

# CMC: Few-shot Novel View Synthesis via Cross-view Multiplane Consistency

Hanxin Zhu\*

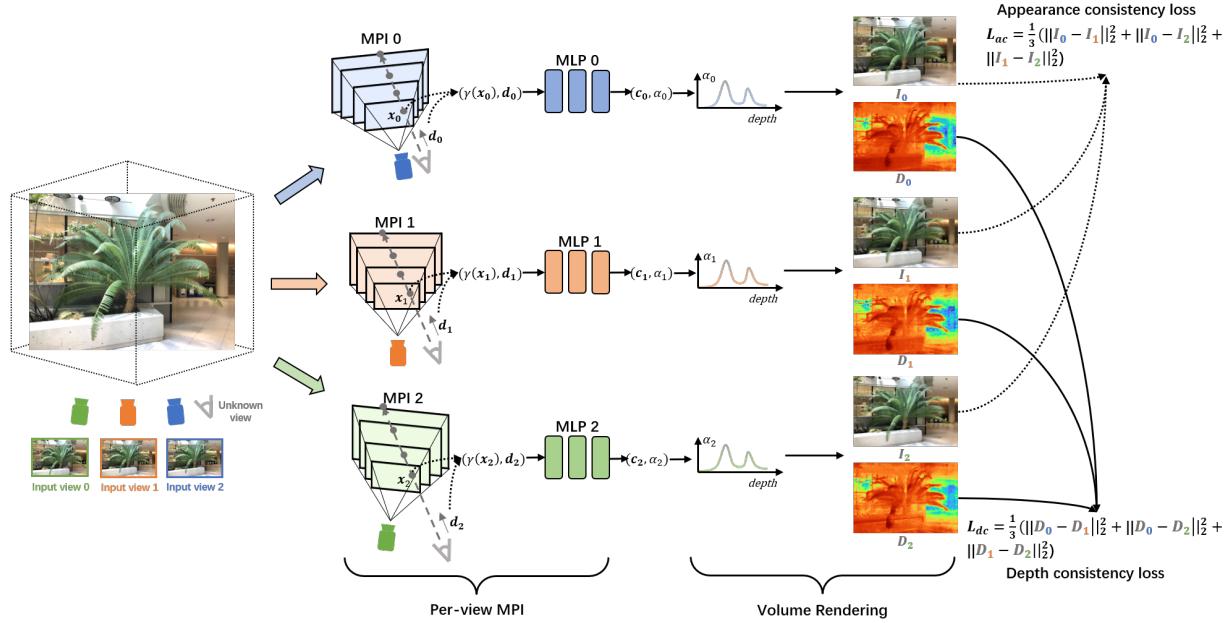
University of Science and Technology of China

Tianyu He<sup>†</sup>

Microsoft Research Asia

Zhibo Chen<sup>‡</sup>

University of Science and Technology of China



**Figure 1: Overview of our proposed method.** Given sparse input view images, we treat every input view as the reference view and construct their corresponding MPI respectively, where each MPI is parameterized by individual MLP (see Sec. 4.1 for details). Since the novel view image can be rendered by any MPI and deserve to have the same colors and depths, we propose the appearance and depth consistency loss to fully utilize cross-view multiplane consistency (see Sec. 4.2 for details).

## ABSTRACT

Neural Radiance Field (NeRF) has shown impressive results in novel view synthesis, particularly in Virtual Reality (VR) and Augmented Reality (AR), thanks to its ability to represent scenes continuously. However, when just a few input view images are available, NeRF tends to overfit the given views and thus make the estimated depths of pixels share almost the same value. Unlike previous methods that conduct regularization by introducing complex priors or additional supervisions, we propose a simple yet effective method that explicitly builds depth-aware consistency across input views to tackle this challenge. Our key insight is that by forcing the same spatial points to be sampled repeatedly in different input views, we are able to strengthen the interactions between views and therefore alleviate the overfitting problem. To achieve this, we build the neural networks on layered representations (*i.e.*, multiplane images), and the sampling point can thus be resampled on multiple discrete planes. Furthermore, to regularize the unseen target views, we constrain the rendered colors and depths from different input views to be the same. Although simple, extensive experiments demonstrate

that our proposed method can achieve better synthesis quality over state-of-the-art methods.

**Index Terms:** Neural Radiance Fields—Few-shot view synthesis—Multiplane Images—Cross-view consistency

## 1 INTRODUCTION

As a fundamental task in computer vision and computer graphics, novel view synthesis aims at rendering novel view images from given several posed input view images [4, 9]. Recently, Neural Radiance Field (NeRF) [25] has gained increasing popularity due to its powerful ability in continuous scene representation and its superior performance of novel view synthesis.

However, the success of NeRF and its variants depends on the number of input views to a large extent [16]. As shown in Fig. 2(a), when just a few input views are given, NeRF tends to overfit input views, resulting in the estimated depths of pixels sharing almost the same value [16, 59]. In principle, this overfitting problem could be alleviated by incorporating priors of different scenes into the neural network [6, 8, 23, 41, 46, 47, 57]. However, these methods require expensive pre-training cost and the pre-trained scenes usually exist domain gap for the target scene [28].

More recently, remarkable progress has also been made toward alleviating the overfitting problem by introducing external supervisions [10, 33, 43], pseudo views [1, 7, 18, 43, 52], or physical priors [16, 17, 28]. For example, Jain et al. [16] introduced semantic consistency between various views to encourage realistic render-

\*e-mail: hanxinzhu@mail.ustc.edu.cn

<sup>†</sup>e-mail: tianyuhe@microsoft.com

<sup>‡</sup>e-mail: chenzhibo@ustc.edu.cn

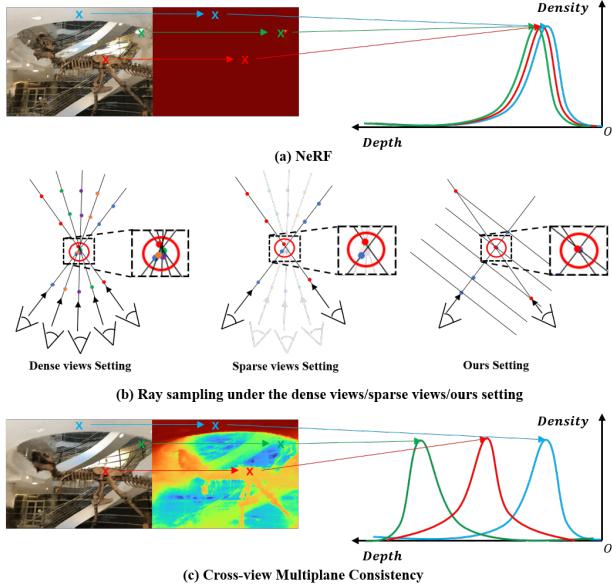


Figure 2: Given a few input views (e.g., 3 input views), (a) NeRF tends to overfit to input views and results in a dramatic performance drop, where the estimated depths of pixels share almost the same value. (b) Our key insight is to ensure the same spatial points can be sampled repeatedly in different input views. (c) Our proposed method can achieve smooth depth estimation by introducing cross-view multiplane consistency, resulting in better synthesis quality.

ings. Niemeyer et al. [28] regularized the geometry and appearance of patches for each unseen view. Although effective, the aforementioned methods either ignore the consistency across multiple views [17, 28] or impose the cross-view consistency solely on the image level [16], thereby limiting the performance.

To tackle this challenge, we make an assumption: due to fewer input views, the sampling point in each ray would rarely be used to render other views, therefore the neural networks tend to memorize colors of each input view instead of learning the underlying geometry [2, 58]. To validate this assumption, we propose Cross-view Multiplane Consistency (CMC), in which we force the sampling points to remain identical when rendering different views, as demonstrated in Fig. 2(b). In this way, the sampling points are able to be rendered to different-view images, resulting in depth-aware consistency across views. More specifically, for each input view, we build individual layered representations (*i.e.*, Multiplane Images) by regarding the input view as the reference view of the Multiplane Images (MPI) [62]. Therefore, based on the discrete multiplane representation, all sampling points are forced to be distributed on the same fixed planes.

Given the multiplane representation for each input view, we aim at imposing cross-view consistency on multiplanes during the optimization. We recognize this in two aspects: 1) for the input views whose ground-truth images are available, we optimize each MPI using a reconstruction loss that minimizes the difference between the rendered input view images and the ground-truth input view images. 2) for the unseen views that lack ground-truth images, we leverage the underlying consistency: the colors and depths that are rendered from different input views (*i.e.*, different MPIs) should maintain the same. As a result, we achieve cross-view multiplanes consistency.

We verify our assumptions and proposals on the common *LLFF* [24] and *Shiny* [50] dataset, where the overfitting problem can be well overcome with a promising improvement in the qualities of

synthetic novel views.

The main contributions of this paper can be summarized as follows:

- We propose to force the sampling points to be the same when rendering different views, which alleviates the overfitting problem of few-shot novel view synthesis.
- To achieve cross-view multiplane consistency, in addition to reconstruction loss for input views, we propose to impose appearance and depth consistency to the unseen views.
- We provide an explanation for the overfitting problem and then give the intuition behind our proposed CMC.
- Our proposed method achieves state-of-the-art performance on various widely adopted datasets.

## 2 RELATED WORK

### 2.1 Novel View Synthesis

As a long-standing problem in computer vision and computer graphics, novel view synthesis has been studied for decades with methods based on image-based rendering [4, 5, 9, 37], light fields [20, 24, 40, 51], point clouds [19, 39, 49, 55] and learning-based representation [11, 12, 32, 63]. Recently people have witnessed an increasing popularity for Neural Radiance Field (NeRF) [25] due to its remarkable performance for novel view synthesis. Given several 2D input view images of a static scene, NeRF can render photorealistic novel view images through coordinate-based implicit neural representation. It has been extended to several different tasks, such as dynamic scenes representation [29, 30], fast training and rendering [13, 27, 50, 56], stylization [15, 26, 45], generalizable scenes representation [6, 23, 47, 53] etc. Though NeRF achieved great synthesis quality, it depends on dense input view images, which would be not suitable for many practical applications. As a result, in this paper, we focus our attention on view synthesis with sparse input views, *e.g.*, few-shot novel view synthesis.

### 2.2 Few-shot NeRF

When only a few input view images with big disparities are available, NeRF easily overfits these input views, as shown in Fig. 2(a). Some generalizable neural fields [6, 8, 47, 57] could avoid this problem by using large-scale cross-scenes datasets to learn scenes priors, while the performance will degrade significantly when there is a large domain gap between the test scenes and the training dataset. [10, 33, 43] proposed to overcome the overfitting tendency of the few-shot setting in a per-scene optimization manner with additional supervision signals, such as sparse depth estimated by Structure-from-Motion [34] or pixel correspondence estimated by [42]. To increase the number of training views available, [1, 7, 18, 52] proposed to use depth-warping to generate novel view images as pseudo labels. [16, 17, 28] made use of physical priors to regularize the scene geometry without any additional supervision signals. Recently, FreeNeRF [54] mitigated the overfitting problem from the perspective of frequency, where a novel frequency annealing strategy on positional encoding was proposed. SimpleNeRF [38] instead leveraged augmented models for better and stable few-shot view synthesis. MixNeRF [36] modeled rays as mixtures of Laplacianssians, followed by FlipNeRF [35] which used flipped reflection rays as additional training sources.

Though these methods would achieve promising results, they either heavily rely on pre-trained neural networks that are usually expensive [33], or only take advantage of physical priors as regularization terms on seen/unseen views independently, without cross-view interactions [17]. Instead, in this paper, we propose to make full use of cross-view consistency to achieve the few-shot novel view synthesis.

### 2.3 Multiplane Images

MPI was first proposed by [62] to expand the small baselines of stereo images. Then [24] extended MPI to view synthesis by constructing local MPIs and blending different MPIs to render novel views. To achieve a fast generation of MPI, DeepView was proposed by [11] through the leverage of learned gradient descent. To model the time-dependent effects of scenes shot at different times, DeepMPI was introduced by [22] in an unsupervised manner. [14, 21, 44] further proposed to use MPI to realize single-view synthesis. Recently [50] has been proposed to model view-dependent effects and to realize real-time rendering. Then [61] proposed to take advantage of MPIs to make a 2D GAN 3D-aware. In this paper, we first apply MPI to few-shot view synthesis, where every input view is treated as the reference view respectively. To enhance the interactions across different views, we propose two new loss functions, *i.e.*, the appearance and depth consistency loss, based on the fact that the rendered colors and depths of the target view by different MPIs should be the same.

## 3 PRELIMINARIES

Our method is built upon Neural Radiance Field (NeRF) [25] and Multiplane Images (MPI) [62]. We elaborate on them in this section.

### 3.1 Neural Radiance Field

NeRF [25] has emerged as a powerful tool for continuous scene representation by encoding scene properties into a neural network  $F_\theta$ , which is usually parameterized by one Multilayer Perceptron (MLP). Input the 3D coordinate  $\mathbf{x} = (x, y, z)$  of a spatial point and one viewing direction  $\mathbf{d} = (d_x, d_y, d_z)$ , NeRF outputs the corresponding color  $\mathbf{c}$  and volume density  $\sigma$ , which is denoted as:

$$\mathbf{c}, \sigma = F_\theta(\gamma(\mathbf{x}), \gamma(\mathbf{d})), \quad (1)$$

where  $\gamma$  is the position encoding operation [25] that aim to recovering high-frequency detail textures.

Given several input view images and their camera parameters, a pixel can be rendered by casting a ray  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$  from the camera origin  $\mathbf{o}$  towards the pixel along direction  $\mathbf{d}$ . Specifically, assuming  $t \in [t_n, t_f]$ , the estimated color  $\mathbf{C}(\mathbf{r})$  of this pixel is formulated as follows:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt, \quad (2)$$

where  $T(t) = \exp(-\int_{t_n}^t \sigma(\mathbf{r}(s)) ds)$ ,  $\sigma$  and  $\mathbf{c}$  are obtained by Eq. 1. NeRF is optimized by minimizing the following loss function:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \|\mathbf{C}(\mathbf{r}) - \mathbf{C}_{gt}\|_2^2, \quad (3)$$

where  $\mathcal{R}$  is a set of sampling rays,  $\mathbf{C}(\mathbf{r})$  is obtained by Eq. 2 and  $\mathbf{C}_{gt}$  represents the ground-truth color.

### 3.2 Multiplane Images

As a layered scene representation, MPI [62] is constructed by a set of frontop-parallel planes with respect to a reference view, where all planes are fixed at specific depths that are distributed equally in the depth space. Considering one MPI with  $D$  planes  $(\mathbf{c}_i, \alpha_i)_{i=1}^D$ , the  $i$ -th plane at depth  $z_i$  can be viewed as a 4-channel RGBA image that contains the color  $\mathbf{c}_i$  and visibility  $\alpha_i$ .

To render a target view based on the MPI of the reference view, each plane of the MPI is warped to the target view  $(\mathbf{c}'_i, \alpha'_i)_{i=1}^D$  using inverse homography warping, followed by an alpha-composition operation [14, 21, 44, 62]. The rendered image  $\mathbf{I}_t$  and depth map  $Z_t$  of target view are denoted as follows:

$$\mathbf{I}_t = \sum_{i=1}^D \left( \mathbf{c}'_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j) \right) \quad (4)$$

$$Z_t = \sum_{i=1}^D \left( z_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j) \right). \quad (5)$$

We build our model on MPI. Therefore, the sampling point can be resampled on multiple discrete planes.

## 4 CROSS-VIEW MULTIPLANE CONSISTENCY

**Motivation.** As shown in Fig. 2(a), NeRF suffers from significant performance degradation when the number of input views is reduced, which also leads to the estimated depths of pixels sharing almost the same value [16, 59]. To tackle this problem, we assume that one plausible reason is that the sampling point in each ray would rarely be used to render other views due to fewer input views. Therefore, it is easier for the neural networks to memorize each input view images [2, 58], rather than learning the underlying geometry. Motivated by this, our key insight is to explicitly build depth-aware consistency across different views.

**Method Overview.** As shown in Fig. 1, to ensure that the sampling points are the same when rendering different views, we build individual layered representation (*i.e.*, Multiplane Images)  $F_\theta^i$  for each input view  $i$  by utilizing the input view  $i$  as the reference view of the Multiplane Images (MPI) [62]. Therefore, all sampling points are distributed on the same fixed planes. Inspired by previous works [21, 50], each MPI is presented by a multilayer perceptron (MLP)  $F_\theta^i$ , which outputs the color and visibility for each plane.

To optimize  $F_\theta^i$  for the input views, we directly minimize the difference between rendered images and the ground-truth ones through a reconstruction loss. While there is no ground-truth image for the unseen views, we introduce an intuition that the colors and depth rendered by different input views should be the same. Specifically, we minimize the difference in the estimated colors and depths that are obtained by different MPIs.

### 4.1 Multiplane Representation for Input Views

As shown in Fig. 1, given several sparse input view images  $\{\mathbf{I}_{in}^i\}_{i=0}^{N-1} \in R^{H \times W \times 3}$  and their corresponding camera extrinsics  $[\mathbf{R}_{in}^i | \mathbf{t}_{in}^i]_{i=0}^{N-1} \in SE(3)$  of a static scene, our goal is to render novel view images photorealistically, where  $H$  and  $W$  are the image height and width,  $N$  is the number of input views available,  $\mathbf{R} \in R^{3 \times 3}$  and  $\mathbf{t} \in R^{3 \times 1}$  represent the rotation matrix and translation vector.

As described in motivation, in this paper we use MPIs to represent the scene. Different from most MPI-based methods that randomly choose one input view as the reference view and the left input views as the target views [14, 21, 44, 50, 62], we propose to treat every input view as the reference view respectively and construct their corresponding MPIs  $\{\mathbf{M}_i\}_{i=0}^{N-1}$  (*i.e.*, per-view MPI), for the purpose of building depth-aware consistency across different input views. We adopt MLPs to present the MPIs following previous works [21, 50].

Specifically, considering the camera parameter  $[\mathbf{R}_t | \mathbf{t}_t]$  of one target view and the  $i$ -th MPI  $\mathbf{M}_i$  corresponding to  $\mathbf{I}_{in}^i$  that has  $D$  planes. When a ray  $\mathbf{r}$  is cast from the camera origin  $\mathbf{o}$  of the target view through one pixel at its image plane whose coordinate is  $(u_t, v_t)$  along direction  $\mathbf{d}_i^{(u_t, v_t)}$ , it will have  $D$  intersections with the  $D$  planes of  $\mathbf{M}_i$ , which are denoted as  $\{\mathbf{x}_i^k = (u_i^k, v_i^k, z_i^k)\}_{k=0}^{D-1}$ , where  $(u_i^k, v_i^k)$  is the pixel coordinate of the  $k$ -th intersection and  $z_i^k$  represents the depth that plane  $k$  is placed. The pixel coordinate of each intersection can be computed by the inverse homography warping operation [14, 21, 44, 50, 62], which is formulated as follows:

$$\begin{bmatrix} u_i^k \\ v_i^k \\ 1 \end{bmatrix} \sim \mathbf{K}_{in}^i \left( \mathbf{R}' - \frac{\mathbf{t}' \mathbf{n}^\top}{z_i^k} \right) \mathbf{K}_t^{-1} \begin{bmatrix} u_t \\ v_t \\ 1 \end{bmatrix}, \quad (6)$$

where  $\mathbf{K}_{in}^i \in R^{3 \times 3}$  and  $\mathbf{K}_t \in R^{3 \times 3}$  are the camera intrinsics for the input view  $\mathbf{I}_{in}^i$  and target view respectively,  $\mathbf{n} = [0, 0, 1]^\top$  is the normal vector of the  $k$ -th plane,  $\mathbf{R}'$  and  $\mathbf{t}'$  are the relative camera extrinsic from the target view to the input view, which is computed as follows:

$$\begin{bmatrix} \mathbf{R}'_{3 \times 3} & \mathbf{t}'_{3 \times 1} \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R}_t & \mathbf{t}_t \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{R}_{in}^i & \mathbf{t}_{in}^i \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix}. \quad (7)$$

With the computed coordinate  $\mathbf{x}_i^k$  of each intersection along the ray  $\mathbf{r}$  whose direction is  $\mathbf{d}_i^{(u_t, v_t)}$ , both  $\mathbf{x}_i^k$  and  $\mathbf{d}_i^{(u_t, v_t)}$  are fed into the MLP  $F_\theta^i$  to estimate its color  $\mathbf{c}_i^k$  and visibility  $\alpha_i^k$  as shown in Fig. 1, which is denoted as:

$$\mathbf{c}_i^k, \alpha_i^k = F_\theta^i(\gamma(\mathbf{x}_i^k), \gamma(\mathbf{d}_i^{(u_t, v_t)})), \quad (8)$$

where  $\gamma$  is the position encoding operation [25] that is formulated as follows:

$$\gamma(\mathbf{x}) = (\sin(2^0 \mathbf{x}), \cos(2^0 \mathbf{x}), \dots, \sin(2^{L-1} \mathbf{x}), \cos(2^{L-1} \mathbf{x})), \quad (9)$$

$L$  is the hand-crafted hyperparameter. Then the color  $\mathbf{C}_i(\mathbf{r})$  and depth  $Z_i(\mathbf{r})$  of the pixel  $(u_t, v_t)$  in the target view can be rendered based on volume rendering by the  $i$ -th MPI.

## 4.2 Cross-view Consistency on Multiplanes

**Reconstruction Loss for Input Views.** Given the rendered color  $\mathbf{C}_i(\mathbf{r})$ , if the target view is one of the input views, then the reconstruction loss (Eq. 3) that minimizes the difference from  $\mathbf{C}_i(\mathbf{r})$  to the ground truth color  $\mathbf{C}_{gt}$  is adopted, which is denoted as follows:

$$\mathcal{L}_{MSE} = \frac{1}{|N|} \frac{1}{|\mathcal{R}|} \sum_{i=0}^{N-1} \sum_{\mathbf{r} \in \mathcal{R}} \|\mathbf{C}_i(\mathbf{r}) - \mathbf{C}_{gt}\|_2^2, \quad (10)$$

where  $\mathcal{R}$  is a set of sampling rays. In Sec. 5.3, we verify that with our multiplane representation, the reconstruction loss alone can overcome the overfitting problem well.

**Appearance and Depth Consistency Loss for Unseen Views.** The above reconstruction loss utilizes consistency across known input views by forcing the spatial points to be sampled on the same planes. To obtain depth-aware consistency across views, we propose the appearance and depth consistency loss across unseen novel views.

Specifically, when the target view is a novel view that has no ground truth color, it still can be rendered by any MPI and deserve to have the same color and depth map, as shown in Fig. 1. Based on such an observation, we propose the following loss functions:

$$\mathcal{L}_{ac} = \frac{2}{|N(N-1)|} \frac{1}{|\mathcal{R}|} \sum_{i=0}^{N-1} \sum_{j=i+1}^{N-1} \sum_{\mathbf{r} \in \mathcal{R}} \|\mathbf{C}_i(\mathbf{r}) - \mathbf{C}_j(\mathbf{r})\|_2^2, \quad (11)$$

$$\mathcal{L}_{dc} = \frac{2}{|N(N-1)|} \frac{1}{|\mathcal{R}|} \sum_{i=0}^{N-1} \sum_{j=i+1}^{N-1} \sum_{\mathbf{r} \in \mathcal{R}} \|Z_i(\mathbf{r}) - Z_j(\mathbf{r})\|_2^2, \quad (12)$$

where  $\mathbf{C}_i/Z_i$  and  $\mathbf{C}_j/Z_j$  represents the rendered colors and depths by the  $i$ -th MPI and  $j$ -th MPI respectively.

As a result, the whole loss function of our proposed method can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{MSE} + \lambda_{ac} \mathcal{L}_{ac} + \lambda_{dc} \mathcal{L}_{dc}, \quad (13)$$

where  $\lambda_{ac}$  and  $\lambda_{dc}$  are hyperparameters that balance the weights of  $\mathcal{L}_{ac}$  and  $\mathcal{L}_{dc}$ .

## 4.3 Weighted Rendering

To render a target view from multiple MPIS, based on the assumption that the closest MPI to the target view should have a greater impact on its rendering process, we adopt a weighted rendering strategy. Specifically, the final output  $\mathbf{C}(\mathbf{r})$  is obtained by calculating a weighted average of the rendering colors from different MPIS, which is denoted as follows:

$$\mathbf{C}(\mathbf{r}) = \sum_{i=0}^{N-1} w_i \cdot \mathbf{C}_i(\mathbf{r}), \quad (14)$$

where  $\mathbf{C}_i(\mathbf{r})$  is the color rendered by the  $i$ -th MPI,  $w_i$  is the weight calculated according to the distance  $\mu_i$  from the  $i$ -th MPI to the target view, which is formulated as follows:

$$w_i = \frac{\mu_i}{\sum_{j=0}^{N-1} \mu_j}, \quad \mu_i = \|\mathbf{o}_t - \mathbf{o}_i\|_2^2, \quad (15)$$

where  $\mathbf{o}_t$  and  $\mathbf{o}_i$  represent the camera origins of the target view and the  $i$ -th MPI respectively.

## 4.4 Analysis on Cross-view Multiplane Consistency

To demonstrate the effectiveness of our proposed method, we make an analysis of CMC in this section. To begin with, we propose an assumption for the overfitting problem of NeRF under the few-shot setting. Specifically, given sparse input views, as shown in Fig. 2(b), a fact is that it is quite difficult for rays of different views to have the same sampling points due to the random uniform sampling strategy of NeRF [25], which is denoted as follows:

$$t_i \sim \mathcal{U} \left[ t_n + \frac{i-1}{M} (t_f - t_n), t_n + \frac{i}{M} (t_f - t_n) \right], \quad (16)$$

where  $t_n$  and  $t_f$  are the near and far bounds,  $M$  is the number of sampling points along the ray and  $t_i$  is the  $i$ -th sampling points. As a result, our assumption is that the sampling points in each ray would only take part in the rendering process of pixels corresponding to this ray, while rarely being used to render other views. Thus, the optimization process of NeRF [25] (Eq. 2) can be viewed as solving the following equation for each ray independently:

$$\mathbf{C}_{gt} = \sum_{i=1}^M T(\sigma_i) f(\sigma_i) \mathbf{c}_i, \quad (17)$$

where  $T$  and  $f$  are both functions of  $\sigma_i$ ,  $\mathbf{C}_{gt}$  is the ground-truth pixel of ray  $\mathbf{r}$ ,  $\sigma_i$  and  $\mathbf{c}_i$  are unknowns to be estimated. Obviously, we have  $2M$  unknowns while only one equation, which means that infinite solutions exist for this problem. Considering the memorization nature of neural networks [2, 58] and Occam's Razor [3, 31], NeRF tends to converge to the simplest way to represent known input views, thus Eq. 17 is assumed to solve the following sparse optimization problem:

$$\begin{aligned} \mathbf{c}_i^*, \sigma_i^* = \arg \min_{\mathbf{c}_i, \sigma_i} & \left\{ \left\| \sum_{i=1}^M T(\sigma_i) f(\sigma_i) \mathbf{c}_i - \mathbf{C}_{gt} \right\|_2^2 \right. \\ & \left. + \sum_{i=1}^M \|\mathbf{c}_i\|_0 + \sum_{i=1}^M \|\sigma_i\|_0 \right\} \end{aligned}, \quad (18)$$

whose solution is

$$\{\mathbf{c}_i^*, \sigma_i^*\} = \begin{cases} \{\mathbf{C}_{gt}, 1\}, & i = 0 \\ \{\mathbf{0}, 0\}, & i \geq 1 \end{cases}, \quad (19)$$

which thus leads to the overfitting problem.

Table 1: Quantitative comparisons on 8 scenes of the *Shiny* dataset.

Method	NeRF [25]			DietNeRF [16]			InfoNeRF [17]			Ours		
Scene	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Cake	15.98	0.514	0.576	18.04	0.556	0.543	14.71	0.469	0.653	17.08	0.564	0.496
Crest	11.50	0.152	0.729	9.74	0.105	0.733	12.28	0.181	0.736	14.54	0.268	0.564
Food	12.65	0.296	0.657	10.30	0.190	0.736	13.25	0.328	0.679	16.00	0.425	0.502
Giants	12.39	0.218	0.730	12.54	0.216	0.733	6.32	0.010	0.776	13.42	0.299	0.651
Pasta	13.95	0.370	0.550	13.96	0.373	0.545	13.84	0.353	0.632	14.89	0.389	0.523
Room	21.19	0.710	0.454	20.01	0.669	0.483	18.99	0.578	0.638	22.59	0.750	0.378
Seasoning	12.27	0.358	0.684	12.05	0.347	0.682	12.62	0.384	0.684	13.05	0.447	0.605
Tools	15.04	0.580	0.500	8.35	0.276	0.717	10.89	0.358	0.65	16.23	0.598	0.411
Average	14.37	0.399	0.610	13.12	0.341	0.646	12.86	0.332	0.681	<b>15.98</b>	<b>0.468</b>	<b>0.516</b>

Based on the analysis above, to overcome such a problem, a direct way is to impose the same point to be sampled in rays of different views. Take two input views  $I_0$  and  $I_1$  as an example, whose camera origins are  $\mathbf{o}_0$  and  $\mathbf{o}_1$  respectively. For one sampling point  $\mathbf{x}_0 = \mathbf{o}_0 + t_0 \mathbf{d}_0$  along ray  $\mathbf{d}_0$  of  $I_0$ , our goal is that  $\mathbf{x}_0$  can also be sampled in  $I_1$  along ray  $\mathbf{d}_1$ , thus guarantee that the same sampling point can take part in the rendering process of pixels in different views. Assuming that the sampling point in ray  $\mathbf{d}_1$  is denoted as  $\mathbf{x}_1 = \mathbf{o}_1 + t_1 \mathbf{d}_1$ , then the problem can be converted into the following formulation (*i.e.*, find the optimal  $t_1$  that can minimize the distances between  $\mathbf{x}_0$  and  $\mathbf{x}_1$ ):

$$t_1^* = \arg \min_{t_1} \|\mathbf{x}_1 - \mathbf{x}_0\|_2^2, t_1 \in [t_n, t_f], \quad (20)$$

where  $\mathbf{x}_0$ ,  $\mathbf{o}_1$  and  $\mathbf{d}_1$  is known. As a result, our goal is to find the optimal  $t_1$  that can satisfy the following formulation:

$$\|\{\mathbf{o}_0^x + t_0 d_0^x, \mathbf{o}_0^y + t_0 d_0^y, \mathbf{o}_0^z + t_0 d_0^z\} - \{\mathbf{o}_1^x + t_1 d_1^x, \mathbf{o}_1^y + t_1 d_1^y, \mathbf{o}_1^z + t_1 d_1^z\}\|_2^2 \leq \delta, \quad (21)$$

where  $\delta \rightarrow 0$ .

Assuming that  $\mathbf{o}_0 = \{\mathbf{o}_0^x, \mathbf{o}_0^y, \mathbf{o}_0^z\}$ ,  $\mathbf{d}_0 = \{d_0^x, d_0^y, d_0^z\}$ ,  $\mathbf{o}_1 = \{\mathbf{o}_1^x, \mathbf{o}_1^y, \mathbf{o}_1^z\}$  and  $\mathbf{d}_1 = \{d_1^x, d_1^y, d_1^z\}$ , then Eq. 21 can be converted into the following formulation:

$$\|\{\mathbf{o}_0^x + t_0 d_0^x, \mathbf{o}_0^y + t_0 d_0^y, \mathbf{o}_0^z + t_0 d_0^z\} - \{\mathbf{o}_1^x + t_1 d_1^x, \mathbf{o}_1^y + t_1 d_1^y, \mathbf{o}_1^z + t_1 d_1^z\}\|_2^2 \leq \delta. \quad (22)$$

Simplifying the above formula, we obtain:

$$t_1^* = \Phi(t_0), \quad (23)$$

where

$$\Phi(u) = ((\mathbf{o}_0^x + u d_0^x - \mathbf{o}_1^x)^2 + (\mathbf{o}_0^y + u d_0^y - \mathbf{o}_1^y)^2 + (\mathbf{o}_0^z + u d_0^z - \mathbf{o}_1^z)^2)^{1/2}, \quad (24)$$

which means that when sampling points in view  $I_0$  are known, then all sampling points in view  $I_1$  should be deterministic. Fortunately, this is exactly the nature of multiplane images. When view  $I_0$  is selected as the reference view to construct the MPI, all the sampling points of different views are deterministic and forced to be distributed on the same planes. As a result, consistency across different views can be well guaranteed. Experiments in Sec. 5.3 also demonstrate the effectiveness of our analysis.

## 5 EXPERIMENTS

We make a comparison with various state-of-the-art methods for few-shot novel view synthesis quantitatively and qualitatively. We also present a detailed analysis of the necessity of adopting per-view

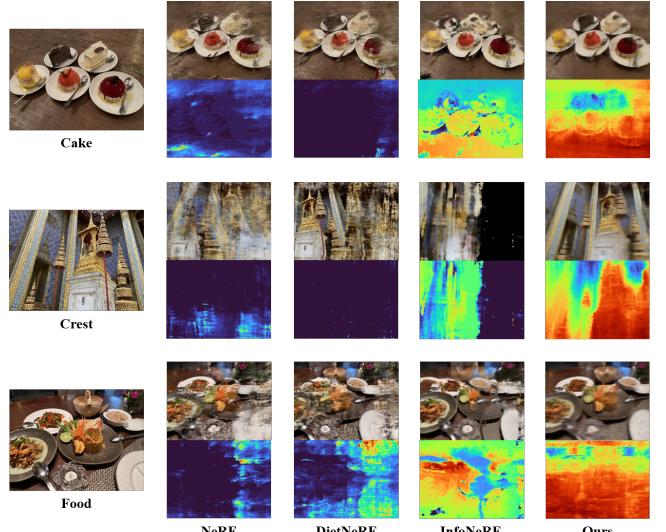


Figure 3: Qualitative comparisons on the *Shiny* dataset, where our proposed method can achieve better novel view synthesis and accurate geometry estimation (*i.e.*, the depth map).

multiplane images and the appearance/depth consistency loss. See supplementary materials for demonstrations of Eq. 18 and Eq. 20, ablation studies on the influence of different numbers of MPI planes, and more visualization results of novel view synthesis. We only evaluate our proposed method on extremely sparse input views, *i.e.*, 3 input views, as it is the most common case.

### 5.1 Implementation Details

**Datasets.** We perform experiments on the *LLFF* dataset [24] and the *Shiny* [50] dataset to validate the effectiveness of our proposed method. Both of the two datasets contain 8 complex real-world scenes with big disparities, while the *Shiny* dataset is more complicated because it has more view-dependent effects such as reflections and refraction. We follow the experimental protocols provided by [28], where the resolution of both input views and target views are  $378 \times 504$ . To make a fair comparison, similar to previous methods, for each scene we choose every 8-th image as the held-out test set and then select 3 images evenly from the remaining images as the input views. Notably, following [28], in our experiment the sampled input views are distributed uniformly in the camera pose space, where the distances across different input views are almost the same. However, our proposed method can also

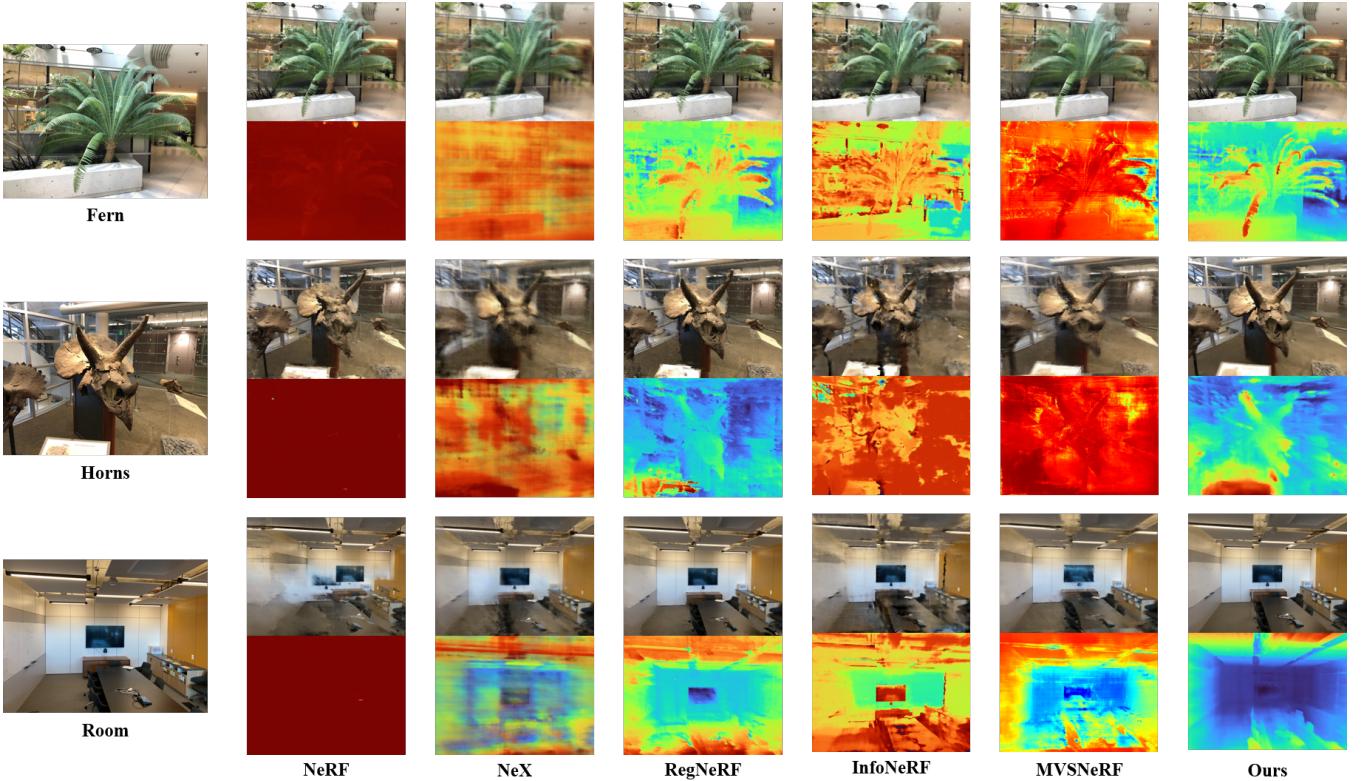


Figure 4: Qualitative comparisons on the *LLFF* dataset. Our proposed method can avoid the overfitting problem, where better novel view synthesis and more continuous depth estimation can be achieved.

be applied to scenarios where input views are randomly selected and exhibit greater spatial separation. This flexibility stems from our individual construction of MPis for each input view, and the capacity of each MPI to render novel views consistently. By assuming that the results of rendered novel views by different MPis remain the same, regions in the novel view overlapping with input views are effectively constrained, producing coherent and reasonable outcomes.

**Training Details.** As we discussed above, we construct per-view MPI with 80 planes for each input view, where each MPI is modeled by one independent four or six-layer leakyrelu-MLP with 256 nodes per layer. We set  $\gamma=10$  for the spatial coordinate  $\mathbf{x}_i^k$  while no position-encoding for the direction vector  $\mathbf{d}_i^{(u_i, v_i)}$ . The initial learning rate is  $5 \times 10^{-3}$  and then gradually reduce to  $1 \times 10^{-4}$ . At the beginning of the training process, we use only the reconstruction loss (Eq. 10) to train the network. After 15 epochs the whole loss function (Eq. 13) with appearance and depth consistency loss is used, where both  $\lambda_{ac}$  and  $\lambda_{dc}$  are set to 1. We train our models with the Adam optimizer with randomly sampled 1024 rays in a batch within 50 epochs by a single NVIDIA RTX 3090 GPU. It takes about 2 hours to train a scene and 10 seconds to render a target view.

**Metrics.** We evaluate the quality of rendered novel view images with Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [48], and Learned Perceptual Image Patch Similarity (LPIPS) [60]. For easier comparison, we also report the average score by calculating the geometric mean of  $10^{-\text{PSNR}/10}$ ,  $\sqrt{1 - \text{SSIM}}$  and LPIPS for the *LLFF* dataset similar to [28].

Table 2: Quantitative comparisons on the *LLFF* dataset. Our proposed method can achieve state-of-the-art performance. ft indicates the results fine-tuned on each scene individually.

Method	PSNR↑	SSIM↑	LPIPS↓	Average↓
NeRF [25]	13.34	0.373	0.451	0.255
NeX [50]	17.36	0.591	0.369	0.163
DietNeRF [16]	14.94	0.370	0.496	0.232
InfoNeRF [17]	14.37	0.349	0.457	0.238
PixelNeRF-ft [57]	16.17	0.438	0.438	0.217
SRF-ft [8]	17.07	0.436	0.529	0.203
MVSNeRF-ft [6]	17.88	0.584	0.327	0.157
GeCoNeRF [18]	18.55	0.578	0.340	0.150
RegNeRF [28]	19.08	0.587	0.336	0.146
MixNeRF [36]	19.27	0.629	0.336	0.134
FlipNeRF [35]	19.34	0.631	0.335	0.133
<b>Ours</b>	<b>19.45</b>	<b>0.659</b>	<b>0.310</b>	<b>0.127</b>

## 5.2 Comparisons with State-of-the-art Methods

### 5.2.1 Results on the *Shiny* Dataset

We first compare our proposed method with vanilla NeRF [25], DietNeRF [16] and InfoNeRF [17] on the challenging *Shiny* dataset proposed by [50] to demonstrate the effectiveness of CMC. We choose 8 real-world scenes from the official shiny and shiny-extended dataset that contain complex view-dependent effects. As shown in Fig. 3, NeRF and DietNeRF will overfit to input views, where the estimated geometry (*i.e.*, the depth map) is quite poor. For InfoNeRF, though it can render a more reasonable depth map, it will fail in more compli-

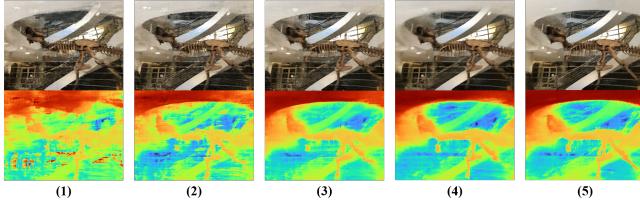


Figure 5: Qualitative comparisons of different choices of loss functions. (1) Single MPI with  $\mathcal{L}_{\text{MSE}}$ . (2) Per-view MPI with  $\mathcal{L}_{\text{MSE}}$ . (3) Per-view MPI with  $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{dc}}^{\text{I}}$ . (4) Per-view MPI with  $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{dc}}^{\text{I}} + \mathcal{L}_{\text{ac}}$ . (5) Per-view MPI with  $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{dc}}^{\text{I}} + \mathcal{L}_{\text{ac}} + \mathcal{L}_{\text{dc}}$ .

cated scenes such as "Crest". On account that DietNeRF only uses a high-level semantic loss on the image level to realize consistency across different views, it will generate repeated contents on the rendered novel view image. Differently, InfoNeRF takes advantage of ray entropy loss to regularize the seen/unseen views independently, where no cross-views interactions exist. As a result, for some occluded areas in the novel view that don't appear in the input views, it is quite difficult for them to estimate reasonable contents. Instead, our proposed method, *i.e.*, CMC, can render accurate depth maps and novel view images by virtue of a fully utilize of cross-view consistency. As demonstrated in Tab. 1, CMC can achieve state-of-the-art performance on all the metrics, which reflects the fact that introducing only physical priors would not be strong enough to deal with complex scenes under the few-shot setting, leverage of cross-view consistency will be helpful for obtaining a more accurate geometry estimation.

### 5.2.2 Results on the LLFF Dataset

Similar to many previous works, we also perform experiments on the common LLFF dataset against many state-of-the-art methods to demonstrate the superiority of our proposed method. Specifically, we compare our method with pretraining-based methods (*i.e.*, Pixel-NeRF [57], SRF [8], MVSNeRF [6]), regularization-based methods (*i.e.*, DietNeRF [16], InfoNeRF [17], RegNeRF [28]), pseudo view-based method (*i.e.*, GeCoNeRF [18]) and NeRF [25].

As verified in Tab. 2, our method can still achieve state-of-the-art performance with a big improvement in SSIM. For qualitative comparisons, as shown in Fig. 4, for methods based on pre-trained network such as MVSNeRF, though they can avoid overfitting to input views to some extent, the rendered novel view images would contain unreasonable artifacts due to the domain gap between training dataset and test scenes. Moreover, for input views with quite big disparities, MVSNeRF still falls into overfitting and estimates wrong geometry, as demonstrated by the scene named "Horns". For regularization-based methods such as InfoNeRF, severe artifacts will exist in the generated novel view images. For RegNeRF, the method with the best performance for few-shot novel view synthesis at present, it can overcome the overfitting problem to a large extent by means of depth smoothing regularization and a well-designed sampling annealing strategy. However, RegNeRF still generates some unreasonable geometry and results in discontinuous depth estimation, as demonstrated by the TV and conference table in the scene named "Room". On the contrary, our proposed method can achieve not only photorealistic novel view synthesis but also quite accurate and continuous depth estimation, without any physical priors serving as the regularization term or any hand-crafted complex sampling strategy to avoid overfitting. In other words, our method can realize few-shot novel view synthesis elegantly with lower complexity, which promises many practical applications.

Table 3: Ablation studies on the choices of different loss functions.

Loss	Sing. MPI	Per-view MPI				
		$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$
$\mathcal{L}_{\text{MSE}}$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$
$\mathcal{L}_{\text{dc}}^{\text{I}}$	$\times$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$
$\mathcal{L}_{\text{ac}}$	$\times$	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$
$\mathcal{L}_{\text{dc}}$	$\times$	$\times$	$\times$	$\times$	$\times$	$\checkmark$
PSNR $\uparrow$	17.56	18.33	18.69	19.24	19.27	<b>19.45</b>
SSIM $\uparrow$	0.597	0.618	0.634	0.656	0.656	<b>0.659</b>
LPIPS $\downarrow$	0.359	0.345	0.334	0.336	0.321	<b>0.310</b>
Average $\downarrow$	0.158	0.146	0.139	0.133	0.130	<b>0.129</b>

### 5.3 Ablation Studies

To verify the importance of constructing per-view MPI and the appearance/depth consistency loss, we perform ablation studies on the choices of loss functions. Specifically, we choose loss functions (Eq. 13) composed of different combinations of the reconstruction loss  $\mathcal{L}_{\text{MSE}}$  (Eq. 10), the depth consistency loss on input views  $\mathcal{L}_{\text{dc}}^{\text{I}}$  (Eq. 12), the appearance consistency loss on novel views  $\mathcal{L}_{\text{ac}}$  (Eq. 11) and the depth consistency loss on novel views  $\mathcal{L}_{\text{dc}}$  (Eq. 12).

**Single MPI.** As shown in Fig. 5, for the setting of single MPI (Sing. MPI), *i.e.*, only one random input view is selected as the reference view and thus only one MPI is constructed to render novel views, though some artifacts exist, the neural network can already avoid overfitting to input views. Actually, a single MPI is a variant of NeRF where only the sampling points in rays of different input views are imposed to be distributed on the same planes. However, such a slight change can achieve nearly 4dB PSNR improvement over NeRF as demonstrated in Tab. 2 and Tab. 3. This observation reflects the effectiveness and superiority of our method, where cross-view multiplane consistency can benefit a lot for accurate geometry recovery.

**Per-view MPI.** To enhance interactions across different input views, we further propose per-view MPI by treating every input view as the reference view and then constructing their corresponding MPis. As shown in Fig. 5 and Tab. 3, with only  $\mathcal{L}_{\text{MSE}}$ , per-view MPI can witness an increase in rendering quality and generate more accurate geometry estimation, which demonstrates the effectiveness of per-view MPI. Then, when we successively add  $\mathcal{L}_{\text{dc}}^{\text{I}}$ ,  $\mathcal{L}_{\text{ac}}$  and  $\mathcal{L}_{\text{dc}}$  into the loss function, a continuous growth of performance can be observed, where more photorealistic novel view images and better depth estimation can be achieved.

## 6 CONCLUSIONS

We present a brand-new technique for few-shot novel view synthesis by cross-view multiplane consistency (*i.e.*, CMC). We propose to address the overfitting problem of few-shot view synthesis by forcing the sampling points to be the identical when rendering different views through multiplane images. This is based on the assumption that given sparse input view images, the sampling point in each ray would rarely be used to render other views and thus cause the neural networks to memorize input views rather than learn the underlying geometry. Then, to enhance interactions across different views, we propose to construct per-view MPI by viewing every input view as the reference view followed by leverage of appearance and depth consistency loss. We further provide an explanation for the overfitting problem and give the intuition behind CMC. To verify our assumption and method, we conduct experiments on a large amount of complex real-world scenes, where our proposed CMC can achieve state-of-the-art few-shot novel view synthesis, without any scene priors or complicated hand-crafted sampling strategy.

## 7 LIMITATIONS AND FUTURE WORKS

**Limitations.** The main limitation of our proposed method is that CMC doesn't perform well on surrounding scenes that contain big camera rotations, such as the Blender dataset [25]. This is because CMC is based on Multiplane Images (*i.e.*, MPI), which is specially designed for forward-facing scenes while not suitable to represent surrounding scenes.

**Future works** Our future works include extending CMC to surrounding scenes using methods such as multisphere representation, where sampling points are forced to be distributed on the same spheres. Moreover, we will try to use only one MLP instead of per-view MPI to represent the scene, which would decrease the training burden to a large extent.

## 8 ACKNOWLEDGEMENT

This work was supported in part by NSFC under Grant 62371434, U1908209, 62021001.

## REFERENCES

- [1] Y. C. Ahn, S. Jang, S. Park, J.-Y. Kim, and N. Kang. Panerf: Pseudo-view augmentation for improved neural radiance fields based on few-shot inputs. *arXiv preprint arXiv:2211.12758*, 2022.
- [2] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.
- [3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam’s razor. *Information processing letters*, 24(6):377–380, 1987.
- [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*, pp. 425–432, 2001.
- [5] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)*, 32(3):1–12, 2013.
- [6] 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*, pp. 14124–14133, 2021.
- [7] D. Chen, Y. Liu, L. Huang, B. Wang, and P. Pan. Geoaug: Data augmentation for few-shot nerf with geometry constraints. In *European Conference on Computer Vision*, pp. 322–337. Springer, 2022.
- [8] J. Chibane, A. Bansal, V. Lazova, and G. Pons-Moll. Stereo radiance fields (srdf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7911–7920, 2021.
- [9] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 11–20, 1996.
- [10] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12882–12891, 2022.
- [11] 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*, pp. 2367–2376, 2019.
- [12] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. Deepstereo: Learning to predict new views from the world’s imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5515–5524, 2016.
- [13] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5501–5510, 2022.
- [14] Y. Han, R. Wang, and J. Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–8, 2022.
- [15] Y.-H. Huang, Y. He, Y.-J. Yuan, Y.-K. Lai, and L. Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18342–18352, 2022.
- [16] A. Jain, M. Tancik, and P. Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5885–5894, 2021.
- [17] M. Kim, S. Seo, and B. Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12912–12921, 2022.
- [18] M. Kwak, J. Song, and S. Kim. Geconerf: Few-shot neural radiance fields via geometric consistency. *arXiv preprint arXiv:2301.10941*, 2023.
- [19] H.-A. Le, T. Mensink, P. Das, and T. Gevers. Novel view synthesis from single images via point cloud transformation. *arXiv preprint arXiv:2009.08321*, 2020.
- [20] M. Levoy and P. Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 31–42, 1996.
- [21] J. Li, Z. Feng, Q. She, H. Ding, C. Wang, and G. H. Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12578–12588, 2021.
- [22] Z. Li, W. Xian, A. Davis, and N. Snavely. Crowdsampling the plenoptic function. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* 16, pp. 178–196. Springer, 2020.
- [23] 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*, pp. 7824–7833, 2022.
- [24] 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)*, 38(4):1–14, 2019.
- [25] 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*, 65(1):99–106, 2021.
- [26] F. Mu, J. Wang, Y. Wu, and Y. Li. 3d photo stylization: Learning to generate stylized novel views from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16273–16282, 2022.
- [27] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.
- [28] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. Sajjadi, A. Geiger, and N. Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5480–5490, 2022.
- [29] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021.
- [30] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. 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.
- [31] C. Rasmussen and Z. Ghahramani. Occam’s razor. *Advances in neural information processing systems*, 13, 2000.
- [32] G. Riegler and V. Koltun. Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12216–12225, 2021.
- [33] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, and M. Nießner. 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.
- [34] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.
- [35] S. Seo, Y. Chang, and N. Kwak. Flipnerf: Flipped reflection rays for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22883–22893, 2023.
- [36] S. Seo, D. Han, Y. Chang, and N. Kwak. Mixnerf: Modeling a ray with mixture density for novel view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20659–20668, 2023.
- [37] S. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *2009 International Conference on Computer Vision*, pp. 1881–1888, 2009.
- [38] N. Somraj, A. Karanayil, and R. Soundararajan. Simplenerf: Regularizing sparse input neural radiance fields with simpler solutions. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.
- [39] Z. Song, W. Chen, D. Campbell, and H. Li. Deep novel view synthesis from colored 3d point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 1–17. Springer, 2020.
- [40] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng. Learning to synthesize a 4d rgbd light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2243–2251, 2017.
- [41] 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*, pp. 15182–15192, 2021.
- [42] P. Truong, M. Danelljan, L. Van Gool, and R. Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5714–5724, 2021.
- [43] P. Truong, M.-J. Rakotosaona, F. Manhardt, and F. Tombari. Sparf: Neural radiance fields from sparse and noisy poses. *arXiv preprint arXiv:2211.11738*, 2022.
- [44] R. Tucker and N. Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 551–560, 2020.
- [45] C. Wang, R. Jiang, M. Chai, M. He, D. Chen, and J. Liao. Nerf-art: Text-driven neural radiance fields stylization. *arXiv preprint arXiv:2212.08070*, 2022.
- [46] D. Wang, X. Cui, S. Salcudean, and Z. J. Wang. Generalizable neural radiance fields for novel view synthesis with transformer. *arXiv preprint arXiv:2206.05375*, 2022.
- [47] 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*, pp. 4690–4699, 2021.
- [48] 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*, 13(4):600–612, 2004.
- [49] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7467–7477, 2020.
- [50] S. Wizadwongsaa, P. Phongthawee, J. Yenphraphai, and S. Suwanjanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8534–8543, 2021.
- [51] D. N. Wood, D. I. Azuma, K. Aldinger, B. Curless, T. Duchamp, D. H. Salesin, and W. Stuetzle. Surface light fields for 3d photography. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 287–296, 2000.
- [52] D. Xu, Y. Jiang, P. Wang, Z. Fan, H. Shi, and Z. Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pp. 736–753. Springer, 2022.
- [53] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5438–5448, 2022.
- [54] J. Yang, M. Pavone, and Y. Wang. 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.
- [55] M. You, M. Guo, X. Lyu, H. Liu, and J. Hou. Learning a unified 3d point cloud for view synthesis. *arXiv preprint arXiv:2209.05013*, 2022.
- [56] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5752–5761, 2021.
- [57] 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*, pp. 4578–4587, 2021.
- [58] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.
- [59] K. Zhang, G. Riegler, N. Snavely, and V. Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.
- [60] 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*, pp. 586–595, 2018.
- [61] X. Zhao, F. Ma, D. Güera, Z. Ren, A. G. Schwing, and A. Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pp. 18–35. Springer, 2022.
- [62] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.
- [63] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 286–301. Springer, 2016.