Bidirectional Optical Flow NeRF: High Accuracy and High Quality under Fewer Views

Shuo Chen, Binbin Yan*, Xinzhu Sang, Duo Chen, Peng Wang, Xiao Guo, Chongli Zhong, Huaming Wan

State Key Laboratory of Information Photonics and Optical Communications,
Beijing University of Posts and Telecommunications
{shuochen365, yanbinbin, xzsang, chenduo, wps1215,2014212810,zclda,wan_huaming }@bupt.edu.cn

Abstract

Neural Radiance Fields (NeRF) can implicitly represent 3D-consistent RGB images and geometric by optimizing an underlying continuous volumetric scene function using a sparse set of input views, which has greatly benefited view synthesis tasks. However, NeRF fails to estimate correct geometry when given fewer views, resulting in failure to synthesize novel views. Existing works rely on introducing depth images or adding depth estimation networks to resolve the problem of poor synthetic view in NeRF with fewer views. However, due to the lack of spatial consistency of the single-depth image and the poor performance of depth estimation with fewer views, the existing methods still have challenges in addressing this problem. So this paper proposes Bidirectional Optical Flow NeRF(BOF-NeRF), which addresses this problem by mining optical flow information between 2D images. Our key insight is that utilizing 2D optical flow images to design a loss can effectively guide NeRF to learn the correct geometry and synthesize the right novel view. We also propose a view-enhanced fusion method based on geometry and color consistency to solve the problem of novel view details loss in NeRF. We conduct extensive experiments on the NeRF-LLFF and DTU MVS benchmarks for novel view synthesis tasks with fewer images in different complex real scenes. We further demonstrate the robustness of BOF-NeRF under different baseline distances on the Middlebury dataset. In all cases, BOF-NeRF outperforms current state-of-the-art baselines for novel view synthesis and scene geometry estimation.

Introduction

Synthesizing novel views from sparse views is a classic ill-posed problem in computer vision, requiring explicit representation of the 3D scene. Unlike explicit 3D scene neural representation methods, implicit 3D scene representation has recently become popular to represent 3D scenes and synthesize novel views continuously. The core idea in implicit 3D scene representation is to use multilayer perceptrons(MLPs) to map 5D coordinates to color and volume density continuously.

NeRF(Mildenhall et al. 2020) is the most representative method in implicit 3D scene representation, which can synthesize photorealistic novel views. Some works optimize NeRF from the speed and quality of view synthesis(Barron et al. 2022; Niemeyer et al. 2022). At the same time, NeRF is also used to solve camera pose, surface reconstruction(Oechsle, Peng, and Geiger 2021; Long et al. 2022), dynamic scene reconstruction, inverse rendering(Yang et al. 2022), and other tasks. However, synthesizing novel views with fewer views remains challenging. The main challenge is depth ambiguity, which means the error depth will erroneously overfit the fewer input images scene. The ambiguity leads to synthesizing wrong novel views.

To overcome such drawbacks of NeRF, current researchers mainly add depth images or depth estimation networks before NeRF network input. Depth-supervised NeRF (Deng et al. 2021) utilizes structure-from-motion (SFM)(Schonberger and Frahm 2016a) to produce sparse 3D points that be used as depth supervision during training. However, SFM can only extract limited and erroneous 3D points in complex scenes, resulting in the inability of NeRF to correctly estimate the scene geometry, thus affecting the quality of novel views. In addition, pixel-NeRF (Yu et al. 2021b) uses the convolutional network to build images feature volume to condition NeRF on the image input. MVS-NeRF (Chen et al. 2021) leverages plane-swept cost volumes for geometry-aware scene reasoning and then combines with NeRF. However, both pixel-NeRF and MVSNeRF need to construct image feature volumes to estimate the scene geometry, resulting in a large memory occupation.

In addition, existing methods also can't effectively solve the problem of novel view details loss in NeRF. Therefore, this paper proposes the BOF-NeRF method to solve the problem of poor view synthesis and novel view details loss with fewer input views. Our approach significantly improves the accuracy of NeRF's scene geometry estimation and view synthesis quality with fewer input views through the proposed optical flow supervision loss and view-enhanced fusion method.

The key to our method is to combine easily computable 2D image optical flow graphs as supervision to guide NeRF to estimate scene geometry. More specifically, the forward and backward optical flow graphs between 2D images are computed to supervise left and right views to learn accurate scene geometry. Since the optical flow graphs between 2D images have errors in non-overlapping and complex back-

*Corresponding author.

ground regions, we combine the camera's intrinsic and color information of the target view to calculate the effective mask in the optical flow graph. Then, the optical flow graph mask is used to determine the weight ratio of optical flow supervision loss. Finally, a two-stage MLP network is used to estimate the geometric information of the scene and synthesize novel views under fewer view inputs.

After estimating the correct scene geometry and synthesizing novel views, we propose a view-enhanced fusion method based on geometry and color consistency to address the problem of NeRF losing novel view details. We use the above network model to synthesize the view and depth image at the given viewpoint. Then, the adjacent left and right input views at the novel viewpoint are determined according to the camera pose. At the same time, the novel depth image and the camera pose are used to determine the corresponding points of the novel view in the adjacent left and right input views. Finally, the difference between the color values of these corresponding points and the novel view is calculated to determine the fusion weight with the novel view, thereby recovering the novel view details. Our contribution can be summarized as follows:

- We propose an optical flow supervision loss designed using easily computable 2D inter-image forward and backward optical flow graphs to address NeRF's problem of wrong scene geometry estimation and poor view synthesis quality under fewer input views.
- We combine the color and depth information of novel views and adjacent input views to propose a view-enhanced fusion method to solve the problem of novel view details loss in NeRF.
- We are able to estimate accurate scene geometry and high-quality novel views on public datasets, which outperform previous state-of-the-art methods favorably.

Related Work

We overview related works from the aspects of explicit 3D scene neural representations, implicit 3D scene neural representations, and optical flow estimation.

Explicit 3D Scene Neural Representations

Previous work mainly addresses scene-specific view synthesis, such as intermediate image synthesis under different shooting angles of the same horizon (Chen and Williams 1993) and image synthesis with prior geometric information (Debevec, Taylor, and Malik 1996). Recently, Convolution Neural Network(CNN) has significantly progressed the task of view synthesis in complex scenes by combining the light field theory to construct the explicit volume neural scene representation. Among them, Multi-plane Image(MPI), Layered-depth Image(LDI) and voxel-base are the most representative methods in explicit 3D scene neural representation.

The core of MPI is to use a CNN to predict a set of front-parallel RGBα planes with uniform disparity sampling within the frustum of each input view and then blend neighboring MPIs by homography warping and alpha compositing to synthesize novel views. In addition to being used for view synthesis under sparse views, the MPI method is also used in single-view view synthesis (Tucker and Snavely 2020) and point cloud rendering (Dai et al. 2020). However, the MPI method is limited by the number of α planes and the image resolution. To address the limitations of MPI methods, Srinivasan (Srinivasan et al. 2019) proposed to build a 3D CNN to increase the frequency of MPI disparity sampling, while Flynn (Flynn et al. 2019) proposed a learning-based gradient algorithm to update and refine the MPI.

Unlike MPI methods, which predict the scene globally and thus make it challenging to process high-resolution images, LDI predicts a more compact local scene representation. LDI stores multiple ordered depths of each pixel and color information corresponding to each depth. Tulsiani1 (Tulsiani, Tucker, and Snavely 2018) proposed to achieve view synthesis from a single view by leveraging more naturally available multi-view supervisory signals. Shih (Shih et al. 2020) proposed a learning-based inpainting model to convert a single RGBD image into 3D images. However, the LDI method can't address the non-Lambertian reflection effects and high-quality view synthesis in occluded regions.

Voxel-based methods optimize CNNs and sample voxel grids to synthesize novel views. Lombardi (Lombardi et al. 2019) proposes a 3D volume representation network from 2D multi-view and a differentiable ray method supporting end-to-end training for dynamic content rendering. Sitzmann1 (Sitzmann et al. 2019) achieves novel view synthesis without 3D supervision by incorporating an adversarial loss function to embed Cartesian 3D grid features. However, voxel-based methods are hard to achieve a trade-off between novel view quality and memory usage.

Implicit 3D Scene Neural Representations

Recently, neural networks have been used to implicitly represent 3D scenes, which can greatly progress surface reconstruction (Genova et al. 2019, 2020)and photorealistic view synthesis tasks while consuming less memory occupation. The 3D mesh or posed depth image is input to the CNN network, which uses the 3D data as supervision to implicitly represent the 3D scene to achieve the surface reconstruction. But none of these methods could achieve photorealistic view synthesis until NeRF was proposed.

NeRF significantly progresses the task of view synthesis by combining fully connected networks, position encoding, and ray integration theory, which is currently the most popular method in the neural implicit 3D representation method. NeRF uses camera intrinsic and extrinsic parameters to determine the 5D coordinates of the ray direction and sampling point position. Then, the 5D coordinate positions are encoded into the network model and predicted by ray integration to output RGB values. So NeRF synthesizes photorealistic novel views that are not limited by the input image resolution.

However, NeRF still has shortcomings to be solved, such as requiring more training views, poor model generalization, slow rendering speed, long training time, and so on. Some works have proposed solutions for the shortcomings of NeRF to improve its performance. Some works achieve view synthesis with fewer views by adding depth images or

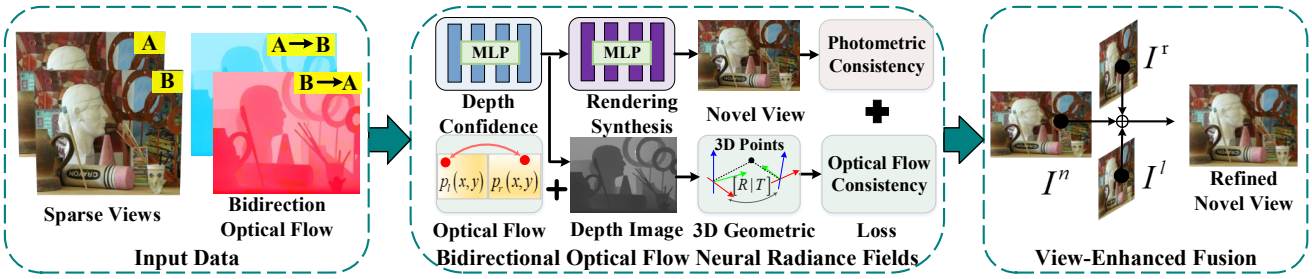

Figure 1: The overview of our BOF-NeRF method. Assuming an input two-view sparse image, our method includes a bidirectional optical flow neural radiance fields and view-enhanced fusion.

depth estimation networks. The IBRNet network framework (Wang et al. 2021) is proposed to learn a generic view interpolation function by combining MLP and ray transformation. KiloNeRF (Reiser et al. 2021) uses a divide-and-conquer approach to use hundreds of small MLPs instead of large MLPs to improve rendering speed. PlenOctrees (Yu et al. 2021a), an octree-based 3D representation of data, is proposed for real-time rendering. A multi-resolution hash coding method is proposed to improve the training speed of NeRF (Müller et al. 2022). At the same time, NeRF is combined with meta-learning to improve the convergence speed of the network(Tancik, Mildenhall, and Wang 2021). However, none of the current works can well solve scene reconstruction and view synthesis with fewer views. Here we can effectively solve this problem by combining the 2D inter-image optical flow graph.

Optical Flow Estimation

Optical flow has critical applications in scene geometry reconstruction and depth estimation. Traditional optical flow algorithms (Chen and Koltun 2016) mainly optimize the trade-off between the data term of visually similar regions and the regularization term imposed by the motion plausibility prior. But it is challenging to solve the accurate calculation of optical flow in complex scenes. Recently deep learning has been used to extract high-dimensional features from 2D images to estimate optical flow images, greatly progressing the optical flow estimation task. Teed (Teed and Deng 2020) proposed the recurrent all-pairs field transforms method to efficiently and robustly estimate the optical flow graph.

In the paper, we propose incorporating 2D image optical flow to supervise and guide NeRF networks to learn scene structure with fewer views. At the same time, the paper proposes a view-enhanced fusion method based on geometry and color consistency. This paper solves the problem of poor view synthesis quality under fewer views and offers an efficient and practical approach to solving the novel view details loss problem in NeRF.

Method

Overview

We first introduce the problem to be solved and the overall scheme of our BOF-NeRF method. Given fewer views I_i, our approach aims to implicitly represent the correct geometry of the scene and synthesize novel views. Fig. 1 illustrates the main modules' bidirectional optical flow neural radiance fields and view-enhanced fusion of our approach. The core step of our method is the bidirectional optical flow neural radiance fields, which enables the network to estimate the correct geometry and novel views under fewer views.

We leverage the optical flow network to estimate the forward and backward optical flow between images to obtain the corresponding points between 2D images. Then, the effective optical flow area is determined by combining the camera's intrinsic parameters and the target view color. Finally, an optical flow loss is constructed using the processed valid optical flow graph, combined with the photometric error loss to guide the network to estimate scene geometry and synthesize novel views. The effectiveness of the BOF-NeRF network comes from the proposed optical flow loss to guide the network to learn the correct scene geometry and synthesize correct novel views with fewer views. In addition, the proposed view enhanced fusion method effectively solves the novel view detail loss problem for NeRF synthesis.

NeRF Volumetric Rendering

Let's briefly review the basic principles and implementation of NeRF. NeRF is a method that does not require geometric information for supervision and only uses input views as supervision. It first performs spatial continuity on the captured scene using the camera's intrinsic and extrinsic parameters to obtain the rays corresponding to the input views. Then NeRF predicts the corresponding RGB c and volume density σ values through a two-stage MLP. Finally, the ray integration obtains the corresponding color value $\hat{C}(r)$ and depth value $\hat{D}(r)$ of each ray:

$$\hat{C}(r) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))c_i \quad (1)$$

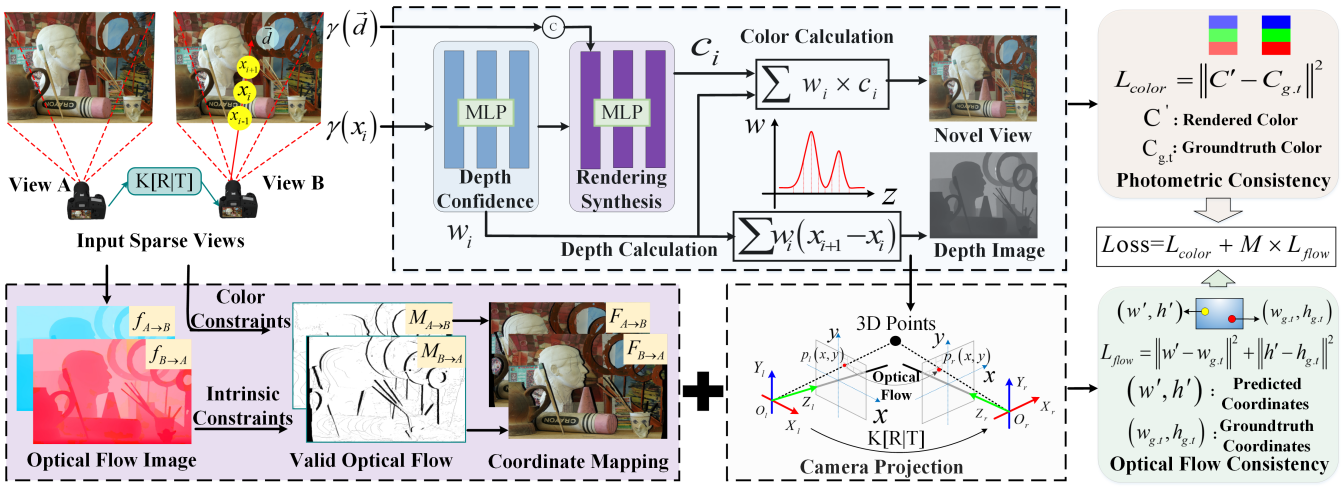

Figure 2: Illustration of our bidirectional optical flow neural radiance fields module. First, we use the COLMAP algorithm to estimate camera intrinsic and extrinsic parameters. Second, we use an optical flow network to estimate the forward and backward optical flow between images. At the same time, we determine the effective area of the optical flow image by combining the camera intrinsics and the target view color. Finally, in the cartesian coordinate system, we use optical flow loss and photometric consistency loss to guide the network to represent the current scene implicitly.

$$\hat{D}(r) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))\delta_i \quad (2)$$

where $T_i = exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j)$ represents the occlusion relationship of the content in the space, δ_i is the distance between adjacent samples. The first-stage network sampling method of NeRF is uniform sampling, and the second-stage sampling method is important and uniform joint sampling. The loss function of NeRF is the sum of the mean square error of the color value predicted by the two-stage network and the true color value:

$$L_{color} = \sum_{r\in R(p)} ||\hat{C}_c(r) - C(r)||_2^2 + ||\hat{C}_r(r) - C(r)||_2^2 \quad (3)$$

where $\hat{C}_c(r)$ is the color value predicted by the coarse network, $\hat{C}_r(r))$ is the color value predicted by the refined network.

Limitations Although NeRF improves the synthesis quality compared with the previous method, there are still some problems. NeRF requires more input views to synthesize novel views that satisfy the geometric relationship. At the same time, it can't correctly estimate the scene geometry and synthesize correct novel views under fewer views. In addition, NeRF is prone to losing novel view details, resulting in blurred synthetic views.

Bidirectional Optical Flow Neural Radiance Fields

To address the inability of neural radiance fields to synthesize correct and photorealistic novel views under fewer views, we propose bidirectional optical flow neural radiance fields, as shown in Fig.2. This section will introduce the specific steps of bidirectional optical flow NeRF. At the same time, we will detail the derivation process and related concepts of the proposed optical flow loss function.

Estimating Camera Pose. In the view synthesis task, the camera pose's accuracy affects the network's degree of convergence and the quality of the synthesized views. Currently, there are two main methods for camera internal and external parameter calibration: checkerboard and COLMAP (Schonberger and Frahm 2016b). The checkerboard method requires a checkerboard to be placed in the scene, resulting in a limited scene. However, the input sparse views I_i in the view synthesis task are often natural unconstrained scenes, so this paper uses the COLMAP algorithm to calculate the camera's intrinsic K_i and extrinsic $[R_i|T_i]$ parameters.

Estimating Image Optical Flow. To solve the problem of view synthesis in NeRF with fewer views, we need to mine the information between input images. Monocular depth estimation networks can estimate depth images for each input view but lack depth consistency between images. At the same time, traditional binocular disparity estimation methods have mismatches at small objects and require filtering. So these methods can't be used to extract information between images. Instead, we found that the state-of-the-art optical flow estimation algorithm RAFT can calculate dense and reasonable correspondences between adjacent wild images. Here, the forward and backward optical flow graphs between adjacent images are calculated.

Optical Flow Mask. Due to the influence of the occlusion relationship and lighting conditions in the captured scene, the calculated optical flow graph has errors in some areas. Therefore, we need to calculate and distinguish between the valid optical flow area and the invalid optical flow area. Assume that the forward optical flow graph of the two-view input image is $f_{a\to b}$, and the backward optical flow graph is $f_{b\to a}$. So the corresponding point coordinates $F_{a\to b}(r, c)$

and $F_{b\to a}(r,c)$ from the original image to the target image in the forward optical flow and the backward optical flow:

$$\begin{aligned} F_{a\to b}(r,c) &= (r,c) + f_{a\to b}(r,c) \\ F_{b\to a}(r,c) &= (r,c) + f_{b\to a}(r,c) \\ r \in [0,H], c &\in [0,W] \end{aligned} \tag{4}$$

where (r,c) is the image coordinates. Considering the invalid optical flow area, the corresponding point of the current image usually exceeds the image resolution. Therefore, the camera intrinsic is used to determine the effective and ineffective areas of optical flow. At the same time, to further ensure the accuracy of optical flow. We propose to use the target color as a constraint to judge the effectiveness of optical flow. So the forward optical flow mask obtained by preprocessing is:

$$\begin{cases} M_{a\to b}(r,c) = 1, if d \le \theta \, and \, F_{a\to b}(r,c) \, in \, [W,H] \\ M_{a\to b}(r,c) = 0, \qquad\qquad ELSE \\ d = abs(Flow_{rgb} - True_{rgb}) \end{cases} \tag{5}$$

where $Flow_{rgb}$ is the corresponding point of the original image in the target image. $True_{rgb}$ is the target ground-truth image of optical flow. d is the absolute difference between the predicted optical flow value and the target's true value. θ is the threshold between optical flow prediction and ground truth.

Optical Flow Loss. To use optical flow graphs to guide the view synthesis network, we need to obtain depth images from different views. Here, we use the depth calculation formula (2) in NeRF to get depth images $depth$ for different views. In NeRF, each ray is represented by the 3D position r_o and the direction r_d. Therefore, the 3D space coordinates can be calculated from the depth:

$$[x,y,z] = r_o + r_d * depth \tag{6}$$

Here, the 3D coordinates need to be mapped to the adjacent left and right input views.

$$[w'h'] = [\frac{u}{z}\frac{v}{z}], \quad [uvz] = KR_1(T_1 + xyz) \tag{7}$$

where w' and h' are the column and row of the depth image predicted by the view synthesis network in the adjacent left and right camera image plane. $[R_1|T_1]$ is the rotation and translation matrix from world coordinates to camera coordinates. The mean square error between the predicted value and the ground truth value $(w_{g.t.}, h_{g.t.})$ of optical flow prediction is used as the optical flow loss:

$$L_{flow} = \sum ||w' - w_{g.t.}||^2 + ||h' - h_{g.t.}||^2 \tag{8}$$

We also add a photometric error loss function to guide the network to synthesize novel views. Therefore, the final loss of the optical flow neural radiance fields is:

$$L = L_{color} + M * L_{flow} \tag{9}$$

where M is the optical flow mask.

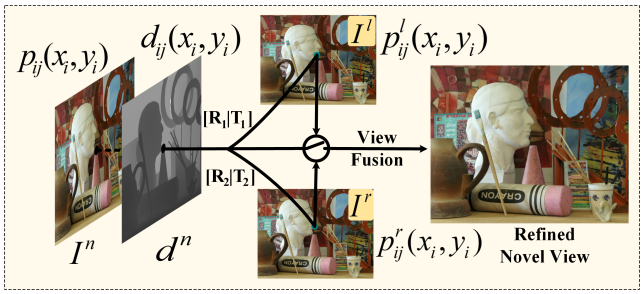

Figure 3: Illustration of our view-enhanced fusion module. We use the pose and depth image of the novel viewpoint to compute the position of the novel view pixels in the left and right adjacent input views. According to the color-constrained threshold w, we fuse the RGB values of the novel view and neighboring input views.

View-Enhanced Fusion

After bidirectional optical flow neural radiance fields processing, we can synthesize correct and photorealistic novel views under fewer input views. However, due to the limitation of the NeRF method itself, the synthesized novel views lose image details. Here, we combine adjacent input images to refine the novel view details of the synthesized views, as shown in Fig. 3. The adjacent left I_{left} and right I_{right} input images are first determined according to the camera pose of the novel view I_{novel}. Second, we utilize the depth image of the novel view from the camera pose to project the novel view coordinates $[x,y,z]$ to the left $p^l_{ij}(x_i,y_j)$ and right $p^r_{ij}(x_i,y_j)$ input images:

$$\begin{cases} p^l_{ij}(x_i,y_j) = K[R_1|T_1][x,y,z]^T \\ p^r_{ij}(x_i,y_j) = K[R_2|T_2][x,y,z]^T \end{cases} \tag{10}$$

Finally, we decide whether to update novel view pixel values according to the difference between the adjacent left I_{left} and right I_{right} pixels of the novel view pixel I_{novel}.

$$\begin{cases} I_{novel} = \dfrac{I_{novel} + M_l * I_{left} + M_r * I_{right}}{1 + M_l + M_r} \\ M_l = 1, \, if abs(p^l_{ij}(x_i,y_j) - p^n_{ij}(x_i,y_j)) \le w \\ M_r = 1, \, if abs(p^r_{ij}(x_i,y_j) - p^n_{ij}(x_i,y_j)) \le w \end{cases} \tag{11}$$

where $p^n_{ij}(x_i,y_j)$ is the novel view pixel value at image coordinates (x_i,y_j). w is the similarity threshold between the pixel values of the novel view and neighboring input views. M_l and M_r are the weights of pixel updates for the left and right views in the novel view.

Experiments

In this section, to evaluate the effectiveness of the proposed method in this paper, we select three typical public datasets for comparison with current mainstream algorithms. First, we quantitatively and qualitatively evaluate the synthetic view quality of our method under fewer input views. Second, we evaluate the synthesis quality of two input views

Horns	PSNR ↑			SSIM ↑			LPIPS ↓		
	2-view	4-view	6-view	2-view	4-view	6-view	2-view	4-view	6-view
NeRF	16.39	22.01	22.39	0.8906	0.9570	0.9612	0.54	0.46	0.45
pixel-NeRF	22.62	21.26	22.06	0.9583	0.9422	0.9518	0.47	0.51	0.47
DS-NeRF	23.19	24.01	24.56	0.9669	0.9708	0.9727	0.45	0.46	0.45
BOF-NeRF	**24.75**	**25.09**	**25.74**	**0.9760**	**0.9767**	**0.9787**	**0.25**	**0.29**	**0.27**
w/o Refine	24.58	24.85	25.49	0.9738	0.9742	0.9766	0.36	0.41	0.40

Table 1: Quantitatively compare the view synthesis quality of Horns scenes in the NeRF-LLFF dataset under different fewer-input views. Our proposed algorithm outperforms other algorithms in PSNR, SSIM, and LPIPS.

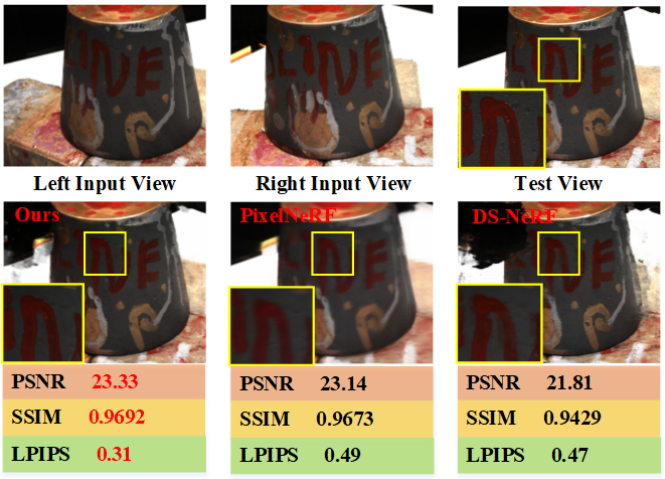

Figure 4: Qualitative and quantitative comparison with two input images on DTU MVS data set. Our method can synthesize clear novel views, and the three indicators of PSNR, SSIM, and LPIPS are better than other methods.

under different baseline distances. Finally, we evaluate the depth accuracy predicted by the proposed method and the recovery effect of view-enhanced fusion.

Datasets

We test our method on three datasets: Middlebury 2005 Stereo datasets (Hirschmuller and Scharstein 2007), NeRF-LLFF datasets, and DTU MVS datasets. (1) Middlebury 2005 Stereo data set contains a series of equally spaced images with horizontal parallax. Among them, a total of nine datasets are included, and each dataset contains seven views under three different exposures. To demonstrate that our method can be synthesized using fewer views under larger baselines, we conduct test evaluations on the Dwarves scene. (2) NeRF-LLFF data set contains eight sets of randomly shot sparse views of different real scenes. This dataset mainly consists of scenes captured in the forward direction. Here, the Horns complex scene is selected to compare the synthetic view quality under two, four, and six viewpoints with the current fewer views synthesis algorithms. (3) DTU MVS data set consists of 128 different scenes. Each scene has been taken from 49 positions, corresponding to the number of RGB images in each scene. We use rectified images with a resolution of 640×512 each. Here, the scan1 scene is selected to compare the synthetic view quality under two viewpoints with the current fewer views synthesis algorithms. Through the above three typical dataset tests, we demonstrate the effectiveness of the proposed method.

Baselines

We select three typical neural radiance fields algorithms for comparison. NeRF is the vanilla neural radiance fields method. pixel-NeRF synthesizes novel views without explicit 3D supervision by extending NeRF. We also compare our method with the state-of-the-art fewer-view synthesis algorithm (DS-NeRF). It solves the problem with sparse 3D point cloud supervision produced by SFM.

Metrics

We use three standard error metrics, Peak Signal-to-Noise Ratio(PSNR), Structural Similarity(SSIM), and Learned Perceptual Image Patch Similarity(LPIPS), to evaluate the quality of the synthesized views. We use projection error P_{error} to quantitatively evaluate the geometric accuracy estimated by the view synthesis network.

Implementation Details

We run our experiments on a PC with a 3.7 GHz Intel Core i9-10900K CPU, 32GB RAM, and NVIDIA GeForce RTX 3090 GPU. Specifically, our BOF-NeRF network is a two-stage network, where each stage is fully connected by a ReLU network with eight layers and 256 channels. The network model is trained via Pytorch (Paszke et al. 2019) using the adam (Kingma and Ba 2014) optimizer with the learning rate set to 5×10^{-4}. At the same time, three typical NeRF algorithms are all retrained on the three datasets to compare with our proposed algorithm. Among them, pixel-NeRF accelerates the convergence speed of the model in different datasets by loading the pre-trained weights of the DTU. The number of ray samples in all networks is set to 64. The proposed method requires 5.1G video memory.

Comparisons on Fewer Images Input

We first select Horns in the NeRF-LLFF scene to test the performance of our algorithm with fewer input views.

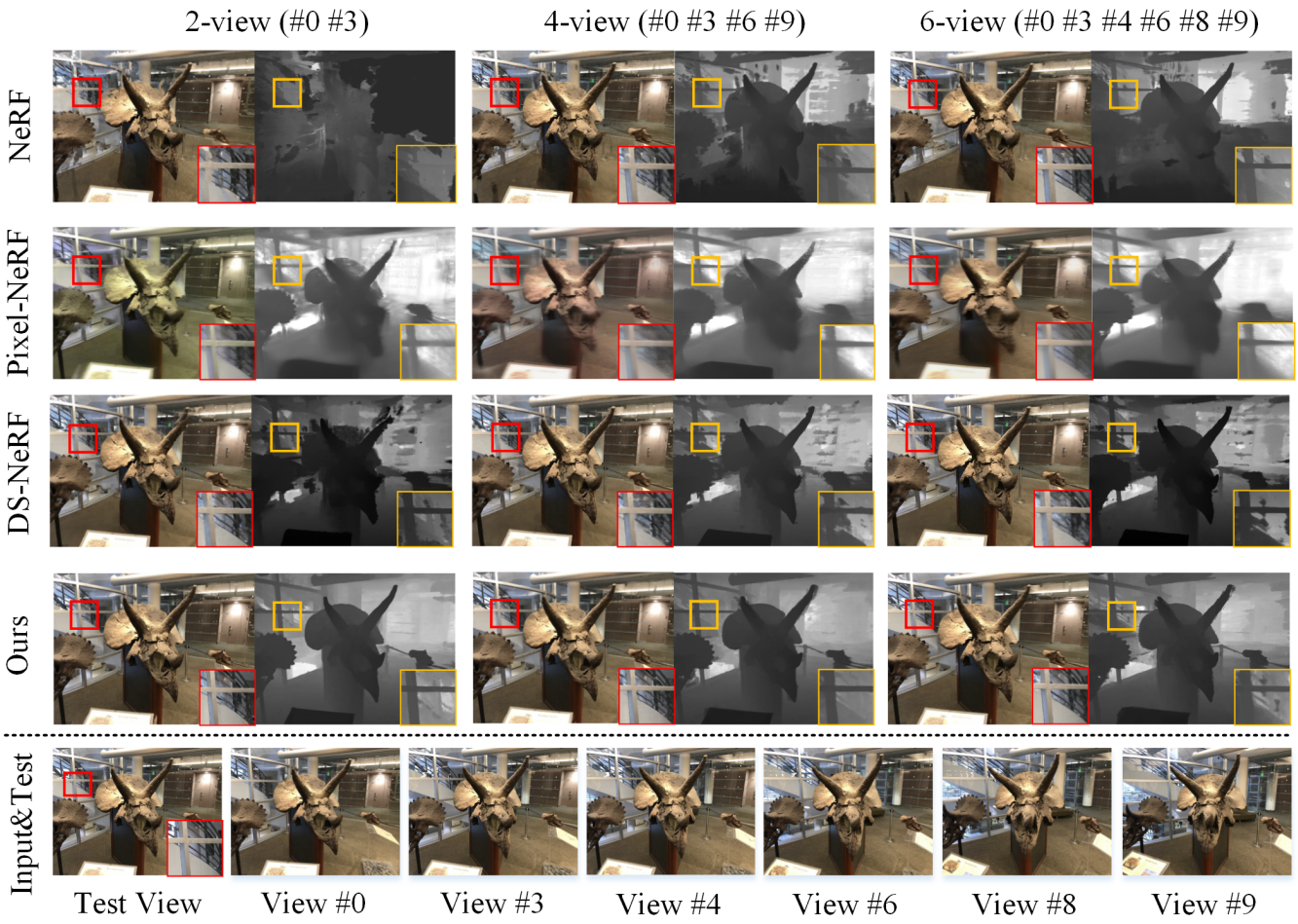

Figure 5: Qualitative comparison of view synthesis quality under different fewer-input views. We train all methods on different input views in the NeRF-LLFF dataset and synthesize color and depth images at test viewpoints. Our proposed method can synthesize refined image details and correct geometric structures under different input views, far exceeding other algorithms.

Middlebury	PSNR ↑				SSIM ↑				LPIPS ↓			
	48cm	64cm	80cm	96cm	48cm	64cm	80cm	96cm	48cm	64cm	80cm	96cm
NeRF	18.50	16.79	15.95	15.92	0.9369	0.9025	0.8885	0.8916	0.44	0.52	0.54	0.54
pixel-NeRF	25.14	23.40	21.92	20.83	0.9760	0.9652	0.9540	0.9375	0.26	0.27	0.33	0.47
DS-NeRF	20.55	19.51	19.49	18.79	0.9383	0.9285	0.9280	0.9227	0.45	0.47	0.48	0.49
BOF-NeRF	25.51	23.57	21.98	21.05	**0.9764**	**0.9663**	0.9555	0.9473	**0.22**	**0.26**	**0.25**	**0.27**
w/o Refine	**25.56**	**23.62**	**22.09**	**21.06**	0.9763	0.9661	**0.9565**	**0.9480**	0.35	0.39	0.32	0.34

Table 2: Quantitatively compare the view synthesis quality of two viewpoints at different baseline distances in the Middlebury scene. Our proposed algorithm outperforms the state-of-the-art algorithm.

From the Fig. 5, it can be observed that 1) our results are more diverse in geometry. That is because we guide the view synthesis network to learn robust scene geometry by mining information between 2D images. 2) The novel views synthesized by our BOF-NeRF method are more detailed and reasonable than other methods. For example, the partially enlarged railing in the picture.

From Tab. 1, we can observe that our method outperforms all comparison methods on three metrics: PSNR, SSIM, and LPIPS. The reason is that our approach can effectively utilize 2D information to learn more accurate geometric details in the view synthesis network. At the same time, our proposed view-enhanced fusion method significantly improves the quality of novel views.

To further demonstrate the proposed method's effectiveness, we conduct quantitative and qualitative comparisons

Figure 6: Rendering quality comparison with different baseline two-image inputs. We train all methods on two-view from different baselines in the Middlebury dataset and synthesize the color and depth images at the test viewpoints. Our proposed method can synthesize clear novel views and smoother depth images than other methods.

ProjectionError(pixel)	Horns($P_{error}\downarrow$)			Middlebury($P_{error}\downarrow$)			
	2-view	4-view	6-view	48cm	64cm	80cm	96cm
NeRF	291.5177	36.0297	35.3826	174.7520	247.6457	317.8365	299.9996
pixel-NeRF	4.2831	4.5553	4.7591	8.0665	9.1144	10.8751	10.6014
DS-NeRF	7.8306	5.0746	4.4831	45.0261	49.2096	53.2879	71.5752
BOF-NeRF	**1.532**	**1.1791**	**1.2961**	**2.028**	**4.1433**	**6.4240**	**9.9601**

Table 3: Quantitative comparison of projection errors of all methods under different scenes, different input viewpoint numbers, and different baseline distances between two viewpoints. Our proposed method significantly outperforms other approaches.

with other algorithms on the DTU MVS data set, as shown in Fig. 4. We can observe from the local zoom-in region that the novel views synthesized by the proposed method are very clear. In addition, PSNR, SSIM, and LPIPS indicators are significantly better than other methods. The main reason is that the optical flow loss can guide the network to learn better scene geometry information.

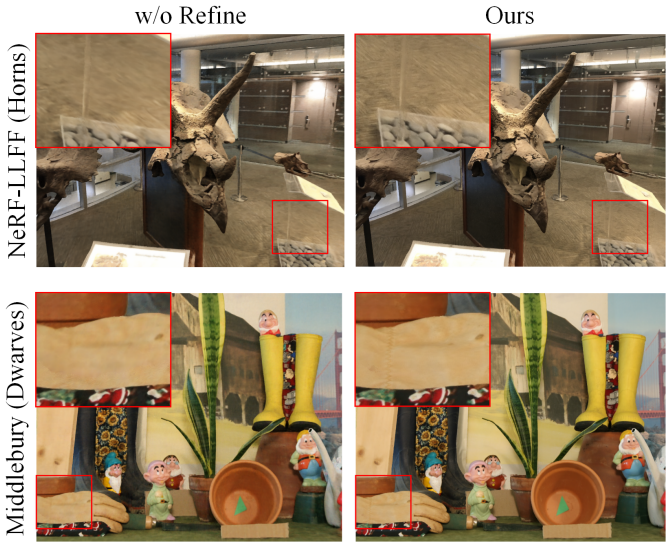

Figure 7: Qualitative comparison of view-enhanced fusion. Our proposed method can recover novel images details.

Comparisons on Different Baseline

We select two viewpoints of the Dwarves scene in the Middlebury dataset with different baseline distances to train all methods and use the intermediate views for testing.

As shown in Fig. 6, our proposed method can synthesize refined novel views and estimate correct scene geometry under different baseline distances from two viewpoints. Meanwhile, our proposed method can observe high-frequency information in synthetic images and edge information in depth and color images. However, NeRF and DS-NeRF methods synthesize wrong novel views and wrongly estimate scene geometry after training, and the quality is worse as the baseline distance increases. In contrast, pixel-NeRF can synthesize correct novel view and depth images with fine-tuning under DTU weights. However, the novel views synthesized by pixel-NeRF are blurred at the edges, and the scene geometry estimation is inaccurate.

From Tab. 2, it can be observed that the three metrics of our method are significantly better than the other methods. Among them, the proposed view-enhanced fusion method can significantly reduce the LPIPS metrics in synthetic views. This is because our proposed method rapidly learns scene-accurate geometry, thereby improving the quality of the synthesized views. In addition, our proposed view-enhanced fusion method based on geometry and color consistency further improves the novel view quality.

Evaluate Depth and High-Frequency Information

Depth Prediction We use the projection error P_{error} to evaluate the scene depth estimation accuracy of all methods under different scenes, input viewpoints, and baseline distances between the two views. As shown in the projection error in Tab. 3, our method achieves the smallest projection error in all the above cases. This further demonstrates the effectiveness of our proposed method.

View-Enhanced Fusion In Tab. 1 and Tab. 2, we have demonstrated by quantitative metrics that our proposed view-enhanced fusion algorithm significantly improves the quality of synthetic views. Here, we qualitatively compared novel view details recovery. From the Horns scene in Fig. 7, our proposed method can recover more texture content. Additionally, in the Dwarves scene, our approach can restore the lines of the glove.

Limitations and Future Work

Although BOF-NeRF is significantly better than other previous works, it is not perfect. In untextured or repeatedly textured scenes, the quality of views synthesized by our method suffers. The proposed method can obtain interpolated novel views but can't address extrapolated novel views. In the future, we will focus on addressing extrapolated view synthesis and view synthesis under weak textures.

Conclusion

In this paper, we propose to combine image optical flow data to guide the network to learn scene geometry and synthesize novel views. Experimental results on indoor and outdoor images show that: 1) It can solve the novel view-incorrect problem of NeRF synthesis under fewer input views without adding any auxiliary devices. 2) We outperform previous neural radiance fields methods on such problems, estimating the correct geometry of the scene and synthesizing correct novel views. 3) Experiments also show that this method can effectively solve the problem of novel view details loss of NeRF method and improve the quality of view synthesis. Overall, we believe our work is critical for understanding NeRF methods that can be applied to image-based 3D reconstruction and relighting.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. 62175017, 62075016).

References

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Chen, A.; Xu, Z.; Zhao, F.; and Zhang. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14124–14133.

Chen, Q.; and Koltun, V. 2016. Full flow: Optical flow estimation by global optimization over regular grids. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4706–4714.

Chen, S. E.; and Williams, L. 1993. View interpolation for image synthesis. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, 279–288.

Dai, P.; Zhang, Y.; Li, Z.; Liu, S.; and Zeng, B. 2020. Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7830–7839.

Debevec, P. E.; Taylor, C. J.; and Malik, J. 1996. 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*, 11–20.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2021. Depth-supervised nerf: Fewer views and faster training for free. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Flynn, J.; Broxton, M.; Debevec, P.; DuVall, M.; Fyffe, G.; Overbeck, R.; Snavely, N.; and Tucker, R. 2019. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2367–2376.

Genova, K.; Cole, F.; Sud, A.; Sarna, A.; and Funkhouser, T. 2020. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4857–4866.

Genova, K.; Cole, F.; Vlasic, D.; and Sarna. 2019. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7154–7164.

Hirschmuller, H.; and Scharstein. 2007. Evaluation of cost functions for stereo matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lombardi, S.; Simon, T.; Saragih, J.; Schwartz, G.; Lehrmann, A.; and Sheikh, Y. 2019. Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics SIGGRAPH*.

Long, X.; Lin, C.; Wang, P.; Komura, T.; and Wang, W. 2022. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, 210–227. Springer.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 405–421. Springer.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5480–5490.

Oechsle, M.; Peng, S.; and Geiger, A. 2021. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5589–5599.

Paszke, A.; Gross, S.; Massa, F.; and Lerer. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*.

Reiser, C.; Peng, S.; Liao, Y.; and Geiger, A. 2021. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14335–14345.

Schonberger, J. L.; and Frahm, J.-M. 2016a. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schonberger, J. L.; and Frahm, J.-M. 2016b. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113. IEEE.

Shih, M.-L.; Su, S.-Y.; Kopf, J.; and Huang, J.-B. 2020. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8028–8038.

Sitzmann, V.; Thies, J.; Heide, F.; and Nießner. 2019. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2437–2446.

Srinivasan, P. P.; Tucker, R.; Barron, J. T.; Ramamoorthi, R.; Ng, R.; and Snavely, N. 2019. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 175–184.

Tancik, M.; Mildenhall, B.; and Wang. 2021. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2846–2855.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, 402–419. Springer.

Tucker, R.; and Snavely, N. 2020. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 551–560.

Tulsiani, S.; Tucker, R.; and Snavely, N. 2018. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 302–317.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; and Barron. 2021. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Yang, W.; Chen, G.; Chen, C.; Chen, Z.; and Wong, K.-Y. K. 2022. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In *European Conference on Computer Vision*, 266–284. Springer.

Yu, A.; Li, R.; Tancik, M.; Li, H.; Ng, R.; and Kanazawa, A. 2021a. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5752–5761.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021b. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.