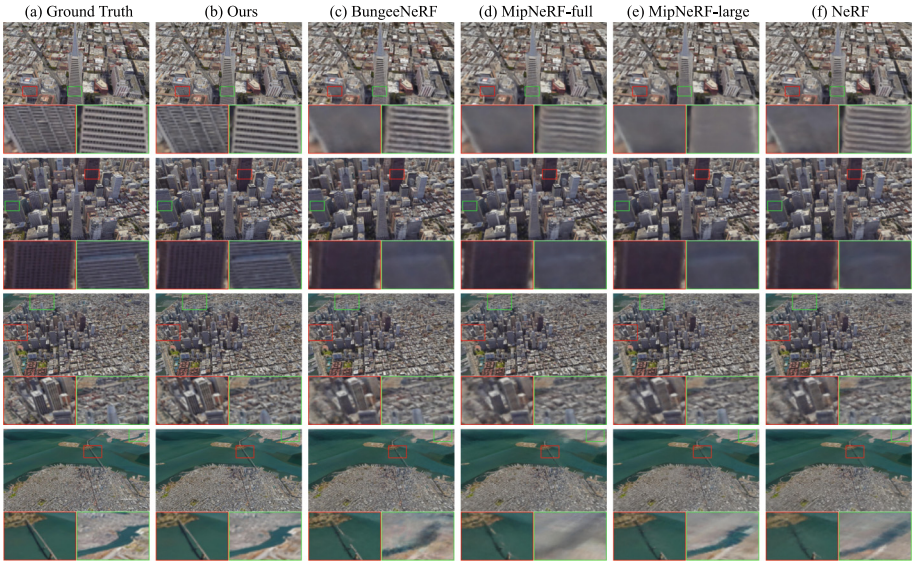

Fig. 4. Qualitative comparisons on Transamerica dataset.

for our tiny MLP. All experiments are conducted in $200k$ iterations with 2048 rays sampled in each iteration on a single RTX 4090 GPU.

Table 1. Quantitative comparisons on 56 Leonard and Transamerica datasets. D denotes MLP_1 depth, d denotes MLP_1 width and *skip* indicates which layer(s) the skip connection is inserted to.

	56 Leonard			Transamerica		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [1] ($D = 8$, $d = 256$, $skip = 4$)	23.570	0.729	0.324	24.046	0.747	0.311
Mip-NeRF-small [35] ($D = 8$, $d = 256$, $skip = 4$)	23.337	0.709	0.354	23.929	0.733	0.330
Mip-NeRF-large [35] ($D = 10$, $d = 256$, $skip = 4$)	23.507	0.718	0.346	24.113	0.744	0.315
Mip-NeRF-full [35] ($D = 10$, $d = 256$, $skip = 4, 6, 8$)	23.665	0.732	0.328	24.113	0.748	0.314
BungeeNeRF [7] ($D = 10$, $d = 256$, $skip = 4, 6, 8$)	24.513	0.815	0.160	24.415	0.801	0.192
Ours ($4k$ iterations, $D = 4$, $d = 64$, $skip = 1$)	24.570	0.837	0.259	25.063	0.846	0.262
Ours (same iter as baselines, $D = 4$, $d = 64$, $skip = 1$)	**30.450**	**0.963**	**0.065**	**30.599**	**0.959**	**0.076**

4.2 Experiment Results

We compare the proposed AG-NeRF against the state-of-the-art method BungeeNeRF [7], vanilla NeRF [1], and three forms of MipNeRF [35].

We report the performance of our method and baselines in Figs. 1, 3, 4 and Table 1. It can be observed that there are significant improvements both in qualitative and quantitative metrics. In comparison with the SOTA method BungeeNeRF [7], the proposed AG-NeRF achieves an increase of about 6 dB in PSNR on

both the 56 Leonard and Transamerica datasets. In close views, baseline methods result in blurry textures due to substantial variation in detail levels, while the proposal can reconstruct sharper geometric and more delicate details. On the other hand, these competitors usually fail to reconstruct complete geometry in remote views, such as slender bridges or densely populated buildings located far from the camera. The proposed AG-NeRF can answer in affirmative to synthesize complete images.

The BungeeNeRF [7] is time-consuming because of its multi-stage training process. Each stage in this sequential training paradigm relies on the checkpoint file generated in the preceding stage, which means each stage has to be fully trained before progressing to the next. It takes up to 5 days to finish training four stages of BungeeNeRF via a single RTX 4090 GPU. In contrast, the proposed method only costs 1 day to converge.

It is worth noting that, with a distinct advantage in leveraging scene priors from our source images, the proposal AG-NeRF facilitates the rapid acquisition of overall geometry of scenes and then gradually recovers texture details. As evidenced on the right of Fig. 1, our model, with shallower MLP, using only $4k$ iterations of training (about a half hour) can achieve competitive performance compared to BungeeNeRF.

Table 2. Ablation study. When the number of source images $V = 0$, the framework is equivalent to the vanilla NeRF ($D = 4$, $d = 64$, $skip = 1$). No attention (AvgPool) means avg-pooling the interpolated feature vectors directly. No attention (MaxPool) means max-pooling the interpolated feature vectors directly.

	56 Leonard			Transamerica		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
No source image ($V = 0$)	19.250	0.370	0.608	20.027	0.416	0.588
No attention (AvgPool)	28.719	0.941	0.101	28.743	0.935	0.120
No attention (MaxPool)	28.994	0.946	0.090	28.950	0.939	0.109
Ours	**30.450**	**0.963**	**0.065**	**30.599**	**0.959**	**0.076**

4.3 Ablation Study

Effectiveness of the Source Images. To validate the effectiveness of source images to provide scene priors, we test our approach on the 56 Leonard dataset using varying numbers of source images. Figure 5 shows that, as the number of source images increases from 2 to 14, our method achieves about 1.3 dB gain in PSNR on the 56 Leonard dataset. When the number of source images exceeds 10, there is a tiny margin on 3D reconstruction result. Therefore, to balance the trade-off between computational efficiency and rendering quality, we choose 10 source images for experiments.

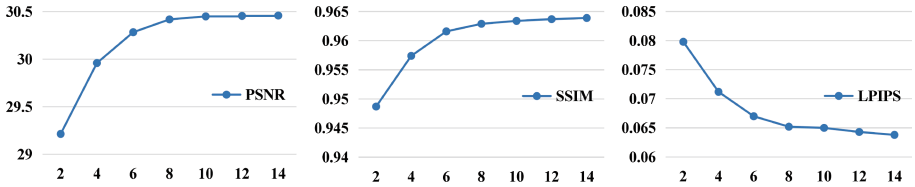

Fig. 5. Comparison on the effect of source image number. The horizontal axis represents the number of chosen source images V.

Furthermore, we conduct another ablation study excluding source image selection method ($V = 0$), that is, vanilla NeRF [1]. This is because the attention-based feature extraction and fusion module fails to take effect without the inputs of source images. The results are summarized in the first row of Table 2, compared to ours, there is about 11 dB decrease in PSNR.

Effectiveness of Attention-based Feature Fusion Approach. To evaluate the effectiveness of our attention-based feature fusion approach in capturing features corresponding to target pixel, avg-pooling or max-pooling are applied for the interpolated feature vectors. As shown from the second and third rows of Table 2, under the same number of source images, avg-pooling case results in an approximately 1.7 dB decrease in PSNR, and max-pooling case leads to about 1.5 dB decrease in PSNR as compared to our model.

5 Conclusion

In this work, we target on rendering remarkable large-scale outdoor scenes captured at varying altitudes. To that end, we propose an end-to-end framework, termed AG-NeRF. NeRF-based approaches typically require hours to days to be trained. Advanced NeRF variants can alleviate this, but tend to be less accessible to the large-scale outdoor scene reconstruction due to the requirement of low-level rendering, complex parameter tuning or a computational bottleneck. It is envisioned that the proposal can model diverse multi-scale scenes with drastically varying city views. Experiments demonstrate that AG-NeRF is competitive to the SOTA in terms of accuracy and speed.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the Fundamental Research Funds for the Central Universities under Grant 2023ZYGXZR085, in part by the Basic and Applied Basic Research Foundation of Guangzhou under Grant 2023A04J1674, in part by The Taihu Lake Innocation Fund for the School of Future Technology of South China University of Technology under Grant 2024B105611004 and in part by the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004.

References

1. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. Springer, pp. 405–421 (2020)
2. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-NeRF: scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8248–8258 (2022)
3. Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-NeRF: scalable construction of large-scale NeRFs for virtual fly-throughs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12922–12931 (2022)
4. Zhenxing, M.I., Xu, D.: Switch-NeRF: learning scene decomposition with mixture of experts for large-scale neural radiance fields. In: The Eleventh International Conference on Learning Representations (2022)
5. Xu, L., Xiangli, Y., Peng, S., Pan, X., Zhao, N., Theobalt, C., Dai, B., Lin, D.: Grid-guided neural radiance fields for large urban scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8296–8306 (2023)
6. Zhang, Y., Chen, G., Cui, S.: Efficient large-scale scene representation with a hybrid of high-resolution grid and plane features (2023). arXiv:2303.03003
7. Xiangli, Y., Xu, L., Pan, Zhao, X.N., Rao, A., Theobalt, C., Dai, B., Lin, D.: BungeeNeRF: progressive neural radiance field for extreme multi-scale scene rendering. In: European Conference on Computer Vision, pp. 106–122. Springer (2022)
8. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: NeRF++: Analyzing and Improving Neural Radiance Fields (2020). arXiv:2010.07492
9. Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the wild: neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7210–7219 (2021)
10. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5470–5479 (2022)
11. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10318–10327 (2021)
12. Lin, H., Peng, S., Zhen, X., Yan, Y., Shuai, Q., Bao, H., Zhou, X.: Efficient neural radiance fields for interactive free-viewpoint video. In: Conference Papers SIGGRAPH Asia, vol. 2022, pp. 1–9 (2022)
13. Li, Z., Wang, Q., Cole, F., Tucker, R., Snavely, N.: Dynibar: neural dynamic image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 4273–4284 (2023)
14. Jiang, Y., Hedman, P., Mildenhall, B., Xu, D., Barron, J.T., Wang, Z., Xue, T.: Alignerf: high-fidelity neural radiance fields via alignment-aware training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 46–55 (2023)
15. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4578–4587 (2021)

16. Yang, J., Pavone, M., Wang, Y.: FreeNeRF: improving few-shot neural rendering with free frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8254–8263 (2023)
17. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12892–12901 (2022)
18. Deng, K., Liu, A., Zhu, J.-Y., Ramanan, D.: Depth-supervised NeRF: fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12882–12891 (2022)
19. Yuan, Y.-J., Lai, Y.-K., Huang, Y.-H., Kobbelt, L., Gao, L.: Neural radiance fields from sparse RGB-d images for high-quality view synthesis. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
20. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. (ToG) **41**(4), 1–15 (2022)
21. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: tensorial radiance fields. In: European Conference on Computer Vision. Springer, pp. 333–350 (2022)
22. Sun, C., Sun, M., Chen, H.-T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5459–5469 (2022)
23. Liu, L., Jiatao, G., Lin, K.Z., Chua, T.-S., Theobalt, C.: Neural sparse voxel fields. Adv. Neural. Inf. Process. Syst. **33**, 15651–15663 (2020)
24. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5501–5510 (2022)
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4) (2023)
26. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2021)
27. Zhao, Z., Jia, J.: End-to-end view synthesis via nerf attention (2022). arXiv:2207.14741
28. Varma, M., Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z.: Is attention all that nerf needs? In: The Eleventh International Conference on Learning Representations (2022)
29. Johari, M.M., Lepoittevin, Y., Fleuret, F.: GeoNeRF: generalizing nerf with geometry priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18365–18375 (2022)
30. Huang, X., Zhang, Q., Feng, Y., Li, X., Wang, X., Wang, Q.: Local implicit ray function for generalizable radiance field representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 97–107 (2023)
31. Schonberger, J.L., Frahm, J.-M.: Structure-from-Motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)

33. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
35. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5855–5864 (2021)