
Volume Feature Rendering for Fast Neural Radiance Field Reconstruction

Kang Han Wei Xiang* Lu Yu

Department of Computer Science and Information Technology
La Trobe University
`{k.han, w.xiang, l.yu}@latrobe.edu.au`

Abstract

Neural radiance fields (NeRFs) are able to synthesize realistic novel views from multi-view images captured from distinct positions and perspectives. In NeRF’s rendering pipeline, neural networks are used to represent a scene independently or transform queried learnable feature vector of a point to the expected color or density. With the aid of geometry guides either in occupancy grids or proposal networks, the number of neural network evaluations can be reduced from hundreds to dozens in the standard volume rendering framework. Instead of rendering yielded color after neural network evaluation, we propose to render the queried feature vectors of a ray first and then transform the rendered feature vector to the final pixel color by a neural network. This fundamental change to the standard volume rendering framework requires only one single neural network evaluation to render a pixel, which substantially lowers the high computational complexity of the rendering framework attributed to a large number of neural network evaluations. Consequently, we can use a comparably larger neural network to achieve a better rendering quality while maintaining the same training and rendering time costs. Our model achieves the state-of-the-art rendering quality on both synthetic and real-world datasets while requiring a training time of several minutes.

1 Introduction

View synthesis is a task that synthesizes unobserved views of a scene from captured multi-view images. This is a long-standing problem that has been studied for several decades in computer vision and computer graphics. The most popular solution to this problem at present is the neural radiance field (NeRF) [15] since it can render views with high fidelity. NeRF represents the density and color of a spatial point in a given direction using a neural network (NN), typically the multilayer perceptron (MLP). With the predicted densities and colors of sampled points along a ray, the NeRF and its variants [2, 3, 24] use the volume rendering technique to aggregate colors to obtain the final rendered pixel color. As many samples are required to render one pixel, the underlying MLP needs to run many times, leading to high computational complexities for both training and rendering. Alternatively, researchers introduced extra learnable features in forms of 3D grids [21, 14], hash tables [16] and decomposed tensors [5, 9, 6, 19] in addition to MLP’s parameters to represent a scene’s density and color fields. As querying the feature vector of a sampled point by interpolation is much faster than one MLP evaluation, only a small MLP is used to transform the queried feature vector to the density or color, leading to substantial speed acceleration compared with pure MLP representations.

Furthermore, modeling the density field independently by computational efficient representations can greatly reduce the number of samples in the color field. The density representations could be occupancy grids [16], hash grids with less parameters [13], and small MLPs [3]. These coarse density

*Corresponding author.

fields enable importance sampling to make the model focus on non-empty points near surface. As a result, only dozens of samples in the color field are needed to render a pixel. For example, Zip-NeRF [4] uses 32 samples after importance sampling through efficient density representations.

Nevertheless, standard volume rendering technique still needs to run a NN dozens of times to render a single pixel. This is why the Instant-NGP [16] uses a small MLP for fast training and rendering, and why the latest Zip-NeRF [4] is comparably slow when using large MLPs to achieve a better rendering quality. The importance of large NNs for high-quality rendering is undoubtedly, considering that the state-of-the-art rendering quality can be achieved by pure MLP representations. Also, the authors of Zip-NeRF [4] and NRFF [9] also highlight the importance of large NNs in modeling the view-dependent effect. Unfortunately, the standard volume rendering technique limits the size of NNs for the sake of fast training and rendering. Towards this end, we propose a novel volume feature rendering (VFR) method, which renders queried feature vectors into an integrated feature vector and then maps it to a final pixel color. In this way, only one single NN evaluation is required to render a pixel. This fundamental change to the standard volume rendering framework enables the use of large NNs for better rendering quality, while requiring a training time in the order of several minutes.

2 Related work

The volume rendering technique integrates the colors of samples along a ray by their densities [12]. A global density field that is differentiable is effective in finding the underlying geometry of a scene from multi-view 2D color observations using volume rendering. This capability is one of the essential reasons to NeRF’s success. As every point in a scene has a density value, the density field is capable of modeling complex geometry. However, standard volume rendering employed by NeRF and its successors [15, 2, 3, 22, 30, 23, 5] suffer from high computational complexity caused by many NN runnings even with importance sampling guided by coarse density fields.

On the other hand, representing the underlying geometry by the signed distance function (SDF) leads to a well-defined surface to enable one sample on the surface for fast rendering [29, 7, 25, 17, 8, 28]. However, the SDF struggles with modeling complex geometry in the context of neural inverse rendering, e.g., trees and leaves, and this is why the best rendering quality is still achieved by volume representations. In this paper, we propose a new volume feature rendering framework that uses density as the geometry representation but shares a similar behavior as the SDF that processes a feature vector by NN only once. Thus, the proposed VFR inherits the advantage of volume representation in modeling complex geometry and also has the strength of the SDF in single NN evaluation.

It is noted that feature accumulation is also discussed in the Sparse Neural Radiance Grid (SNeRG) [10] mainly for the purpose of modeling the specular effect. However, the diffuse color is still obtained using the standard volume rendering method in [10], which requires many times of MLP evaluations. Besides, the SNeRG needs to train a NeRF first (require days to train) and then bakes the trained NeRF to a sparse grid with specular features. In this work, we demonstrate that this pre-training is not necessary, and the accumulated feature vector can be used to predict the final view-dependent color instead of only the specular color. Our method can be directly optimized from scratch instead of baking the optimized features in the SNeRG. As a result, our method requires significantly less training time but achieves a much better rendering quality compared with the SNeRG.

Integrating feature vectors of samples along the ray is natural in transformer models [20, 26]. For example, Suhail *et al.* proposed an epipolar transformer to aggregate features extracted from reference views on epipolar lines [20]. However, the computational cost of transformer is much higher than MLP and that transformer-based method is significantly (8 times) slower than MLP-based Mip-NeRF on the same hardware. In this work, we propose a volume feature rendering framework that aggregates feature vectors using densities. The proposed aggregation only involves a weighted combination of feature vectors such that our method is significantly faster than transformer-based feature aggregation.

3 Model architecture

3.1 Standard volume rendering

Standard volume rendering integrates color along a cast ray according to density [12]. A high density of a point on the ray indicates a high probability of hitting the surface. The NeRF and its various

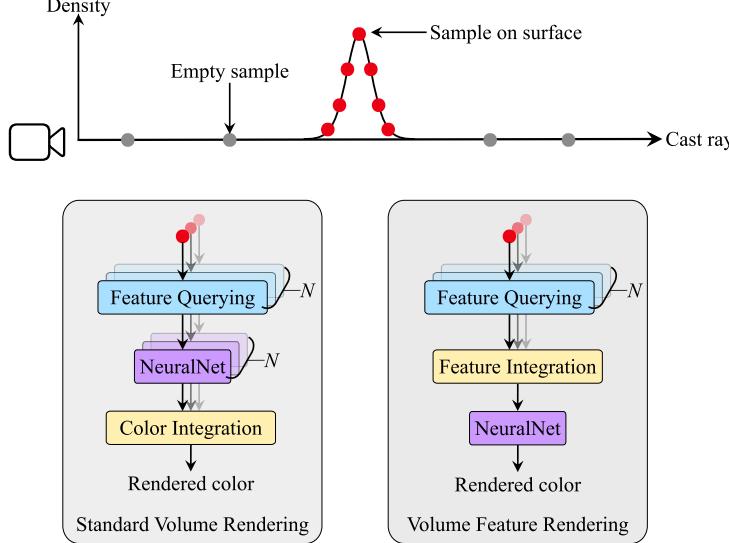


Figure 1: Standard volume rendering needs to run NeuralNet many times for all the samples on the surface, while the proposed volume feature rendering only needs to evaluate NeuralNet once.

variants adopt this standard volume rendering technique, where the color and density of a point are typically predicted by a neural network. For a cast ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, where \mathbf{o} is the camera origin and \mathbf{d} is the view direction, the rendered color $C(\mathbf{r})$ using standard volume rendering is:

$$C(\mathbf{r}) = \int T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt, \text{ where } T(t) = \exp\left(-\int_0^t \sigma(\mathbf{r}(s)) ds\right) \quad (1)$$

where $\sigma(\mathbf{x})$ and $\mathbf{c}(\mathbf{x}, \mathbf{d})$ are the density and color at position \mathbf{x} , respectively. In practice, the above integral is solved by sampling and integration along the cast ray. After transforming the densities to weights by $w(t) = T(t)\sigma(\mathbf{r}(t))$, the color will be rendered as follows:

$$C(\mathbf{r}) = \sum_{i=1}^N w_i \text{NeuralNet}(F(\mathbf{x}_i), \mathbf{d}) \quad (2)$$

where \mathbf{x} is the position of a sample point on the ray and F is a function to query the corresponding feature vector of \mathbf{x} . Pure MLP-based methods including the NeRF [15], Mip-NeRF [2], Ref-NeRF [24] and Mip-NeRF 360 [3] represent F by use of MLPs. However, this representation imposes a computational burden as N times MLP evaluations are required to render a ray. Alternatively, recent research shows that modeling F by extra learnable features is significantly faster than pure-MLP based representations, because linear interpolation of learnable features is much more efficient than MLP evaluations. These learnable features are typically organized in the forms of the 3D grids, hash tables, and decomposed tensors. With the learnable features, a small NN can achieve the state-of-the-art rendering quality but accelerate both the training and rendering time significantly.

Furthermore, hierarchical importance sampling guided by coarse geometries is employed to greatly reduce the number of samples in (2). This importance sampling can be achieved by modeling the density fields independently [21, 5, 9], occupancy grid [16], and proposal networks [3, 4]. As shown in Fig. 1, importance sampling makes the model focus on valuable samples on the surface. For instance, by using two levels of importance sampling, the Zip-NeRF reduces the final number of valuable sampling to 32. However, the NN still needs to run 32 times to deliver the best rendering quality. Although integrating colors predicted by the NN is in line with the concept in classical volume rendering, the many times NN running is the main obstacle for fast training and rendering. On the other hand, the importance of a large NN in realistic rendering especially in modeling the view-dependent effect is highlighted in recent work [3, 9]. Considering the fact that valuable samples on the surface are close in terms of spatial position, the queried feature vectors are also very similar as they are interpolated according to their position. As such, evaluating the NN with similar input feature vectors is not necessary. In the next subsection, we will introduce the proposed volume feature rendering method that integrates these feature vectors to enable a single evaluation of the NN.

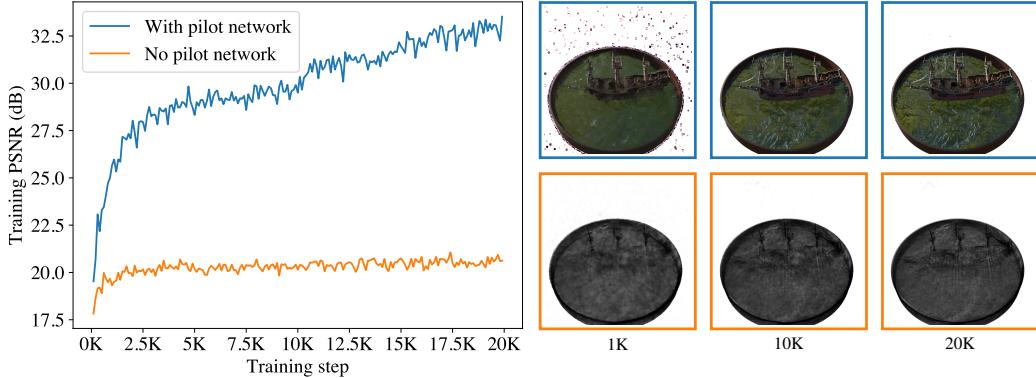


Figure 2: The proposed VFR fails to converge on the scene of *ship* from the dataset in [15]. A small pilot network functioning as standard volume rendering by a few training steps in the early training stage resolves this problem.

3.2 Volume feature rendering

As shown in Fig. 1, the proposed volume feature rendering method integrates queried feature vectors instead of predicted colors to form a rendered feature vector. A subsequent NN is then applied to transform the feature vector to the final rendered color. By doing this, the NN only needs to run once. Specifically, we take the NN out of the integral as follows:

$$C(\mathbf{r}) = \text{NeuralNet} \left(\left(\sum_{i=1}^N w_i F(\mathbf{x}_i) \right), \mathbf{d} \right). \quad (3)$$

This fundamental change to the standard volume rendering framework relieves the rendering framework from the high computational cost caused by many NN evaluations. Consequently, we can use a larger network to achieve a better rendering quality while maintaining a similar training time.

3.3 Pilot network

It is reasonable to integrate feature vectors of non-empty samples on the surface in the proposed VFR. However, at the beginning of training, the model does not have a coarse geometry to focus on samples on the surface. The VFR will integrate feature vectors of all samples during the early training stage. We found that this feature integration works well for most scenes but fails to converge for some scenes. Such failure is more likely to occur when employing a comparably large NN. The VFR normally converges using the NN with two hidden layers each of 256 neurons, but it does not converge when using 4 hidden layers. Fig. 2 illustrates a failure case on the scene of *ship* from the NeRF synthetic dataset [15].

As increasing the size of the NN only has a slight implication to reconstruction and rendering speeds but can lead to a better rendering quality thanks to our proposed VFR framework, we are motivated to use a large NN. We hypothesize that the reason for the aforementioned convergence failure is that the model cannot find the correct geometry due to the feature integration of all samples in the early training stage. Motivated by this insight, we design a pilot network that functions as standard volume rendering to aid the model in finding a coarse geometry in the early training stage. A small pilot network and a few training steps are sufficient to achieve this goal. In this work, we use the pilot network with two hidden layers, each with 64 neurons. The pilot network is only used in the first 300 training steps. After this training, the pilot network will be discarded, and we can use the proposed VFR to train the model normally. The number of pilot training steps is small compared with the total training steps, e.g., 20K steps. Thus, it has a negligible effect on the overall training time. As shown in Fig. 2, the model converges normally with the aid of the pilot network.

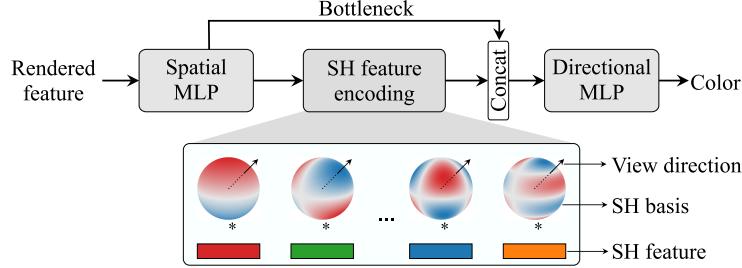


Figure 3: Architecture of the designed neural network for modeling the view-dependent effect. Each spherical harmonics basis function is multiplied by a feature vector to encode the input view direction.

3.4 View-dependent effect

We use the spherical harmonics (SH) feature encoding method to encode view direction for modeling view-dependent effect. This approach can be seen as a variant of the rendering equation encoding in [9]. Different from the vanilla SH encoding [16] employed in the literature, we multiply each encoded SH coefficient by a feature vector as shown in Fig. 3. Mathematically, an SH feature encoding \mathbf{e}_l^m can be expressed as:

$$\mathbf{e}_l^m = \mathbf{f}_l^m Y_l^m(\mathbf{d}) \quad (4)$$

where \mathbf{f}_l^m is an SH feature vector and Y_l^m is the SH basis function with degree l and $m = \{-l, -(l-1), \dots, 0, \dots, l-1, l\}$. \mathbf{f}_l^m is a feature vector predicted by a spatial MLP from the rendered feature vector. All resultant SH encodings form a comprehensive encoding vector, which will be concatenated with a view-independent bottleneck feature vector. The concatenated one will feed into a directional MLP to predict the final rendered color. The NN in (2) thus includes the spatial and directional MLPs.

4 Implementation details

We implement the proposed VFR using the NerfAcc library [13], which is based on the deep learning framework PyTorch [18]. The learnable features are organized by a multiresolution hash grid (MHG) [16]. This feature representation models the feature function F in (3) that accepts the input of the position and outputs a feature vector. In the MHG, the number of feature channels in each layer is two and the positions of voxels will be hashed to a one-dimensional table if the number of voxels is larger than 2^{19} . The NN contains a spatial MLP and a directional MLP. The spatial MLP has two layers and the directional MLP has four layers, all with 256 hidden neurons. The size of the bottleneck feature is set to 256. We use the GELU [11] instead of the commonly used ReLU [1] activation function, as we found the GELU results in a slightly better quality.

The density of a sample \mathbf{x}_i is derived from the queried feature vector $F(\mathbf{x}_i)$ by a tiny density mapping layer. On the real-world 360 dataset [3], the mapping layer is a linear transform from the feature vector to the density value. On the NeRF synthetic [15] dataset, we found a tiny network with one hidden layer and 64 neurons yields a slightly better rendering quality. The density mapping layers have negligible impact on the overall training time for both datasets.

We employ a hierarchical importance sampling strategy on both the NeRF synthetic [15] and real-world 360 datasets [3]. The occupancy grid is used for efficient sampling on the NeRF synthetic dataset. For the 360 dataset, we follow the most recent Zip-NeRF [4] that uses two levels of importance sampling, each with 64 samples. These two density fields are implemented by two small

Table 1: Experimental settings on the NeRF synthetic and real-world 360 datasets. #Levels mean the number of hash grid levels.

Dataset	#Batch	#Steps	#MHG	#Levels	Efficient samp.	SH degree	#SH feature
NeRF synthetic[15]	2^{18} points	20K	13M	16	Occupancy grid	4	4
Real-world 360[3]	2^{13} rays	40K	34M	32	Proposal net	7	8

Table 2: Rendering quality in PSNR on the NeRF synthetic dataset. Our VFR trained with 6K steps in 0.97 minutes achieves a similar quality compared with existing methods.

	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	MEAN	Time
NeRF [15]	34.08	25.03	30.43	36.92	33.28	29.91	34.53	29.36	31.69	hours
Mip-NeRF [2]	35.14	25.48	33.29	37.48	35.70	30.71	36.51	30.41	33.09	hours
TensoRF [5]	35.76	26.01	33.99	37.41	36.36	30.12	34.61	30.77	33.14	10 mins
Instant-NGP [16]	35.10	23.85	30.61	37.48	35.87	29.08	36.22	30.62	32.35	4.2 mins
NerfAcc [13]	35.71	25.44	33.95	37.37	35.67	29.60	36.85	30.29	33.11	4.3 mins
Our VFR: 6K	35.65	26.29	35.28	36.82	35.12	29.92	36.30	28.80	33.02	0.97 mins
Our VFR: 20K	36.81	26.63	36.27	38.74	37.01	31.44	38.89	31.19	34.62	3.3 mins

MHGs, where the number of hashing levels is 10 and the feature channel in each level is one. The final number of samplings on the radiance field (also implemented by an MHG) is 32. The models are trained using the Adam optimizer with a learning rate of 0.01 and the default learning rate scheduler in NerfAcc [13]. We use PSNR, SSIM [27] and LPIPS [31] to evaluate the rendered image quality. The running times are measured on one RTX 3090 GPU unless specified. Table 1 summarizes other experimental settings in this work.

5 Experimental results

5.1 NeRF synthetic dataset

The proposed VFR achieves the best rendering quality but uses the minimum training time compared with state-of-the-art fast methods on the NeRF synthetic dataset. Since the original Instant-NGP [16] uses the alpha channel in the training images to perform background augmentation, we report the results of the Instant-NGP from NerfAcc [13], where such augmentation is not employed for fair comparison. As shown in Table 2, our VFR outperforms the Instant-NGP by 2.27 dB and NerfAcc by 1.51 dB, but only uses 3.2 minutes for training. Compared with TensoRF [5], our method only uses 33% of its training time while surpassing TensoRF by 1.48 dB in PSNR.

Specifically, the proposed VFR significantly outperforms the compared methods on all scenes. On most scenes, over 1 dB PSNR improvements are achieved by our proposed method compared with the Instant-NGP and NerfAcc. Noticeably, over 2 dB quality improvements on the scenes of *ficus* and *mic* are observed in Table 2. In addition, our VFR only needs 6K training steps to reach a similar rendering quality compared with the existing algorithms. Such 6K training steps can be completed within 1 minute on one RTX 3090 GPU, which is less than the 25% of training times of the Instant-NGP and NerfAcc, and only 10% of the training time in the TensoRF.

This significant improvement in rendering quality stems from the increased size of the NN. The MLP in the Instant-NGP and NerfAcc has 4 layers and 64 neurons in each layer. By comparison, we use a 6-layer MLP with 256 neurons. The computational cost of our MLP is roughly 24 times higher than MLPs in the Instant-NGP and NerfAcc. However, thanks to the proposed VFR framework, our MLP only needs to run one time instead of many times as in the standard volume rendering framework employed in the compared methods. As can be observed from Table 2, the single NN evaluation property in our VFR leads to a minimum training time even when we use a much larger NN.

Fig. 4 provides a visual comparison of the rendered novel views by the NerfAcc (based on the Instant-NGP) and our VFR. On the compared scenes, both NerfAcc and the proposed VFR can recover good geometries. However, for scenes *drums* and *ficus*, obvious visual artefacts appear in the rendered views from NerfAcc while our results do not have such visual artefacts. Besides, our VFR produces more realistic visual effects such as shadow and reflection, as observed on the scenes of *hotdog* and *mic*. The presented error maps and objective PSNRs in Fig. 4 also demonstrate the advantage of our VFR in terms of rendering quality. While NerfAcc and our method use the same underlying feature representation, i.e., hash grids, we believe the quality advantage of our method stems from the large NN enabled by the proposed VFR framework.

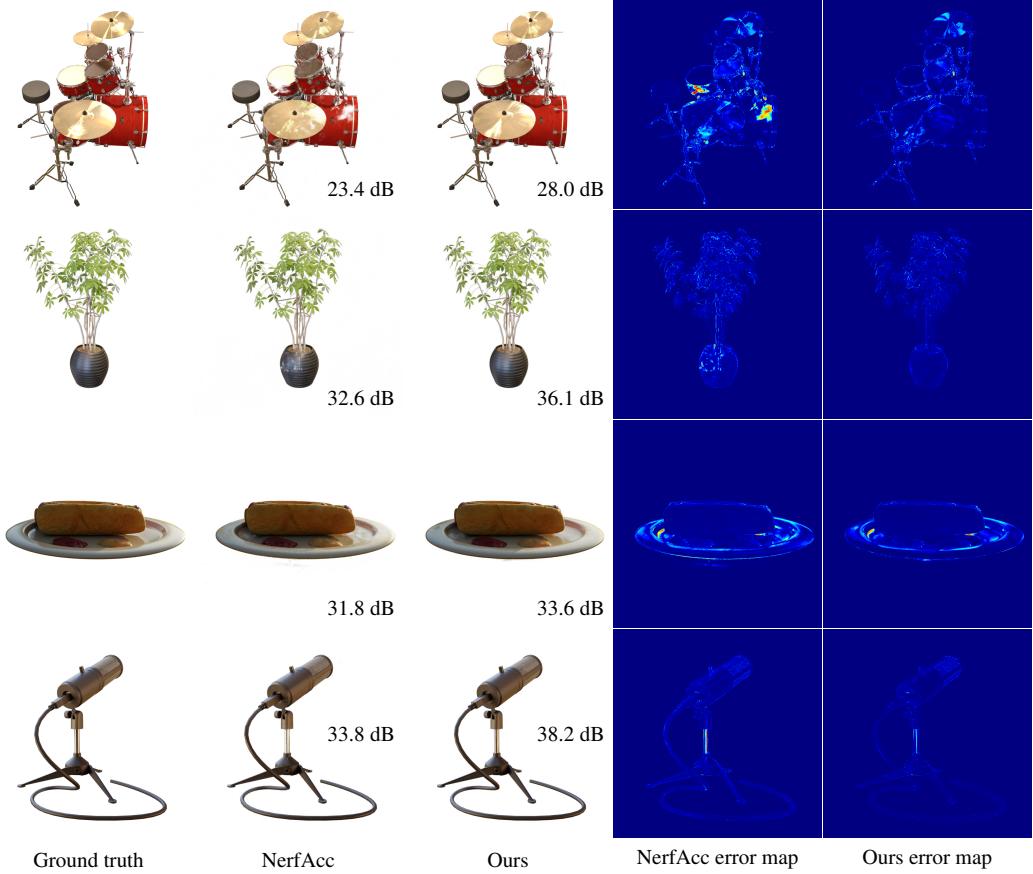


Figure 4: Visual comparison of the synthesized novel views on the scenes of (from top to bottom) *drums*, *ficus*, *hotdog*, and *mic* in the NeRF synthetic dataset. Readers are encouraged to zoom in for a detailed comparison.

5.2 Real-world 360 dataset

We also experiment on 7 real-world unbounded scenes from the 360 dataset [3]. The scenes of *flowers* and *treehill* in this dataset are not included as they are not publicly available. We use the unbounded ray parameterization method proposed in the Mip-NeRF 360 [3], where points are mapped to a sphere if their radius to the scene center is larger than one. The latest Zip-NeRF [4] reports on the rendering qualities and training times of state-of-the-art algorithms on this dataset. So we use their results of NeRF [15], Mip-NeRF [2], NeRF++ [30], Mip-NeRF 360 [3], and Instant-NGP [16] for comparison. As the training times of these methods in [4] are measured on 8 V100 GPUs, we follow the Zip-NeRF that rescales the training time by multiplying the number of used GPUs. We acknowledge that this approach may not be rigorous, but the rescaled times provide a rough comparison in terms of training efficiency.

As shown in Table 3, our proposed VFR is able to achieve state-of-the-art rendering quality but using the minimum training time compared with the existing methods. Although the Mip-NeRF 360 is able to synthesize novel high-fidelity views, it requires days to train on the high-end GPUs because it is based on a pure MLP representation. Using learning features organized in hash grids, the Instant-NGP achieves a comparable rendering quality but reduces the training time from days to hours. The most recent Zip-NeRF uses multisampling for high-quality anti-aliasing view rendering at the expense of training time relative to the Instant-NGP. Our method achieves a slightly better rendering quality in terms of the mean PSNR. However, we achieve this state-of-the-art rendering quality with a significantly less training time. As a comparison, our model only uses 11 minutes to train on one RTX 3090 GPU, while Zip-NeRF's training time is approximately 7.7 hours on one V100 GPU.

Table 3: Rendering quality in PSNR on the real-world 360 dataset. The training times of various methods are simply multiplied by the number of used GPUs to provide a rough comparison. #Feat. represents the number of learnable features in the underlying hash grids, which is estimated by the number of levels \times the number of voxels each level \times the number of feature channels.

	Outdoor				Indoor				MEAN	#Feat.	Time	Hardware
	Bicycle	Garden	Stump	Bonsai	Counter	Kitchen	Room					
NeRF [15]	21.76	23.11	21.73	26.81	25.67	26.31	28.56	24.85	N/A	days	V100	
Mip-NeRF [2]	21.69	23.16	23.10	27.13	25.59	26.47	28.73	25.12	N/A	days	V100	
NeRF++ [30]	22.64	24.32	24.34	29.15	26.38	27.80	28.87	26.21	N/A	days	V100	
Mip-NeRF 360 [3]	24.38	26.95	26.40	32.97	29.47	32.07	31.53	29.11	N/A	days	V100	
Instant-NGP [16]	22.79	25.26	24.80	31.08	26.21	29.00	30.31	27.06	84M	1.2 hrs	V100	
Zip-NeRF [4]	25.62	27.77	27.23	34.01	29.51	32.40	32.20	29.82	84M	7.7 hrs	V100	
NerfAcc [13]	24.26	26.89	26.31	32.35	27.39	31.84	31.80	28.69	34M	11 mins	RTX 3090	
Our VFR: 20K	24.20	26.88	26.16	34.13	29.51	32.98	32.50	29.48	34M	5.7 mins	RTX 3090	
Our VFR: 40K	24.55	27.33	26.42	34.82	30.00	33.33	33.00	29.92	34M	11 mins	RTX 3090	



Figure 5: Visual comparison of synthesized novel views on the scenes of (from top to bottom) *bicycle*, *bonsai*, and *kitchen* on the real-world 360 dataset [3].

It should be noted that we achieve this high-quality view rendering using an entirely different approach than the Zip-NeRF. As the Zip-NeRF focuses on anti-aliasing, multisampling is a reasonable solution but leads to a higher computational cost. Besides, the authors of the Zip-NeRF found a larger NN helps in improving the rendering quality. Since the Zip-NeRF still employs the standard volume rendering method, a larger NN will undoubtedly increase the training time. Our proposed VFR, however, relieves the rendering framework from the high computational cost caused by many network evaluations. As only one network evaluation is required in our VFR, NN evaluation is no longer the computing bottleneck in the rendering framework. As a result, the VFR enables one to achieve a similar rendering quality as the Zip-NeRF but with a significantly less training time.

The visual comparison in Fig. 5 shows that our VFR is able to render more realistic views compared with NerfAcc (based on Instant-NGP). Our model provides a more accurate geometry reconstruction on the scene of *bicycle*. As can be observed from Fig. 5, the rendering results of NerfAcc contain obvious visual artefacts on the scenes of *bonsai* and *kitchen*, while our results are more realistic and closer to the ground truth. The reflective effect in the highlighted patches on the scene of *kitchen* is also well modeled by our model thanks to our novel SH feature encoding method.

Table 4: Ablation study. SH means spherical harmonics for view direction encoding (VDE) while SHFE stands for SH feature encoding. We use the NerfAcc’s implementation of Instant-NGP. MLP’s size is represented by layers \times neurons. Times are measured in minutes by 20K training steps.

	Activation	MLP	VDE	NeRF synthetic				Real-world 360			
				Time	PSNR↑	SSIM↑	LPIPS↓	Time	PSNR↑	SSIM↑	LPIPS↓
A) Instant-NGP	ReLU	4 \times 64	SH	4.4	33.11	0.9614	0.053	5.6	28.39	0.8147	0.245
B) Instant-NGP	GELU	4 \times 64	SH	4.4	33.43	0.9634	0.049	5.6	28.43	0.8152	0.245
C) Instant-NGP	GELU	4 \times 128	SH	5.6	33.92	0.9663	0.045	6.8	28.70	0.8205	0.239
D) Our VFR	ReLU	4 \times 64	SH	2.9	33.17	0.9619	0.053	5.0	28.20	0.8013	0.262
E) Our VFR-GELU	GELU	4 \times 64	SH	3.0	33.61	0.9641	0.048	5.0	28.26	0.8064	0.254
F) Our VFR-larger net	GELU	6 \times 64	SHFE	3.0	34.05	0.9678	0.043	5.2	29.06	0.8215	0.240
G) Our VFR-larger net	GELU	6 \times 128	SHFE	3.0	34.33	0.9697	0.041	5.4	29.28	0.8266	0.236
H) Our VFR-SH	GELU	6 \times 256	SH	3.2	34.46	0.9701	0.040	5.5	28.91	0.8192	0.245
I) Our full VFR	GELU	6 \times 256	SHFE	3.3	34.62	0.9713	0.038	5.7	29.48	0.8301	0.233

5.3 Ablation study

Table 4 shows the ablation study of the proposed methods. Compared with the commonly used ReLU, we found that GELU activation improves the rendering quality for both the standard volume rendering (A, B) and our VFR (D, E). Larger MLP (B, C) does increase the quality using the standard volume rendering but at the expense of more training time. For our VFR, when increasing the size of the NN by using more hidden neurons (F:64 neurons, G:128, I:256), the rendering quality constantly improves but the training time only slightly increases. Replacing the SH encoding (H) with the SH feature encoding (I) also leads to a better quality with negligible extra training time.

6 Limitations

Similar to many other neural rendering methods, the proposed VFR requires per-scene optimization to deliver high-fidelity view rendering. Although our VFR can be optimized in several minutes for one scene, we are still far from realizing real-time reconstruction for 3D video applications. As for the rendering time, our method can render 4~5 frames per second at the resolution of 800 \times 800 on the NeRF synthetic dataset, similar to the rendering speed of NerfAcc. On the real-world 360 dataset, the rendering speed is about 1~2 frames per second at the resolution of 1300 \times 840.

Elaborate implementation optimization can help in achieving high-quality real-time rendering using our method. However, compared with the recent work that specifically focuses on real-time rendering such as the BakedSDF [29] and MobileNeRF [7], the rendering speed of our VFR needs to be further improved. The BakedSDF and MobileNeRF represent a scene’s surface to enable one feature querying and one neural network evaluation, at the expense of some rendering quality decline compared with volume representation. Our VFR uses the volume representation with one NN evaluation to deliver the state-of-the-art rendering quality, but the feature querying still needs to be conducted many times for a cast ray. Considering the fact that the queried feature vectors of non-empty samples are finally integrated into one feature vector using our VFR, we believe the number of feature querying can also be reduced to a small value, e.g., from 32 to 5, to achieve high-quality and real-time rendering simultaneously.

7 Conclusion

We proposed the volume feature rendering method for high-quality view synthesis. The proposed VFR integrates the queried feature vectors of non-empty samples on a ray and then transforms the integrated feature vector to the final rendered color using one single neural network evaluation. We found the VFR may fail to converge on some scenes and thus introduced the pilot network to guide the early-stage training to tackle this problem. Our proposed VFR is able to achieve the state-of-the-art rendering quality while using significantly less training time. The proposed VFR can be easily used to replace the standard volume rendering in many neural rendering methods based on volume

representation, with the benefits of either a better rendering quality by employing a larger neural network or a significantly reduced training time thanks to one single neural network evaluation.

References

- [1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.
- [2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.
- [4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-NeRF: Anti-aliased grid-based neural radiance fields. *arXiv preprint arXiv:2304.06706*, 2023.
- [5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022.
- [6] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. Factor fields: A unified framework for neural fields and beyond. *arXiv preprint arXiv:2302.01226*, 2023.
- [7] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. *arXiv preprint arXiv:2208.00277*, 2022.
- [8] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022.
- [9] Kang Han and Wei Xiang. Multiscale tensor decomposition and rendering equation encoding for view synthesis. *arXiv preprint arXiv:2303.03808*, 2023.
- [10] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021.
- [11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [12] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH*, 18(3):165–174, 1984.
- [13] Rui long Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022.
- [14] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020.
- [15] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.
- [16] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

- [17] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021.
- [18] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Proceedings of Thirty-first Conference on Neural Information Processing Systems Workshop*, 2017.
- [19] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. *arXiv preprint arXiv:2211.11610*, 2022.
- [20] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8269–8279, 2022.
- [21] Cheng Sun, Min Sun, and Hwann-Tzong 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*, pages 5459–5469, 2022.
- [22] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.
- [23] Matthew Tancik, Vincent Casser, Xincheng Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.
- [24] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.
- [25] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, 2021.
- [26] Yunxiao Wang, Yanjie Li, Peidong Liu, Tao Dai, and Shu-Tao Xia. Next: Towards high quality neural radiance fields via multi-skip transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 69–86. Springer, 2022.
- [27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [28] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021.
- [29] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P Srinivasan, Richard Szeliski, Jonathan T Barron, and Ben Mildenhall. Bakedsdf: Meshing neural sdbs for real-time view synthesis. *arXiv preprint arXiv:2302.14859*, 2023.
- [30] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.
- [31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.