

# SC-NeRF: Self-Correcting Neural Radiance Field with Sparse Views

Liang Song, Guangming Wang, Jiuming Liu, Zhenyang Fu, Yanzi Miao, and Hesheng Wang

**Abstract**—In recent studies, the generalization of neural radiance fields for novel view synthesis task has been widely explored. However, existing methods are limited to objects and indoor scenes. In this work, we extend the generalization task to outdoor scenes, trained only on object-level datasets. This approach presents two challenges. Firstly, the significant distributional shift between training and testing scenes leads to black artifacts in rendering results. Secondly, viewpoint changes in outdoor scenes cause ghosting or missing regions in rendered images. To address these challenges, we propose a geometric correction module and an appearance correction module based on multi-head attention mechanisms. We normalize rendered depth and combine it with light direction as query in the attention mechanism. Our network effectively corrects varying scene structures and geometric features in outdoor scenes, generalizing well from object-level to unseen outdoor scenes. Additionally, we use appearance correction module to correct appearance features, preventing rendering artifacts like blank borders and ghosting due to viewpoint changes. By combining these modules, our approach successfully tackles the challenges of outdoor scene generalization, producing high-quality rendering results. When evaluated on four datasets (Blender, DTU, LLFF, Spaces), our network outperforms previous methods. Notably, compared to MVSNeRF, our network improves average PSNR from 19.369 to 25.989, SSIM from 0.838 to 0.889, and reduces LPIPS from 0.265 to 0.224 on Spaces outdoor scenes.

**Index Terms**—Novel view synthesis, Generalization, Multi-view stereo, Multi-head attention.

## I. INTRODUCTION

**N**OVEL view synthesis (NVS) is a promising and long-standing problem that plays a fundamental role in both the computer vision [39], [42], robotic [40] and graphics [1].

NVS aims to capture visual information from a sparse set of reference views to render an unseen target view. Early methods [2], [3] produce a target view by interpolating in the ray [2] or pixel plane [3]. Subsequent works [4], [38] have exploited dense input views or geometric constraints, such as epipolar consistency [4], for depth-aware warping of the input

\*This work was supported by the Fundamental Research Funds for the Central Universities (Grant No.2020ZDPY0303), the General Program of National Natural Science Foundation of China (Grant No.61976218). The first two authors contributed equally. Corresponding Author: Yanzi Miao and Hesheng Wang.

L. Song, Z. Fu, and Y. Miao are with Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, School of Information and Control Engineering, Advanced Robotics Research Center, China University of Mining and Technology, Xuzhou 221116, China.

G. Wang, J. Liu, and H. Wang are with Department of Automation, Key Laboratory of System Control and Information Processing of Ministry of Education, Key Laboratory of Marine Intelligent Equipment and System of Ministry of Education, Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai Jiao Tong University, Shanghai 200240, China.



Figure 1. Comparison with previous methods IBRNet [12] and MVSNeRF [10] on Spaces. We train both their and our networks on DTU and generalize to outdoor scenes in Spaces. The left, middle, and right images respectively show the rendering result of [12], [10], and ours.

views [38]. However, these methods are susceptible to artifacts caused by occlusion, the density of input views, and inaccurate geometry. To solve this problem, the multiplane image (MPI) approaches [5], [6] offer real-time rendering and generalization capabilities by representing the scene using a set of parallel planes derived from several input images. Nevertheless, when the perspective difference between the input view and target view is significant, there may be occurrences of edge rendering overlap [37].

Recently, Neural radiance fields (NeRF) [7] and subsequent works [8], [9] have the strong ability to produce realistic new view synthesis results. However, there are two main drawbacks: 1) It requires densely captured images for each scene. 2) It needs to be trained from scratch to overfit the new scene, with no generalization to unknown scenes.

To address the aforementioned shortcomings of NeRF, many methods [11]–[14] usually build a large composite dataset to fit the network to different scenarios, including object, indoor, and outdoor scenarios. However, recent works [10], [12] can not effectively generalize to outdoor scenes when trained on only object-level datasets. When MVSNeRF [10] generalizes to an outdoor scene, black artifacts appear in the sky or border, as shown in the blue box in Figure 1. This is because the space scale and structure between the training scene and the test scene is extremely different and there maybe also exists reflective material in the outdoor scene. When the perspective gap between the input view and the target view increases, the result rendered by the IBRNet [12] method will appear blank at the boundary, as shown in the red box in Figure 1.

To solve these problems, we propose SC-NeRF, a novel approach that can be well generalized to different scenes by reconstructing radiation fields from only three unstructured multi-view input images. The SC-NeRF is trained only in an

object-level dataset, while it can be generalized to a variety of different scenarios, especially outdoor scenarios. Due to the strong generalization ability, the SC-NeRF avoids time-consuming per-scene optimization and can directly regress realistic images from novel viewpoints of outdoor scenes.

To be specific, a low-resolution 3D geometric cost volume is constructed from sparse multi-view input images. This geometric cost volume can provide continuous geometric priors, when there is a non-covisual region between the input and target perspectives. In order to solve the problem of artifacts in the rendered outdoor scene, the rendered features are corrected in terms of appearance and geometry. Specifically, a multi-head attention mechanism is leveraged to correct rendered characteristics using direction embedding as query, geometric or appearance features as key, and rendered features as value. Although it alleviates the shadow problem in the distance to some extent, it will cause shadow transfer in the render view. This is mainly because using only the direction as the query can not effectively get complete structure information of the scene. Therefore, we combine the rendered depth value with direction embedding as query, effectively solving the shadow transfer problem.

Our approach is completely differentiable, which can be trained in end-to-end manner from sparse view inputs. Our experiments show that with just three input views, our network can synthesize photo-realistic images on DTU [15], Blender [7], LLFF [16], Spaces [6]. Overall, our contributions are as follows:

- We propose a novel end-to-end network for synthesizing realistic images from sparse input views. We firstly propose a geometry correction module based on multi-head attention. It can address the issue of black artifacts in rendered views, caused by inconsistencies in scale and structure between training and testing scenes.
- Building on the geometry correction module, we also design an appearance correction module to alleviate boundary blank and ghosting artifacts in rendered views caused by relatively large viewpoint changes.
- We validate the effectiveness of our model on four datasets, including Blender, LLFF, DTU, and Spaces. Notably, on the outdoor scenes in the Spaces dataset, our model outperforms MVSNeRF by 34.17% in terms of PSNR, and IBRNet by 19.9%.

## II. RELATED WORK

### A. Novel View Synthesis via NeRF

In recent years, various neural scene representations have been proposed to implement view synthesis [7], [17], [18], [41], [43]. NeRF [7] has achieved very impressive results in novel view synthesis by optimizing the 5D neural radiation field of a scene. However, it must be optimized for each new scenario, which takes hours or days to converge.

There are some methods proposed to extend NeRF's generalization capabilities [10]–[13], [19]. GRF [11] projects the learned local image features onto three-dimensional points to obtain a general and rich point representation. MVSNeRF [10] utilizes plane sweep cost volume for neural radiation

field reconstruction. NeuRay [14] enables the construction of radiation fields to focus on visible image features by modeling the visibility of 3D points in the input view. However, none of these methods consider how to train the network only on an object-level dataset and be generalized to outdoor scenes.

### B. Multi-View Stereo

Multi-view stereo (MVS) is a core problem in the field of computer vision. Multi-view stereo matching reconstruction can be regarded as the inverse process of taking pictures of a certain scene. Its purpose is to restore the real 3D scene through images taken from different viewpoints. A large number of traditional methods [20]–[24] use hand-crafted similarity metrics and regularization methods to calculate dense correspondence of scenes. These methods can achieve good results on non-Lambertian surfaces and scenes without weakly textured regions. However, the artificially designed similarity metrics become unreliable in weakly textured regions, thus leading to incomplete reconstruction results. Recently, deep learning techniques [25]–[28] have been introduced. Among these, MVSNet [25] applies a 3D CNN for depth estimation on the plane scan cost of the reference view, and achieves high-quality 3D reconstruction. Subsequent works [26]–[28] extend this technique to recurrent planar sweeps [26], point-based densification [27], and cascaded cost volume [28] for improving the effect of reconstruction. We follow their ideas to build a geometrically consistent cost volume. This ensures that the network meets the consistency of multiple views, so that the network can focus on information from different views and also learn geometric priors when rendering the novel views.

### C. Transformer in NeRF

Recently, there have been some attempts to incorporate the transformer [29] architecture into the NeRF model. IBRNet [12] proposes ray transformers that dynamically correlate appearance information from multiple source views. NerFormer [30] proposes to use transformers to aggregate features between different views on the ray and learn radiance fields from the aggregated features. GNT [31] directly regresses colors to synthesize views without need for NeRF's volumetric rendering. GPNR [32] improves generalization by using several stacked "patch-based" transformers to aggregate global features. Different from the above methods, our method mainly uses transformers to correct the features for better generalization to different scenes. Specifically, we use the transformer to correct the geometric and appearance features, which make better use of information from different perspectives and improve the generalization ability of network and geometric reconstruction.

## III. METHOD

Given several sparse source views, our method uses volume rendering to synthesize a target view in a new camera pose. The core problem is how to obtain the density and colors of the continuous space by using the information from the input views, and how to make this representation generalize to other scenes, especially outdoor scenes.

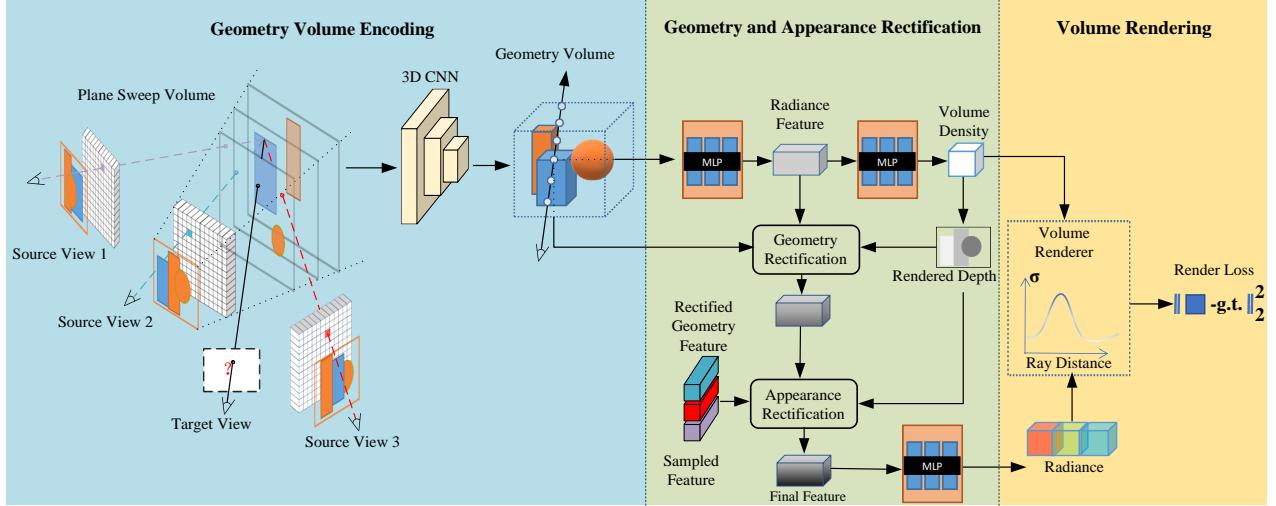


Figure 2. Overview of SC-NeRF. We first extract image features and warp them onto a plane sweep, then use 3DCNN to build a geometric Volume. Second, we use the geometric feature rectification and appearance feature rectification modules to obtain the final radiation features. Finally, we use an MLP to regress the volume density and RGB radiance values of any sampling point in space, and use volume rendering to obtain the rendered view.

The overview of our SC-NeRF is shown in Fig. 2. For the sparse input  $M$  views ( $M = 3$ ), we first warp the extracted image features into the reference perspective and construct a geometric encoding volume using 3DCNN (Sec. III-A). Then, we obtain the final radiance features through the geometric feature correction module and the appearance feature correction module based on the multi-head attention mechanism (Sec. III-B). Finally, we use an multi-layer perceptron (MLP) to regress the volume density and RGB radiance from the corrected radiance features. These volume properties are passed through the volume rendering formula to obtain the final rendered images (Sec. III-C).

### A. Geometry Volume Encoding

Inspired by the recent MVSNeRF [10], we construct the encoding volume  $V$  at the reference view, allowing for geometry-aware scene understanding.

First of all, a 2D CNN  $G_1$  is used to extract the local appearance features of the input images. In our network, each input image  $I_i \in R^{H_i \times W_i \times 3}$  is converted into a 2D feature map  $F_i \in R^{H_i/4 \times W_i/4 \times C_1}$  by a down-sampled convolution operation:

$$F_i = G_1(I_i), \quad (1)$$

where  $H_i$  and  $W_i$  are the image height and width, and  $C_1$  is the number of image feature channels.

Then, we transform the features of the source view into the reference view by the homographic warping operation. Given the camera intrinsic [ $K$ ] and extrinsic parameters [ $R, T$ ], we use the homographic warping:

$$H_i(z) = K_i \cdot (R_i \cdot R_r^T + \frac{(t_r - t_i) \cdot n_r}{z}) \cdot K_r^{-1}, \quad (2)$$

where  $H_i$  is the matrix warping from the view  $i$  to the reference view  $r$  at depth  $z$ .  $K_i$  and  $K_r$  are the intrinsic matrices.  $n_r$  denotes the unit normal vector.  $R$  and  $t$  are the

camera rotation and translation matrices. Each feature map  $F_i$  can be warped to the reference view by:

$$F_{i,z}(u, v) = Pad(F_i)(H_i(z)[u, v, 1]^T), \quad (3)$$

where  $F_{i,z}$  is the warped feature map at depth  $z$ , and  $(u, v)$  represents a pixel location in the reference view.  $Pad$  indicates the feature image edge padding operation. In this work, we parameterize  $(u, v, z)$  using the normalized device coordinate (NDC) at the reference view.

We leverage the variance-based method [10] to compute the cost from the warped feature maps on the  $D$  sweeping planes. In particular, for each position  $P$ , its cost feature vector is computed by:

$$P(u, v, z) = Var(F_{i,z}(u, v)), \quad (4)$$

where  $Var$  is the variance operation.

Finally, we use a 3D CNN network  $G_2$  with a U-Net structure to encode the cost volume mentioned above. This process is expressed by:

$$V = G_2(P), \quad (5)$$

where  $V$  is encoding geometry volume. This encoded volume contains the geometry feature of the scene, and is later continuously interpolated and converted into volume density.

### B. Geometric and Appearance Features Rectification

Black artifacts appear in the rendering result of outdoor view, which are caused by the following reasons: 1) The spatial range of the object-level training set is much smaller than that of the outdoor test scenes. 2) There are some non-Lambertian reflective objects in the outdoor scene, which lead to the deviation of the feature. At the same time, when the viewpoint changes drastically, there may be significant differences between multiple viewpoints, resulting in incomplete coverage of details and textures on the object surface in a single viewpoint image. These uncovered details and textures

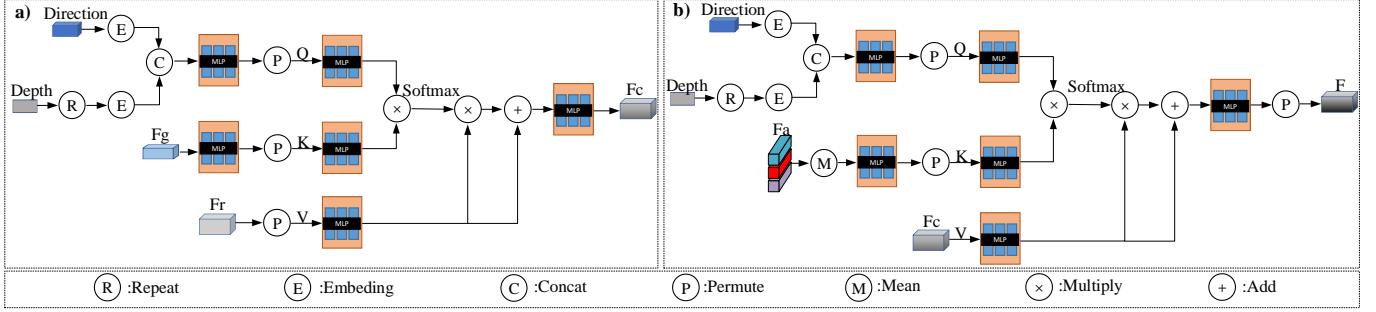


Figure 3. Geometry and appearance rectification modules. a) For the appearance features rectification module, direction embedding and depth embedding be used as query, sampled volume features  $F_g$  are as key, the radiance features  $F_r$  is taken as a value. b) For the appearance features rectification module, we also use the direction embedding and depth embedding as the query, the sampled image feature  $F_a$  as the key, and the corrected feature  $F_c$  as the value, so as to build a multi-head attention mechanism.

can lead to visual artifacts, such as ghosting and boundary blanking. To overcome these challenges, we design geometric and appearance feature rectification modules based on multi-head attention mechanisms.

1) *Geometric feature Rectification*: Given an arbitrary 3D location  $x$ , an MLP  $M_1$  be used to obtain radiance features  $F_r$ ,

$$F_r = M_1(E(x), s), \quad (6)$$

where  $s$  is the neural feature trilinearly interpolated from the volume  $V$  at the location  $x$ .  $E(\cdot)$  indicates embedding operation. Then, the corresponding volume density  $\sigma$  is regressed by an MLP  $M_2$ ,

$$\sigma = M_2(F_r). \quad (7)$$

We obtain the rendered depth  $\hat{D}$  of the pixel corresponding to the  $k$  sampled points based on the volume rendering formula:

$$\hat{D} = \sum_{k=1}^N T_k(1 - \exp(-\sigma_k))z_k, \quad (8)$$

$$T_k = \exp\left(-\sum_{j=1}^{k-1} \sigma_j\right). \quad (9)$$

Our geometric feature correction module is shown in Fig.3 (a). First, we normalize the rendered depth and get depth embedding. At the same time, we do the same operation with the ray direction. An MLP  $M_3$  is used to process depth embedding and direction embedding to obtain query values  $Q$ :

$$Q = M_3(E(\hat{D}) \oplus E(d)), \quad (10)$$

where  $\oplus$  is concat operation,  $d$  is direction vector. It is worth noting that if we only use direction embedding as the query value, the rendered view will have black shadows and white holes.

Then, the radiance features  $F_r$  serve as the value  $V$  and the volume features  $s$  as the key  $K$ . The matching matrix of attention is calculated in the sampling point channel rather than the feature channel. The reason for doing this is to allow the model to independently learn which depth sample points contribute more to the rendering. This enables assigning higher weights to these depth sample points. By focusing on assigning weights to individual points rather than features for

each sample point, we can better capture the informative points and improve the rendering quality. Therefore, we choose to pay attention in the sampling point dimension and calculate the attention weight for each sampling point to overcome the above problem.

Finally, we generate corrected features  $F_c$  through multiple attention mechanisms:

$$F_c = \text{Multihead}(Q, K, V), \quad (11)$$

$$\text{Multihead}(Q, K, V) = (\text{head}_1 \oplus \dots \oplus \text{head}_h)W^O \quad (12)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (13)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (14)$$

where  $QW_i^Q$  represents the weight matrix of the query ( $Q$ ) for the  $i$ -th attention head,  $KW_i^K$  represents the weight matrix of the key ( $K$ ) for the  $i$ -th attention head,  $VW_i^V$  represents the weight matrix of the value ( $V$ ) for the  $i$ -th attention head.  $W^O$  is the feature matrix,  $\oplus$  is contact operation. This corrected feature can adapt to the change of scene space.

2) *Appearance feature Rectification*: Considering that large changes in viewpoint can lead to boundary blanking and artifacts in rendering, we also correct the appearance features through the multi-head attention mechanism.

We reproject the pixel feature  $F_i$  back onto the sample points along the light. We preliminarily regard pixel features as the appearance features of each position in the three-dimensional space along the direction of light. With this formula, each 3D point can theoretically have the corresponding 2D appearance feature. Specifically, given a three-dimensional point  $x$ , the observed 2D image  $I_i$  with camera intrinsics  $K$  and camera pose  $\xi$ , the corresponding 2D appearance feature  $F_a$  can be retrieved through the following reprojection operation:

$$F_a = \pi(\{(F_i \oplus I_i), \xi, K\}, \{x, y, z\}), \quad (15)$$

where the function  $\pi(\cdot)$  follows the principle of multi-view geometry [33]. If the point is inside the image, we simply select the nearest pixel using bilinear interpolation and index its features for the 3D point. If the point is outside the image, we assign a zero vector to the 3D point, which means there is no information observed.

Table I

QUANTITATIVE RESULTS OF NOVEL VIEW SYNTHESIS AT SYNTHETIC AND REAL DATA. WE SHOW AVERAGED RESULTS OF PSNRs, SSIMs AND LPISS ON THREE DIFFERENT DATASETS [7], [16], [25]. WE COMPARE OUR METHOD WITH CONCURRENT NERF-BASED METHODS [10], [12], [13].

Model	Synthetic Data(Blender [7])			Real Data(DTU [15])			LLFF [16]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
PixelNeRF [13]	7.39	0.658	0.411	19.31	0.789	0.382	11.24	0.486	0.671
IBRNet [12]	22.44	0.874	0.195	26.04	0.917	0.190	21.79	0.786	0.279
MVSNeRF [10]	23.62	0.897	0.176	26.63	0.931	0.168	21.93	0.795	<b>0.252</b>
Ours	<b>23.753</b>	<b>0.90</b>	<b>0.174</b>	<b>26.834</b>	<b>0.934</b>	<b>0.167</b>	<b>22.087</b>	<b>0.80</b>	0.253

Table II

QUANTITATIVE RESULTS OF NOVEL VIEW SYNTHESIS AT OUTDOOR DATA. WE EVALUATED THE PSNR, SSIMS, LPIPS INDICATORS UNDER DIFFERENT DIFFICULTY LEVEL SETTINGS OF THE SPACES DATASET [6]. WE COMPARE OUR METHOD WITH CONCURRENT NERF-BASED METHODS [10], [12].

Model	Small			Medium			Large		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
IBRNet [12]	21.667	0.844	0.297	21.002	0.824	0.311	20.346	0.797	0.349
MVSNeRF [10]	19.369	0.838	0.265	19.196	0.797	0.310	19.604	0.772	0.336
Ours	<b>25.989</b>	<b>0.889</b>	<b>0.224</b>	<b>24.962</b>	<b>0.856</b>	<b>0.260</b>	<b>24.584</b>	<b>0.843</b>	<b>0.272</b>

As in the geometric rectification section, we also use direction embedding and depth embedding as query values  $Q$ . In order to make better use of the appearance information of each source perspective, we calculate the mean of the appearance features of the three samples.

We use the mean appearance feature as the key  $K$  and the correction feature  $F_c$  as the value  $V$ . We use formula (11) to obtain the final corrected radiance feature  $F$ . This process allows our network to use the input source view appearance feature to correct the appearance feature  $F_a$  of the render view.  $F$  can effectively adapt to both the appearance of the scene and the geometry of the scene. Finally we use an MLP  $M_4$  decoding structure similar to NeRF to get radiance value  $c$ :

$$c = M_4(F, E(\hat{D}), E(d)). \quad (16)$$

### C. Rendering and Training

The method described in the previous sections generates corrected radiance  $c$  and density  $\sigma$  values. To render the color  $\hat{rgb}$  of ray through the scene, we first query the color and density of  $N$  samples on the ray, and then accumulate the color and density along them:

$$\hat{rgb} = \sum_{k=1}^N T_k (1 - \exp(-\sigma_k)) c_k, \quad (17)$$

$$T_k = \exp\left(-\sum_{j=1}^{k-1} \sigma_j\right), \quad (18)$$

where  $\hat{rgb}$  is the final pixel color output, and  $T_k$  represents the volume transmittance.

This volume rendering is completely differentiable, so SC-NeRF can regress the final pixel color at the target view point from the sparse input views in an end-to-end manner. We use the  $\ell_2$  norm of the rendered pixel versus the real pixel as a loss.

$$L = \|rgb - \hat{rgb}\|_2^2, \quad (19)$$

Table III

QUANTITATIVE RESULTS OF NOVEL VIEW SYNTHESIS AT OUTDOOR DATA. WE EVALUATED THE PSNR, SSIMS, LPIPS INDICATORS UNDER DIFFERENT DIFFICULTY LEVEL SETTINGS OF THE DTU DATASET [15]. WE COMPARE OUR METHOD WITH CONCURRENT NERF-BASED METHODS [10], [12], [13].

Method	Abs err↓	Acc (0.01)↑	Acc (0.05)↑
PixelNeRF [13]	0.239	0.039	0.187
IBRNet [12]	1.62	0.000	0.001
MVSNeRF [10]	0.035	0.717	0.866
Ours	<b>0.022</b>	<b>0.770</b>	<b>0.914</b>

where  $rgb$  is the ground truth pixel color sampled from the target image  $I_t$  at a novel viewpoint.  $\|\cdot\|_2^2$  denote the  $\ell_2$  norm.

## IV. EXPERIMENTS

We evaluate our SC-NeRF on four datasets, namely DTU [15], Blender [9], LLFF [16], and Spaces [6]. In addition, we conduct the ablation studies of our network to demonstrate the effectiveness of each component. Extensive experiments show that our method exceeds the current state-of-the-art methods, especially when generalized to outdoor scenes.

### A. Experimental Settings

1) *Dataset and evaluation settings:* We train our network only on DTU dataset [15]. We follow the MVSNeRF [10] data partitioning method and divide the data into 88 training scenes and 16 testing scenes. We also evaluate our method on additional synthetic [9] and real datasets [16], following MVSNeRF. In order to evaluate the generalization ability of the network to outdoor scenes, we also select corresponding scenes from the Spaces dataset [6] for evaluation. In particular, as shown in Figure 5, we select 8 outdoor scenes in the Spaces dataset which contains 100 scenes captured by a 16-camera rig, and set three different levels of difficulty through the spatial gap between the source and target views. The performance is evaluated by PSNR, SSIM [34] and LPIPS [35] metrics.

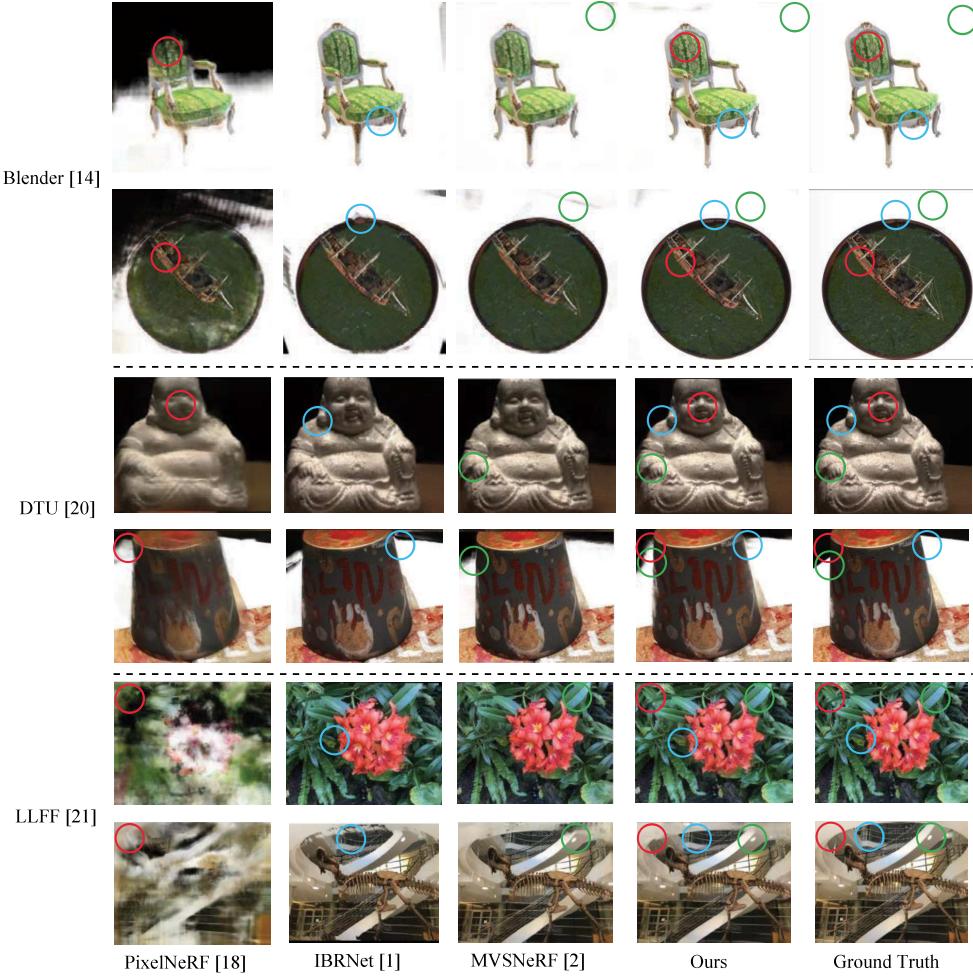


Figure 4. Rendering quality comparison at object level and indoor data. We show the visual comparison results of our method and other NeRF-based generalization methods [10], [12], [13] on 3 different test sets [7], [15], [16]. For each data set, we select two sets of scenarios to show. From the red circle, it can be observed that PixelNeRF [13] has a poorer rendering effect. From the blue circle, it can be seen that IBRNet [12] lacks sufficient detail in handling edge details. From the green circle, it can be noticed that MVSNeRF [10] is slightly inferior in rendering background details.

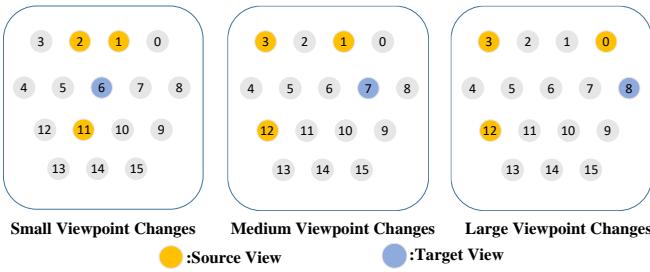


Figure 5. Three levels of viewpoint settings with increasing difficulty. We build three levels of difficulty rendering based on the gap between the source view and the target view. The orange ones represent the input 3 source views, and the blue ones represent the target view to be rendered. We call the different difficulty levels as "small", "medium" and "large" respectively.

2) *Implement Details*: The dimension of image features  $F_i$  extracted by 2D CNN is set as 32. The depth sampling planes for the homographic warping operation are 128. The volume feature  $V$  dimension of 3D CNN encoding is set as 8. The number of sampling points for each ray is 128. The dimensions of position embedding, orientation embedding  $E(\hat{d})$  and depth

embedding  $E(\hat{D})$  are 63, 33 and 11, respectively. The number of heads of multi-head attention used for appearance correction and geometry correction is 4. All training and evaluation experiments are conducted on a single RTX 3090 GPU with PyTorch1.10.1. We randomly select 1024 pixels from a novel viewpoint as a batch and apply the Adam optimizer with an initial learning rate of 0.0005.

### B. Comparison Results

We compare with three recent NeRF-based works, PixelNeRF [13], IBRNet [12], and MVSNeRF [10] that also aim to improve generalization ability of NeRF. We input the three source views to retrain the three models on the DTU data for the fair comparison. We choose four groups of images in each scene from the three datasets [7], [16], [25] for testing, and finally evaluate the performance with the mean PSNR, SSIM, and LPIPS. We show the quantitative results in Tab. I. To further compare the generalization capability of our network in outdoor environments, we compare ours with the MVSNeRF and IBRNet in Tab. II. For a more intuitive comparison of experimental effects, we show visualization comparison in Fig.

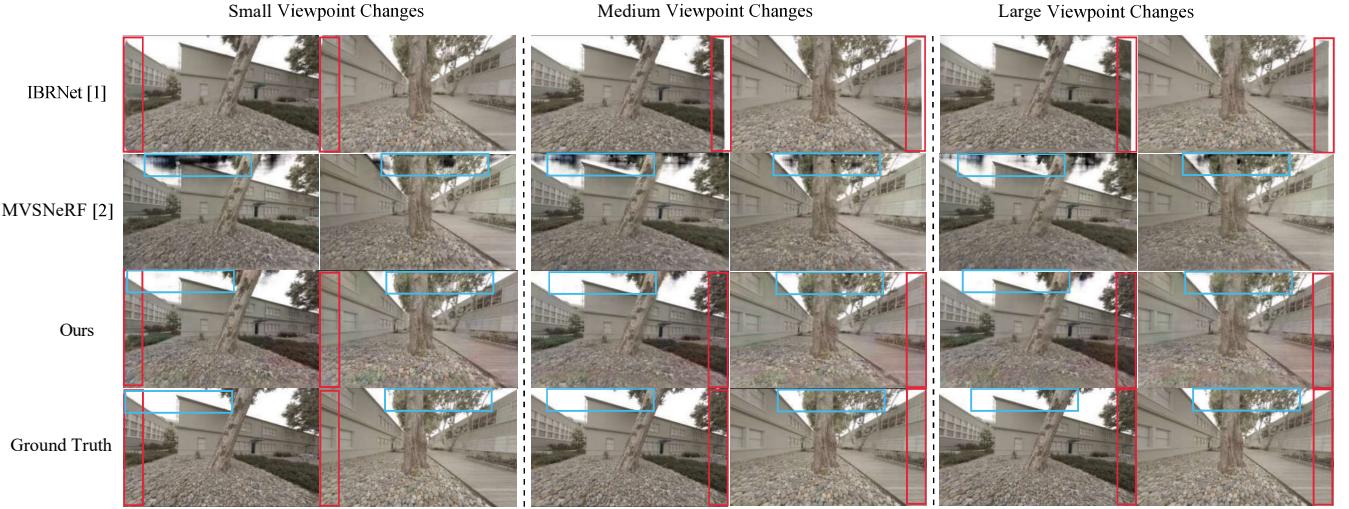


Figure 6. Rendering quality comparison at outdoor data. We show the results of a visual comparison of our method with two state-of-the-art methods [10], [12] on the Space dataset [6] for three settings of different difficulty levels. From the red box, it can be observed that as the viewing angle increases, IBRNet [12] produces blank spaces at the edges of the rendered image. From the blue box, it can be seen that MVSNeRF [10] exhibits black pseudo-shadows in areas such as the sky. Our method can effectively solve these issues in outdoor scenes.

Table IV

QUALITATIVE ABALATION STUDY. WE EVALUATED THE PSNR, SSIMS, LPIPS INDICATORS UNDER DIFFERENT DIFFICULTY LEVEL SETTINGS OF THE SPACES DATASET [6]. WE COMPARE THE PERFORMANCE OF THE DIFFERENT COMPONENTS OF OUR METHOD.

Model	Small			Medium			Large		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline(BL)	19.369	0.838	0.265	19.196	0.797	0.310	19.604	0.772	0.336
BL+Appearance-V(A_V)	23.037	0.806	0.393	22.9	0.796	0.400	23.107	0.869	0.439
BL+Appearance-VD(A_VD)	23.894	0.836	0.310	23.338	0.808	0.332	23.148	0.798	0.341
BL+Geometry-V(G_V)	19.175	0.822	0.335	18.706	0.799	0.359	18.582	0.787	0.370
BL+Geometry-VD(G_VD)	25.528	0.880	0.241	24.220	0.825	0.298	23.384	0.810	0.317
BL+Appearance+Geometry(AG_VD)	24.145	0.845	0.294	23.374	0.789	0.335	22.469	0.772	0.353
<b>BL+Geometry+Appearance(GA_VD)</b>	<b>25.989</b>	<b>0.889</b>	<b>0.224</b>	<b>24.962</b>	<b>0.856</b>	<b>0.260</b>	<b>24.584</b>	<b>0.843</b>	<b>0.272</b>

4 and Fig. 6. We also present the comparative results of the depth estimates at DTU dataset in Tab. III.

1) *Comparisons of view synthesis at synthetic and indoor data:* Quantitative results in Tab. I show that our SC-NeRF performs the best in all datasets. Although our model is only trained on DTU, it can be well generalized to other two datasets with highly different distributions of scenes and views. On the Blender, DTU, and LLFF datasets, the PSNR evaluation values are higher than those of PixelNeRF by 16.363, 7.524, and 10.847 respectively. They are also higher than IBRNet by 1.313, 0.794, and 0.297 respectively, and higher than MVSNeRF by 0.133, 0.204, and 0.157 respectively. As shown in Fig. 4, PixelNeRF has obvious blurring and artifacts when generalized to other scenes. This is because they only consider introducing 2D image features into the NeRF model, but don't consider the scene geometry. IBRNet has achieved well generalization results due to the introduction of rays transformer, but some artifacts still appear in the details. The view rendered by MVSNeRF tends to contain artifacts around the background because its cost volume is built for a specific reference view where the camera setback may not cover the target view sufficiently. The main reason for the superiority of our model is that we not only consider the geometric features of the scene, but also its appearance features, and

use the rectification-based strategy to make the two features mutually optimized. This correction mechanism can improve the rendering effect in indoor scenes for better generalization.

2) *Comparisons of view synthesis at outdoor data:* To evaluate the ability to generalize to outdoor scenes, we test IBRNet, MVSNeRF, and our model on the Spaces dataset with three difficulty level settings. From Tab. II, it can be clearly found that the generalization to outdoor scenes has a significant decline in the performance of their method compared with the test scene in the object level and indoor scenes. This can be seen in the degradation of their method's performance when generalized to outdoor scenarios. However, our method can achieve the best performance. On the three difficulty levels (small, medium, large) of the Spaces dataset, the PSNR obtained from testing is higher than IBRNet by 20%, 18.9%, and 20.1% respectively. It is also higher than MVSNeRF by 34.2%, 30%, and 25.4% respectively. From Fig. 6, we can analyze the reasons for the bad generalization of their two methods to outdoor scenes. Due to the increasing gap between the input view of the scene and the rendered target view, it can be found that IBRNet can not render the non-common view regions. Therefore, the rendered view will have a rendering blank area in the border outline. From the second line of Fig. 6, it can be found that MVSNeRF will have artifacts

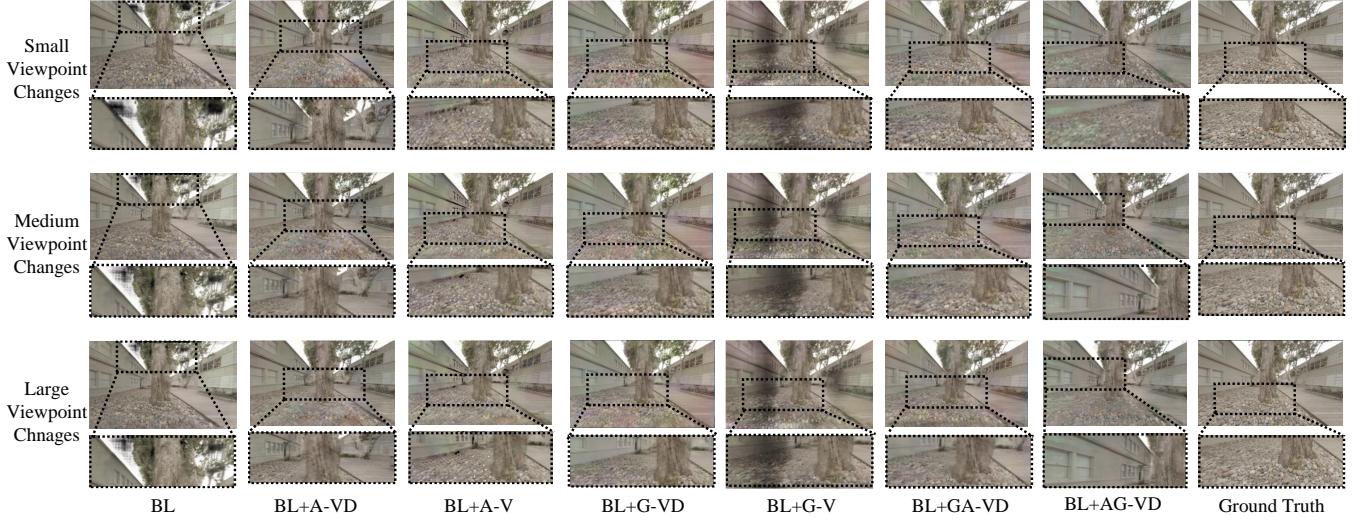


Figure 7. Qualitative ablation study. We show the visualization results of adding geometry correction module and appearance correction module on Baseline. We mark the area of our concern with a box to highlight problems such as rendering holes and artifact transfer.

in the sky. This is because of a huge difference in scene depth between the training and testing set. Differently, we can effectively alleviate the depth inconsistency between the training and testing set by normalizing the rendered depth and embedding it as a part of the query value. At the same time, we use the appearance correction strategy to effectively use the appearance characteristics of different views to overcome the problem of rendering loss in non-common view areas. It can be seen from both qualitative results and quantitative metrics that our model outperforms the two state-of-the-art methods on the above comparison in outdoor scene rendering performance.

### C. Comparisons of Depth Reconstruction

In order to evaluate whether the model effectively learns the ability to model the geometry of the scene. We reconstruct the depth as [9] by weighting the depth values of the sampling points on the ray and the volume density. We compare our approach with three NeRF-based methods [10], [12], [13]. It can be seen from the Tab. III that our method achieves the best results of the estimation of the depth of the new view. It can be observed that the absolute error obtained from the tests is reduced by 90.8% compared to PixelNeRF, by 98% compared to IBRNet, and by 37% compared to MVSNeRF. Since only the local features of the image are used and the geometric structure of the scene is not considered, the rendering depth of PixelNeRF has 20 times larger errors. It is worth noting that although IBRNet can render the target view well and has strong generalization ability, it does not learn the 3D model of the scene in essence, but only an interpolation synthesis in appearance. So it suffers from extremely poor depth estimations. Due to the geometric correction strategy adopted by our method, our method can outperform MVSNeRF's depth estimation metrics.

### D. Ablations and Analysis

Tab. IV and Fig. 7 summarize the quantitative and qualitative results of SC-NeRF for different architecture choices at different difficulty levels on the Spaces dataset. We take MVSNeRF as our baseline. We will only add the appearance correction module with orientation embedding as query value on the baseline called "Appearanc-V". "Appearance-VD" means adding an appearance correction module with orientation embedding and rendering depth embedding as query values. Similarly, for the geometry correction module, we distinguish them by different query values, namely "Geometry-V" and "Geometry-VD".

From Tab. IV, we find that when only using direction embedding as the query, the geometric rectification will make the indicators decrease, while the appearance rectification can effectively improve the PSNR. This shows that direction embedding as a query can effectively correct appearance characteristics. When we combine depth embedding with direction embedding as a multi-head attention query, we find that it can greatly improve the performance of geometric rectification, and also promote the appearance rectification. This shows that providing geometric depth information of the scene can effectively improve the performance of novel view rendering.

It can be seen from Fig. 7, when only the direction is used as the query value, white holes will appear in appearance rectification, and black shadow transfer will appear in geometric rectification. However, when the depth embedding is also used as the query value, it can solve the two problems of the above.

Only using the appearance rectification module can improve the PSNR metrics very well, but the structural index SSIM will be lower than the Baseline. On the contrary, if only the geometric correction module is used, the SSIM and LPIPS indicators can be improved very well.

By comparing the sequence of geometric correction and appearance correction, we can find that the feature planning

and correction before the appearance correction of the feature can obtain better rendering performance. Therefore, our final model firstly corrects geometric features, and then appearance features.

## V. CONCLUSIONS

We propose a new generalizable approach for neural rendering. It provides a more practical neural rendering technique using a small number of images as input. Through the proposed geometric feature correction module and appearance feature correction module, our network can be trained on only object-level scenes to effectively generalize to outdoor scenes. We show that our rectification strategy can provide valuable geometry and appearance cues, leading to state-of-the-art performance under several challenging settings on four benchmark datasets.

## REFERENCES

- [1] K. N. Ngan, T. Sikora, and M.-T. Sun, “Visual communications and image processing 2000,” *Visual Communications and Image Processing 2000*, vol. 4067, 2000.
- [2] M. Levoy and P. Hanrahan, “Light field rendering,” in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 31–42.
- [3] S. Chen and L. Williams, “View interpolation for image synthesis proceedings of the 20th annual conference on computer graphics and interactive techniques,” 1993.
- [4] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, “Unstructured lumigraph rendering,” in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 425–432.
- [5] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, “Stereo magnification: Learning view synthesis using multiplane images,” in *SIGGRAPH*, 2018.
- [6] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker, “Deepview: View synthesis with learned gradient descent,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2367–2376.
- [7] 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.
- [8] L. Liu, J. Gu, K. Zaw Lin, T.-S. Chua, and C. Theobalt, “Neural sparse voxel fields,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 651–15 663, 2020.
- [9] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, “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.
- [10] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, “Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 124–14 133.
- [11] A. Trevithick and B. Yang, “Grf: Learning a general radiance field for 3d representation and rendering,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 182–15 192.
- [12] 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.
- [13] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, “pixelnerf: 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] 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.
- [15] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs, “Large scale multi-view stereopsis evaluation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 406–413.
- [16] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, “Local light field fusion: Practical view synthesis with prescriptive sampling guidelines,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.
- [17] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” *Acm Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.
- [18] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, “Neural volumes: Learning dynamic renderable volumes from images,” *ACM Trans. Graph.*, vol. 38, no. 4, jul 2019. [Online]. Available: <https://doi.org/10.1145/3306346.3323020>
- [19] J. Chibane, A. Bansal, V. Lazova, and G. Pons-Moll, “Stereo radiance fields (srfs): Learning view synthesis for sparse views of novel scenes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7911–7920.
- [20] J. S. De Bonet and P. Viola, “Poxels: Probabilistic voxelized volume reconstruction,” in *Proceedings of International Conference on Computer Vision (ICCV)*, vol. 2, 1999, p. 3.
- [21] C. H. Esteban and F. Schmitt, “Silhouette and stereo fusion for 3d object modeling,” *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 367–392, 2004.
- [22] Y. Furukawa and J. Ponce, “Accurate, dense, and robust multiview stereopsis,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2009.
- [23] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, “Pixelwise view selection for unstructured multi-view stereo,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 501–518.
- [24] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, “A comparison and evaluation of multi-view stereo reconstruction algorithms,” in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*, vol. 1. IEEE, 2006, pp. 519–528.
- [25] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, “Mvsnet: Depth inference for unstructured multi-view stereo,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.
- [26] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, “Recurrent mvsnet for high-resolution multi-view stereo depth inference,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5525–5534.
- [27] R. Chen, S. Han, J. Xu, and H. Su, “Point-based multi-view stereo network,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1538–1547.
- [28] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, “Cascade cost volume for high-resolution multi-view stereo and stereo matching,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2495–2504.
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [30] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, “Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 901–10 911.
- [31] M. V. T. P. Wang, X. Chen, T. Chen, S. Venugopalan, and Z. Wang, “Is attention all that neRF needs?”, in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: <https://openreview.net/forum?id=xE-LtsE-xx>
- [32] M. Suhail, C. Esteves, L. Sigal, and A. Makadia, “Generalizable patch-based neural rendering,” in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 2022, pp. 156–174.
- [33] A. M. Andrew, “Multiple view geometry in computer vision,” *Kybernetes*, vol. 30, no. 9/10, pp. 1333–1341, 2001.
- [34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

- [36] S. Chen, J. Li, Y. Zhang, and B. Zou, "Improving neural radiance fields with depth-aware optimization for novel view synthesis," *arXiv preprint arXiv:2304.05218*, 2023.
- [37] K. Zhou, W. Li, Y. Wang, T. Hu, N. Jiang, X. Han, and J. Lu, "Nerflix: High-quality neural view synthesis by learning a degradation-driven inter-viewpoint mixer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 363–12 374.
- [38] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3703–3712.
- [39] G. Wang, X. Wu, S. Jiang, Z. Liu, and H. Wang, "Efficient 3d deep lidar odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5749–5765, 2023.
- [40] G. Wang, C. Zhang, H. Wang, J. Wang, Y. Wang, and X. Wang, "Unsupervised learning of depth, optical flow and pose with occlusion from 3d geometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 308–320, 2020.
- [41] Y. Xu, X. Xu, J. Jiao, K. Li, C. Xu, and S. He, "Multi-view face synthesis via progressive face flow," *IEEE Transactions on Image Processing*, vol. 30, pp. 6024–6035, 2021.
- [42] K. Lv, H. Sheng, Z. Xiong, W. Li, and L. Zheng, "Pose-based view synthesis for vehicles: A perspective aware method," *IEEE Transactions on Image Processing*, vol. 29, pp. 5163–5174, 2020.
- [43] A. P. S. Kohli, V. Sitzmann, and G. Wetzstein, "Semantic implicit neural scene representations with semi-supervised training," in *2020 International Conference on 3D Vision (3DV)*, 2020, pp. 423–433.



**Liang Song** received the B.S. degree from Xuzhou University of Technology, Xuzhou, China, in 2021. He is currently pursuing the master degree in China University of Mining and Technology. His current research interests include computer vision and SLAM, in particular, Neural Radiance Field.



**Guangming Wang** received the B.S. degree from Department of Automation from Central South University, Changsha, China, in 2018. He is currently pursuing the Ph.D. degree in Control Science and Engineering with Shanghai Jiao Tong University. His current research interests include SLAM and computer vision, in particular, neural radiance fields and neural rendering.



**Jiuming Liu** received the B.S. degree from Department of Automation from Harbin Institute of Technology, Harbin, China, in 2022. He is currently pursuing the master degree in Control Science and Engineering with Shanghai Jiao Tong University. His current research interests include SLAM and computer vision, in particular, Neural Radiance Field, and transformer.



**Zhenyang Fu** received the B.S. degree from Department of Automation from China University of Mining and Technology, Xuzhou, China, in 2023. His current research interests include intelligent visual perception and computer vision, in particular, 3D scene reconstruction and intelligent robots.



**Hesheng Wang** received the B.Eng. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2002, and the M.Phil. and Ph.D. degrees in automation and computer-aided engineering from The Chinese University of Hong Kong, Hong Kong, in 2004 and 2007, respectively. He is currently a Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. His current research interests include visual servoing, service robot, computer vision, and autonomous driving. Dr. Wang is an Associate Editor of IEEE Transactions on Automation Science and Engineering, IEEE Robotics and Automation Letters, Assembly Automation and the International Journal of Humanoid Robotics, a Technical Editor of the IEEE/ASME Transactions on Mechatronics. He served as an Associate Editor of the IEEE Transactions on Robotics from 2015 to 2019. He was the General Chair of the IEEE RCAR 2016, and the Program Chair of the IEEE ROBIO 2014 and IEEE/ASME AIM 2019.



**Yanzi Miao** received the Ph.D. degree in control science and engineering in 2009 from the China University of Mining and Technology, Xuzhou, China. As being a joint-PhD candidate and a visiting scholar, she worked in the Department of Informatics, University of Hamburg, Germany, in 2007 and 2017, respectively. Her is currently a Professor at the School of Information and control Engineering, China University of Mining and Technology. Her current research interests include Intelligent Perception and Fusion, Machine Vision and Active Olfaction. She has served as the Technical Co-Chair of IEEE RCAR 2019.