MASKNERF: MASKED NEURAL RADIANCE FIELDS FOR SPARSE VIEW SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although Neural Radiance Fields (NeRF) has achieved impressive 3D reconstruction with dense views, its performance degrades significantly when the training views are sparse. We observe that under the sparse view setting, it is important to learn the correspondence of pixels among different views, i.e., the 3D consistency, to improve the reconstruction quality. To achieve this, we first propose the Hard-Mask that utilizes the depth information to locate pixels with correspondence relationship and then assigns higher loss weights on these pixels. The key idea is to achieve pixel-wise differentiated optimization of NeRF based on the 3D consistency among target views and source views instead of treating each pixel equally. This optimization strategy helps NeRF-based algorithms to learn fine-grained object details with limited data. To deal with the absence of accurate depth information, the Degenerated Hard-Mask is proposed to estimate the correspondence relationship based on the trend of training losses. Our proposed method can serve as a plug-in component for existing NeRF-based view-synthesis models. Extensive experiments on recent representative works, including NeRF (Mildenhall et al., 2020), IBRNet (Wang et al., 2021) and MVSNeRF (Chen et al., 2021), show that our method can significantly improve the model performance under sparse view conditions (e.g., up to 70% improvement in PSNR on DTU dataset).

1 INTRODUCTION

Novel view synthesis is a long-standing problem in computer vision and graphics, which aims to render photo-realistic images of unseen point-views. Recently, with the success of coordinate-based representation learning in 3D vision, the field of novel view synthesis has gained increasing popularity (Jang & Agapito, 2021; Li et al., 2021; Liu et al., 2021; Rematas et al., 2021). In particular, one representative work, Neural Radiance Fields (NeRF) (Mildenhall et al., 2020), produces realistic results through training an coordinate-based neural network with the help of dense nearby views for each static scenario. The key factor of high quality NeRF (Huang et al., 2022) is in the requirement of dense views, where it explicitly learns the correspondences of pixels, i.e., 3D consistency, in different views with supervision from the multi-view images. In contrast, it is difficult for a NeRF model to learn such correspondence relationship in the sparse view setting (as few as three views) as limited supervision information is available, which leads to significantly performance degradation and thus limits the ability to extend NeRF-based models to real-world scenarios.

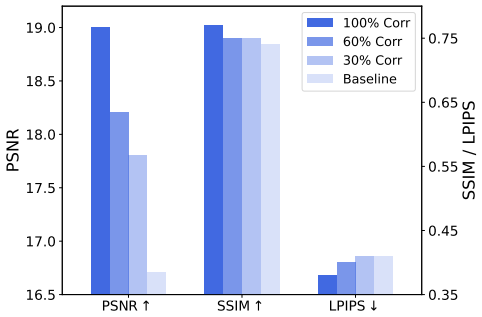


Figure 1: Performance (PSNR↑, SSIM↑, LPIPS↓) comparison of NeRF with different levels of correspondence information. Using more correspondence leads to better model performance.

To alleviate the limitation of NeRF in the sparse view setting, two lines of works have been proposed. The first line of works pre-trains NeRF on large-scale data sets consisting of many scenes and fine-tune the model with sparse view data (Chen et al., 2021; Chibane et al., 2021; Jang & Agapito, 2021; Li et al., 2021; Liu et al., 2021; Rematas et al., 2021; Trevithick & Yang, 2021; Wang et al., 2021; Yu et al., 2021). The other line of works introduces extra regularization into the optimization

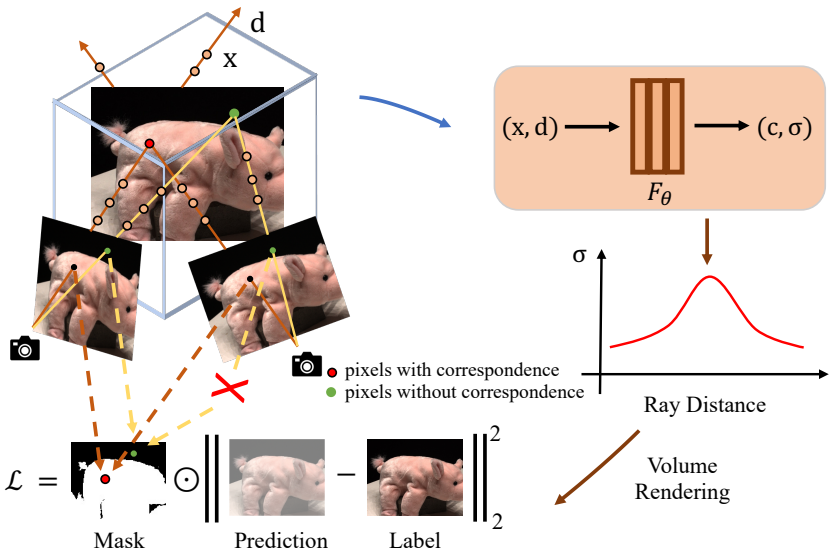


Figure 2: **The demonstration of proposed masks.** We utilize the depth correspondence (or training loss) among different views to mask pixels satisfying 3D correspondence (the red point) or not (the green point) and construct the loss $\mathcal{L}$ based on the mask information.

of NeRF in the sparse view setting, such as the "free" depth supervision (Deng et al., 2021; Jain et al., 2021; Roessle et al., 2021; Niemeyer et al., 2021; Wei et al., 2021; Kim et al., 2022). These works usually focus on the optimization of pixel-level color and depth within a single view, but ignore the 3D consistency relationship of pixels among different views. However, 3D consistency is important especially in appearance and geometry information reconstruction tasks under sparse view setting, such as monocular adjacent view synthesis (Rockwell et al., 2021) and monocular depth estimation (Godard et al., 2019; Zhou et al., 2022).

The 3D consistency, i.e., the 3D correspondence between source and target views, refers to the relationship of a set of pixels that are obtained by projecting the same 3D scene point into different views (one pixel in one view). For pixels satisfying the correspondence relationship, the predicted color must look similar and the predicted depth must satisfy the homography warping relationship. As shown in Fig. 1, we observe consistent performance improvements when we gradually increase the amount of 3D correspondence information in the optimization of NeRF under the sparse view setting. This evaluation validates the importance of 3D consistency (see more details in Sec. 3.2). However, how to explicitly incorporate the 3D consistency into the optimization process of NeRF remains as a challenging problem. Consequently, we propose to explicitly incorporate the correspondence relationship between source and target views into the optimization process of NeRF to improve the performance in sparse views setting.

Specifically, we propose Masked Neural Radiance Fields (MaskNeRF), an efficient component that regularizes the optimization of NeRF through masks for sparse view synthesis (as shown in Fig. 2). Instead of treating all pixels equally in target views, we propose the Hard-Mask to enforce the optimization of NeRF optimization to focus more on pixels that satisfy the correspondence relationship. Our Hard-Mask first selects pixels satisfying 3D correspondence between target views and source views based on the geometry information derived from depth and then assign higher loss weights on these pixels during the training process. However, accurate depth information may not be available for some scenarios. To deal with this situation, the Degenerated Hard-Mask is proposed to predict the correspondence relationship based on the trend of the training loss information. Our proposed method serves as a plug-in component for existing NeRF-based view-synthesis methods. Experiments show that our proposed methods can improve the performance of representative NeRF methods, e.g., NeRF (Mildenhall et al., 2020), IBRNet (Wang et al., 2021) and MVSNeRF (Chen et al., 2021) by up to 70% in PSNR, 60% in SSIM and 26% in LPIPS on various datasets.

2 Related Works

To deal with sparse-view scenarios, two lines of research have been proposed to improve the generalizability of NeRF. The first line of research focuses on incorporating prior knowledge, while the other line introduces additional ground-truth information during the training process.

View Synthesis with Prior Knowledge. This line of research takes prior knowledge into account by pre-training neural network with large amount of data, and decreases the need of dense views for rendering novel 3D scenarios. SRF (Chibane et al., 2021), GRF (Trevithick & Yang, 2021), Point-NeRF (Xu et al., 2022) IBRNET (Wang et al., 2021) and PixelNeRF (Yu et al., 2021) make use of pre-trained model to extract feature maps of source views, which are then adopted to form appearance and geometry features for points in target views. Inspired by multi-view depth estimation tasks (Yang et al., 2020), Neural rays (Liu et al., 2022) and MVSNeRF (Chen et al., 2021) exploit pre-trained model to construct a geometry-aware cost volume during the inference process, which are then exploited by a decoder to reconstruct RGB images. Through the use of prior knowledge, these algorithms can effectively deal with the lack of information under sparse views setting. However, an evident performance decrement on the training scenarios with dense views and that on the testing scenarios with sparse views can still be observed. The proposed MaskNeRF incorporates correspondence relationship to regularize the optimization process. In this way, MaskNeRF improves the performance of models under sparse view setting without increasing computational request.

View Synthesis with Additional Information. To deal with limited information provided by sparse views, this line of research introduces additional information to assist view synthesis process. DS-NeRF (Deng et al., 2021) and GeoNeRF (Johari et al., 2022) introduces geometry constrain to the optimization of NeRF with the help of the ground truth depth information or "free" depth extracted from Structure-From-Motion (SFM) solvers like COLMAP (Schonberger & Frahm, 2016). The "free" depth loss can also be constructed with dense depth priors estimated from ScanNet (Roessle et al., 2021). CodeNeRF (Jang & Agapito, 2021), DoubleField (Shao et al., 2022), ShaRF Rematas et al. (2021) and Improving (Darmon et al., 2022) focus on object-centric scenarios with ground truth shape information. They jointly optimize appearance information and shape information, which helps to build better correspondence among views under sparse view setting. Meanwhile, DietNeRF (Jain et al., 2021) proposes to use CLIP (Dosovitskiy et al., 2020) to extract semantic information as the additional supervision. While, the semantic information can only be obtained from low-resolution images and thus can provide high-level information only. In addition, RegNeRF (Niemeyer et al., 2021) and RapNeRF (Zhang et al., 2022) also introduce regularization to further improve the performance of NeRF-based model, while none of them have utilized the across view 3D consistency in their regularization and our proposed strategies could be easily combined with their methods. Even though these algorithms achieve impressive performance under certain conditions, the requirement of additional accurate information is not always feasible for most of the datasets, e.g., on LLFF NeRF Real dataset (Mildenhall et al., 2019) where ground truth depth, shape and semantic information are not provided. To deal with this problem, in addition to the Hard-Mask which requires estimated depth information, we also introduce the Degenerated Hard-Mask which can be directly used without the need of any additional information during the optimization of NeRF-based models.

3 METHOD

3.1 BACKGROUND

Neural Radiance Fields. The Radiance Field learns a continuous function which takes as input the 3D location $\mathbf{x}$ and unit direction $\mathbf{d}$ of each point and predicts the volume density $\sigma \in [0, \infty)$ and color value $\mathbf{c} \in [0, 1]^3$. In NeRF (Mildenhall et al., 2020), this continuous function is parameterized by a multi-layer perception (MLP) network $F_\theta : (\gamma(\mathbf{x}), \gamma(\mathbf{d})) \rightarrow (\mathbf{c}, \sigma)$, where the weight parameters θ are optimized to generate the volume density σ and directional emitted color $\mathbf{c}$, γ is the predefined positional embedding applied to $\mathbf{x}$ and $\mathbf{d}$, which maps the inputs to a higher dimensional space.

Volume Rendering. Given the Neural Radiance Field (NeRF), the color of any pixel is rendered with principles from classical volume rendering (Kajiya & Von Herzen, 1984) the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ cast from the camera origin $\mathbf{o}$ through the pixel along the unit direction $\mathbf{d}$. In volume rendering, the volume density $\sigma(\mathbf{x})$ can be interpreted as the probability density at an infinitesimal distance at location $\mathbf{x}$. With the near and far bounds t_n and t_f, the expected color $\hat{C}_\theta(\mathbf{r})$ of camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is defined as

$$\hat{C}_\theta(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt, \text{ where } T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds), \tag{1}$$

where $T(t)$ denotes the accumulated transmittance along the direction $\mathbf{d}$ from t_n to t. In practice, the continuous integral is approximated by using the quadrature rule (Max, 1995) and reduced to the traditional alpha compositing. The neural radiance field is then optimized by constructing the photometric loss $\mathcal{L}$ between the rendered pixel color $\hat{C}_\theta(\mathbf{r})$ and ground truth color $C(\mathbf{r})$:

$$\mathcal{L} = \frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} ||\hat{C}_\theta(\mathbf{r}) - C(\mathbf{r})||_2^2, \tag{2}$$

where $\mathcal{R}$ denotes the set of rays, and $|\mathcal{R}|$ is the number of rays in $\mathcal{R}$.

3.2 Preliminary: Pixel-wise 3D Correspondence

In this section, we demonstrate the importance of considering the correspondence, i.e., 3D consistency, in the optimization process. With no loss of generality, we define $\mathcal{M}$ to be the set containing pixels satisfying correspondence relationship and $\mathcal{T}$ to be the correspondence relationship between pixel (i, j) and $(m, n) := \mathcal{T}((i, j))$. The 3D appearance consistency is defined in Definition 3.1. Similarly, we also define the 3D geometry consistency and details are shown in Appendix. B. By involving the proposed mask in Sec. 3.3, we select and assign larger loss weights to pixels that satisfy the homography warping relationship between source views and target views, i.e., the correspondence relationship. We compare the performance (PSNR↑, SSIM↑, LPIPS↓) of assigning larger weights to different portions (30%, 60%, 100%) of pixels satisfying the correspondence relationship in the DTU data set. The baseline is the original NeRF model that treats all pixels equally during the optimization process. As shown in Fig. 1, assigning large weights to more pixels satisfying correspondence leads to better model performance. More details can be found in Appendix C.

Definition 3.1 (Appearance Consistency) *The appearance consistency refers to the color difference between the pixel $(i, j) \in \mathcal{M}$ (in the left view of Fig. 3) and its corresponding pixel $(m, n) := \mathcal{T}((i, j))$ (in the right view of Fig. 3) should be smaller than a threshold value ϵ_c, i.e.:*

$$||C_\theta(\mathbf{r}_{ij}) - C_\theta(\mathbf{r}_{mn})||_2^2 \leq \epsilon_c, \tag{3}$$

where $C_\theta(\mathbf{r}_{ij})$ and $C_\theta(\mathbf{r}_{mn})$ are color labels of pixel (i, j) and (m, n).

3.3 Regularizing Neural Radiance Fields Through Masks

Based on the importance of 3D correspondence, we propose two forms of masks to enforce NeRF-based algorithms to focus on the 3D correspondence relationship.

Hard-Mask. Given a series of images for the specific scenario, MaskNeRF derives the Hard-Mask which masks pixels satisfying 3D correspondence relationship between source views and target views. With no loss of generality, we show the derivation of Hard-Mask in two views. As shown in Fig. 3, MaskNeRF samples a bunch of pixels $\{(i, j)\}$ with coordinates $\{\mathbf{x}_{ij}^{lp} = [i, j, 1]^T\}$ in the left camera coordinate, where l denotes the left camera view and p denotes the pixel coordinate. For each pixel (i, j), one camera ray is cast from the camera origin $\mathbf{o}$ along with the ray direction $\mathbf{d}$. With the estimated depth s_{ij}^l of pixel (i, j) in the left camera view, the world coordinate of the intersection point $\mathbf{x}_{ij}^{lw}$ can be derived as

$$\mathbf{x}_{ij}^{lw} = (\mathbf{R}^l)^{-1}\mathbf{K}^{-1} \cdot (s_{ij}^l \cdot \mathbf{x}_{ij}^{lp}), \tag{4}$$

where $\mathbf{R}^l$ is the world-to-camera transformation matrix of the left camera view, $\mathbf{K}$ is the camera intrinsic matrix.

To get the pixel coordinate of the intersection point $\mathbf{x}_{ij}^{lw}$ in the right view, the estimated world coordinate $\mathbf{x}_{ij}^{lw}$ is transformed into the image plane of the right camera view with the world-to-camera transformation matrix $\mathbf{R}^r$ and camera intrinsic matrix $\mathbf{K}$ as follows:

$$s_{mn}^{'} \cdot \mathbf{x}_{mn}^{rc} = \mathbf{K}\mathbf{R}^r\mathbf{x}_{ij}^{lw}, \tag{5}$$

where $\mathbf{x}_{mn}^{rc} = (m, n, 1)$ is the pixel coordinate by projecting the intersection point $\mathbf{x}_{ij}^{lw}$ onto the right camera image plane, $s_{mn}^{'}$ is the estimated depth of the intersection point $\mathbf{x}_{ij}^{lw}$ in the right camera.

Pixels (i, j) and (m, n) are masked as pixels with 3D correspondence relationship when 1) the pixel (m, n) is not out of the boundary of the right image plane and 2) the transformed depth $s_{mn}^{'}$ and

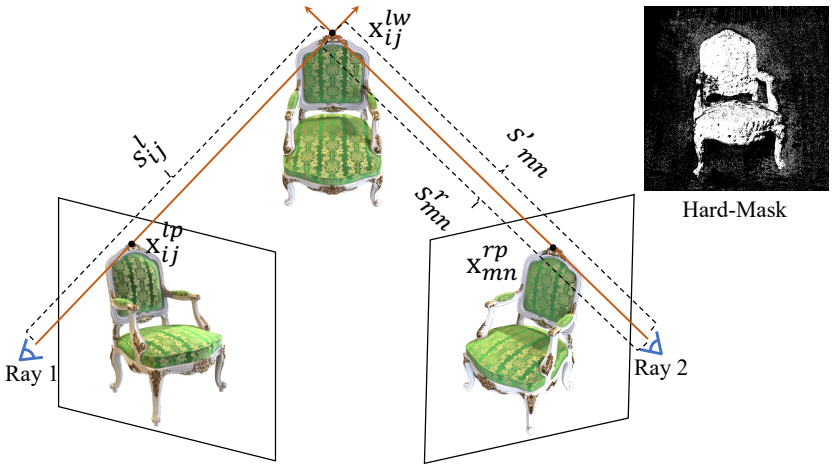


Figure 3: Illustration of deriving Depth-based Hard-Mask. We first derive the world coordinate $\mathbf{x}_{ij}^{lw}$ of pixel (i, j) in the left view and then project the world coordinate $\mathbf{x}_{ij}^{lw}$ into the right view, which leads to pixel (m, n). If the difference of projected depth $s_{mn}^{'}$ and depth label s_{mn}^{r} is less than a threshold α, the pixel (i, j) and (m, n) are marked as pixels satisfying 3D correspondence. By convention, depth is defined as the coordinate value along z axis in the corresponding camera coordinate system.

depth s_{mn}^{r} of pixel (m, n) are sufficiently close. Pixel (i, j) is regarded as a pixel that does not satisfy 3D correspondence under the sparse view setting and is excluded from Hard-Mask when it cannot find a pixel that satisfies the above condition in all training views. Following the above derivation, we set a threshold α to mask pixels with 3D correspondence relationship as follows:

$$|s_{mn}^{r} - s_{mn}^{'}| < \alpha \rightarrow \text{pixel } (i, j) \in \mathcal{M}, \mathcal{T}((i, j)) = ((m, n)). \tag{6}$$

where $\mathcal{M}$ is defined to be the set containing masked pixels and $\mathcal{T}$ defines the correspondence relationship between pixel (i, j) and $(m, n) := \mathcal{T}((i, j))$.

With the derived Hard-Mask, the loss function is defined as

$$\mathcal{L}_h = \frac{1}{|\mathcal{R}|} \left\{ \sum_{\mathbf{r} \in \mathcal{R} \cap \mathcal{M}} ||\hat{C}_\theta(\mathbf{r}) - C(\mathbf{r})||_2^2 + \lambda \sum_{\mathbf{r} \notin \mathcal{R} \cap \mathcal{M}} ||\hat{C}_\theta(\mathbf{r}) - C(\mathbf{r})||_2^2 \right\}, \tag{7}$$

where $\mathcal{R}$ denotes the set of rays, the coefficient $\lambda \ll 1$ controls the loss ratio of emphasizing the pixels satisfying the correspondence relationship.

As shown in Proposition 1, the above loss function, which focuses on the pixels selected by the Hard-Mask, implicitly emphasizes the appearance consistency in the optimization of NeRF. The proof is provided in Appendix A. We also show that it emphasizes the geometry consistency in the optimization of NeRF (see Appendix B for more details).

Definition 3.2 (Consistency of Estimated Appearance) *The consistency of estimated appearance refers to the predicted color difference between pixel $(i, j) \in \mathcal{M}$ and pixel $(m, n) := \mathcal{T}((i, j))$ (in the left/right view of Fig. 3) should be smaller than a threshold value ϵ_c, i.e.:*

$$||\hat{C}_\theta(\mathbf{r}_{ij}) - \hat{C}_\theta(\mathbf{r}_{mn})||_2^2 \leq \epsilon_c, \tag{8}$$

where $\hat{C}_\theta(\mathbf{r}_{ij})$ and $\hat{C}_\theta(\mathbf{r}_{mn})$ are predicted color of pixel (i, j) and (m, n).

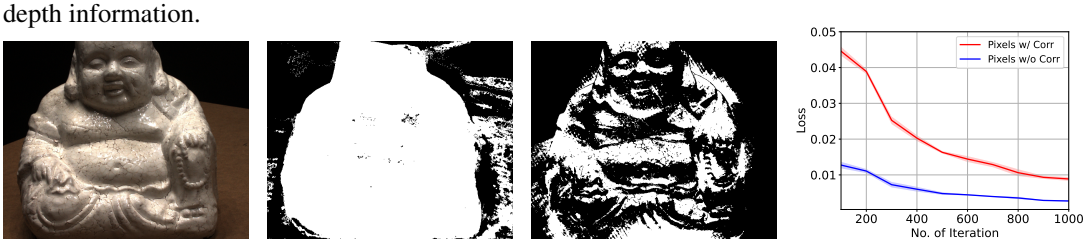
Proposition 1 (Appearance Consistency Regularization) *Directly minimizing appearance consistency in Definition 3.2 leads to trivial solution $\hat{C}_\theta(\mathbf{r}_{ij}) = \hat{C}_\theta(\mathbf{r}_{mn}) = 0$. Focusing on minimizing the errors between predict color values and their ground truth for pixels included by the Hard-Mask as in Eqn. (7) would help to emphasize the appearance consistency:*

$$||\hat{C}_\theta(\mathbf{r}_{ij}) - C(\mathbf{r}_{ij})||_2^2 + ||\hat{C}_\theta(\mathbf{r}_{mn}) - C(\mathbf{r}_{mn})||_2^2 \geq \frac{1}{4}||\hat{C}_\theta(\mathbf{r}_{ij}) - \hat{C}_\theta(\mathbf{r}_{mn})||_2^2 - \epsilon_c/2. \tag{9}$$

The above estimated Hard-Mask locates pixels satisfying 3D correspondence relationship, which enforces NeRF to focus on the optimization of 3D consistency. However, the quality of the estimated

Hard-Mask depends on the precision of the estimated depth. In absense of accurate depth information, we propose a Degenerated Hard-Mask which is directly extracted from the NeRF's training loss.

Degenerated Hard-Mask (DH-Mask). The Degenerated Hard-Mask is proposed based on the observation that pixels satisfying 3D correspondence (masked by Hard-Mask) tend to have a larger loss in the optimization. As shown in Fig. 4, we compare the loss of pixels masked by Hard-Mask, i.e., pixels with correspondence, and the loss of those without correspondence. As can be seen, pixels masked by Hard-Mask have a significantly larger loss (Fig. 4 (d)). Based on the above observation, we extract the DH-Mask by masking pixels with the largest top-K loss values at the beginning of the training stage (e.g., when the number of iteration is 500). The resulting DH-Mask has a large overlapping part when compared with the Hard-Mask. One example of Degenerated Hard-Mask is illustrated in Fig. 4 (c) by setting K to be 30% of the total number of pixels. Experimental results suggest that DH-Mask has competitive performance in comparison with Hard-Mask with accurate depth information.



(a) Image (b) H-Mask (c) DH-Mask (d) Loss Curve

Figure 4: (a) Original image. (b) Hard-Mask. (c) Degenerated Hard-Mask. (d) Loss curve for pixels with correspondence relationship and that of pixels without correspondence relationship.

4 EXPERIMENTS

Datasets. We evaluate our proposed methods on the real-world multi-view DTU dataset (Jensen et al., 2014), Forward-Facing LLFF dataset (Mildenhall et al., 2019) and Realistic Synthetic NeRF dataset (Mildenhall et al., 2020). For DTU, we follow PixelNeRF (Yu et al., 2021) to split the data into 88 training scenes and 16 testing scenes. The 88 DTU training scenes are used to pre-train the IBRNet (Wang et al., 2021) and MVSNeRF (Chen et al., 2021) models. For each testing scene in three data sets, we select 3/6 views from 20 nearby views as training views and randomly select 4 views as validation and testing views. Similar to MVSNeRF (Chen et al., 2021), we evaluate all methods on the DTU dataset with the object masks applied to the rendered and ground truth images.

Evaluation Metrics: For performance comparison, we report the mean of peak signal-to-noise ratio (PSNR) (Sara et al., 2019), structural similarity index (SSIM) (Wang et al., 2004) and Learned Perceptual Image Patch Similarity (LPIPS) perceptual metric (Zhang et al., 2018).

Implementation Details. We compare our methods with state-of-the-art NeRF (Mildenhall et al., 2020), IBRNet (Wang et al., 2021) and MVSNeRF (Chen et al., 2021). We also consider NeRF with depth serving as supervision (Deng et al., 2021) for a fair comparison. For all NeRF (Mildenhall et al., 2020) based methods which do not require pre-training, we directly train the model from scratch for each target scene. In our experiments, we use the depth extracted from a pre-trained MVSNeRF (Chen et al., 2021) to derive the Hard-Mask. For a fair comparison, both NeRF and our method use this depth as the supervision. All mentioned methods (NeRF, NeRF† (NeRF with depth), NeRF + DH-Mask, NeRF + H-Mask) were trained with 50000 iterations. For H-Mask, the threshold α is set to be 0.1 and λ is set to be 0.1 on DTU, LLFF and NeRF Synthetic data set. In DH-Mask, the portion top-K is set as top-50%. For methods that require pre-training, we directly use the released code and checkpoint of MVSNeRF and retrain IBRNet on the DTU data. Then we conduct a test-time optimization on DTU, LLFF and NeRF Synthetic testing scenes. We run each method with four random seeds and report the mean results. More implementation details are provided in Appendix D.

Initialization for Stable Optimization. During our experiments, we observe that NeRF is prone to a catastrophic failure at the initialization stage in which MLP emits negative values before the ReLU activation. In this case, all predicted σ values are zero and gradients back-propagated from the loss function to MLP parameters are zero and thus leads to the failure of the optimization. To address the above failure, Mip-NeRF (Barron et al., 2021) proposed to use a softplus function to yield

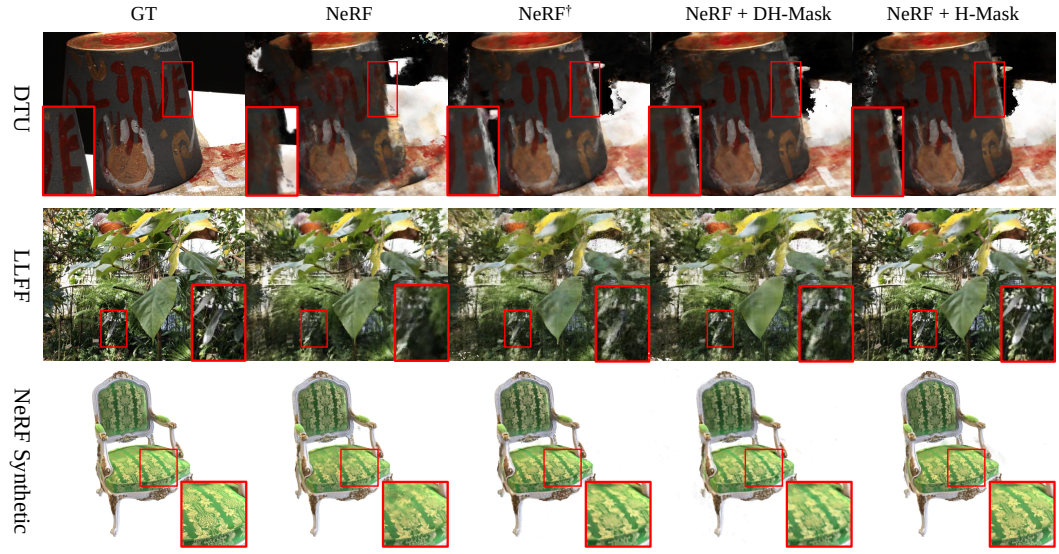


Figure 5: Novel View Synthesis Results on DTU, LLFF and NeRF Synthetic data sets with 3 views as input. We observe that the baselines suffer from blur results, while our Hard-Mask and Degenerated Hard-Mask can produce sharp results with fine-grained details.

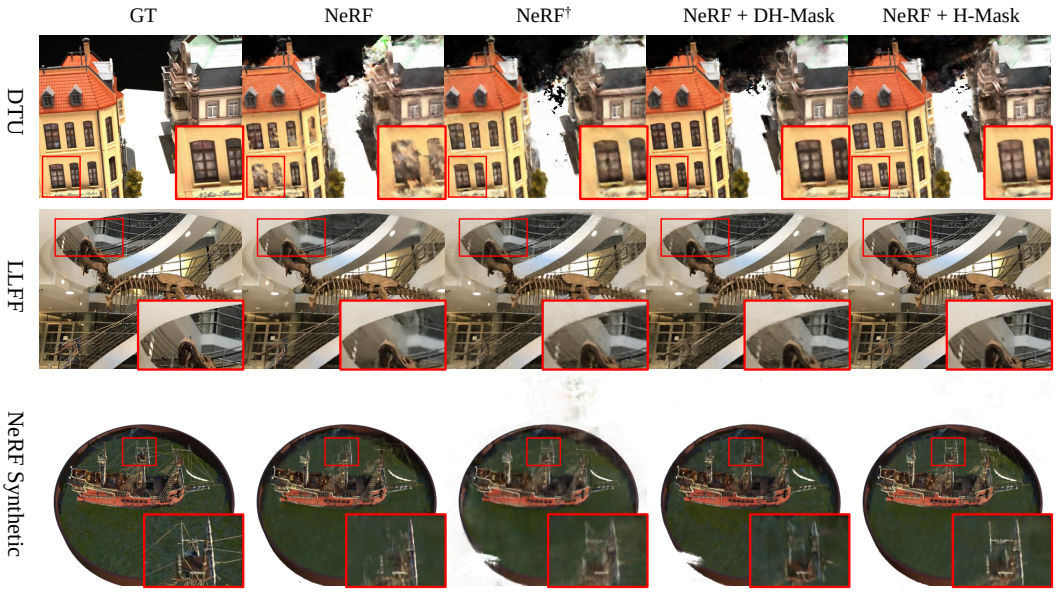


Figure 6: Novel View Synthesis Results on DTU, LLFF and NeRF Synthetic data sets with 6 views as input. With Depth supervision, NeRF† can render 3D object with the correct geometry, but it suffers from rendering sharp details. In contrast, our approach renders both correct geometry and sharp details.

a stable optimization. However, we observe that NeRF overfits to training views by using the softplus function in the sparse view setting. In this paper, we propose to modify the initialization of bias parameters in the MLP to guarantee both stable optimization and good generalization ability. During our experiments, we find initializing the value of bias parameters in MLP using a uniform distribution between 0 and 1 leads to acceptable results. The comparison results are reported in Appendix E.

4.1 View Synthesis Results in the Sparse View Setting

In this experiment, we first evaluate the performance achieved by the original version of above-mentioned models with sparse view settings and then compare it with their improved version using proposed DH-Mask and H-Mask. Quantitative results are shown in Tab. 1. Several conclusions are drawn as follows. (1) For 3 input view settings, our proposed H-Mask could largely improve the performance of the original NeRF, e.g., 70% relative PSNR improvement is achieved on the DTU

Table 1: Performance (PSNR, SSIM and LPIPS) comparison among state-of-the-art NeRF methods on DTU, NeRF Synthetic and Forward-Facing data sets. ↑ means the larger is better; ↓ means the smaller is better.

Method	Setting	Real Data (DTU)			Synthetic Data (NeRF)			Forward-Facing (LLFF)		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF (Mildenhall et al., 2020)	3-view	11.40	0.50	0.49	14.59	0.82	0.29	12.52	0.34	0.60
NeRF† (Deng et al., 2021)		11.80	0.52	0.49	15.13	0.82	0.30	13.10	0.35	0.62
NeRF + DH-Mask		18.90	0.77	0.38	20.71	0.86	0.31	21.19	**0.74**	**0.40**
NeRF + H-Mask		**19.44**	**0.80**	**0.36**	**20.94**	**0.87**	**0.29**	**21.44**	**0.74**	**0.40**
IBRNet (Wang et al., 2021)		19.22	0.83	0.29	14.20	0.75	0.38	21.32	0.77	0.30
IBRNet + DH-Mask		20.17	0.84	**0.28**	14.98	0.78	**0.34**	21.72	0.78	0.30
IBRNet + H-Mask		**20.59**	**0.85**	**0.28**	**15.16**	**0.79**	**0.34**	**22.05**	**0.79**	**0.29**
MVSNeRF (Chen et al., 2021)		19.17	0.80	0.34	15.12	0.82	0.29	18.99	0.68	0.41
MVSNeRF + DH-Mask		22.96	0.89	**0.22**	21.21	0.86	0.25	21.68	0.79	0.30
MVSNeRF + H-Mask		**23.27**	**0.87**	0.25	**21.77**	**0.87**	**0.23**	**21.88**	**0.79**	**0.29**
NeRF (Mildenhall et al., 2020)	6-view	13.77	0.55	0.46	17.35	0.85	0.25	13.71	0.38	0.57
NeRF† (Deng et al., 2021)		14.17	0.56	0.45	17.95	0.87	0.25	14.29	0.39	0.56
NeRF + DH-Mask		21.73	0.83	0.35	23.40	0.91	0.25	22.34	0.76	0.39
NeRF + H-Mask		**22.25**	**0.84**	**0.34**	**23.80**	**0.91**	**0.24**	**22.74**	**0.78**	**0.38**
IBRNet (Wang et al., 2021)		24.51	0.92	0.18	16.54	0.81	0.32	23.45	**0.83**	**0.22**
IBRNet + DH-Mask		25.19	0.91	0.19	17.73	**0.84**	**0.28**	23.66	**0.83**	0.23
IBRNet + H-Mask		**25.36**	**0.93**	**0.18**	**17.78**	**0.84**	**0.28**	**23.67**	**0.83**	**0.22**
MVSNeRF (Chen et al., 2021)		24.48	0.90	0.21	23.19	0.90	0.19	22.65	0.81	0.28
MVSNeRF + DH-Mask		24.61	0.89	0.21	23.93	0.90	0.20	23.07	0.81	0.28
MVSNeRF + H-Mask		**24.95**	**0.90**	**0.20**	**24.18**	0.90	**0.13**	**23.84**	**0.84**	**0.23**

Figure 7: Novel View Synthesis Results produced by MVSNeRF and our methods. MVSNeRF produces results with poor lighting when the testing view is far from the training view while our methods can render images with better lighting effects by utilizing the 3D correspondence between source and target views.

data set. Besides, when compared with NeRF† which directly introduces depth constrain, our Mask could bring larger performance improvement through emphasizing the optimization of pixels with 3D correspondence relationship. (2) By deriving mask information from the trend of training loss, DH-Mask achieves competitive (only a bit worse) results when compared with H-Mask. Both of DH-Mask and H-Mask bring obvious performance improvement for 3-view experiments with variants of NeRF. (3) When given more views (6 views) as input, though overall good rendering results are achieved by baseline methods, our approaches can still produce consistent improvements among three data sets. As shown in Fig. 5 and Fig. 6, our methods can render images with both correct geometry and fine-grained details which is close to the ground truth, while NeRF suffers from blur effects around edges, as blur parts tend to have a large loss and our proposed masks emphasize the optimization of this part to render sharp details.

We also investigate the performance of DH-Mask and H-Mask with IBRNet (Wang et al., 2021) and MVSNeRF (Chen et al., 2021), which require the pre-training and per-scene optimization. As shown in Tab. 1, IBRNet and MVSNeRF produce better results than the vanilla NeRF in the 3/6 view setting. However, we still observe some inconsistent results when the testing view is far from the training views. For example, as shown in Fig. 7, MVSNeRF produces images with poor lighting, while our proposed MaskNeRF can predict the correct lighting effect of the pixels in the target view using the correspondence of pixels with similar lighting effect in source views. Instead of directly predicting the pixel-wise color value like NeRF, MaskNeRF aims to learn better 3D correspondence among the source views and the target views and then employ the 3D correspondence to predict target pixel information during the inference stage. On the other hand, our method not only provides better PSNR, SSIM, and LPIPS but also renders images with better lighting conditions without additional computational cost.

Table 2: Performance (PSNR, SSIM and LPIPS) comparison among different λ values used in Hard-Mask on the DTU dataset with 3 training views as input.

λ	Real Data (DTU)		
	PSNR↑	SSIM↑	LPIPS↓
0.20	19.18	0.76	0.39
0.10	19.44	0.80	0.36
0.05	18.71	0.76	0.40
0.00	18.29	0.75	0.41

Table 3: Performance (PSNR, SSIM and LPIPS) comparison among different Top-K values used in Degenerated Hard-Mask on the DTU dataset with 3 training views as input.

K	Real Data (DTU)		
	PSNR↑	SSIM↑	LPIPS↓
80%	18.20	0.75	0.39
50%	18.90	0.77	0.38
30%	18.04	0.74	0.41
10%	17.58	0.72	0.43

4.2 Ablation Study

Ablation of λ in H-Mask. As shown in Tab. 2, we investigate the performance of different λ values (0, 0.05, 0.1, 0.2) in Hard-Mask on the DTU data set with 3 training views as input. Our experimental results suggest that $\lambda = 0.1$ leads to the best performance. From the table, we could learn that either enlarging or decreasing λ values would influence the performance of H-Mask. The main reason could be that enlarging λ equals to alleviate the emphasis on pixels learned by proposed Masks. Meanwhile, decreasing λ means ignoring more pixels not satisfying the 3D correspondence, while, these pixels may play an important role in the accuracy computation, e.g., the black background in DTU dataset as in Fig. 7.

Ablation of top-K Ratio in DH-Mask. In DH-Mask, top-K Ratio controls the portion of pixels masked by Degenerated Hard-Mask according to the trend of training loss. We compare the NeRF performance by selecting different portion (10%, 30%, 50%, 80%) of pixels and observe that masking top-50% pixels leads to the best performance, which is used in our main experiments. The reason for this phenomenon is similar to that of performance about different λ in Hard-Mask.

Additional Supervision on Unseen Views using Hard-Mask. With the the proposed Hard-Mask, we can locate pixels satisfying 3D correspondence relationship between unseen views and seen views. This can further allow us to reconstruct the color and depth information in unseen views with the help of the homography warping relationship between unseen views and seen views. One example of the reconstructed image is shown in Fig. 8(b). It could serve as an additional supervision for the optimization process and improve the performance, e.g., 1.06 absolute PSNR improvement on the DTU data set. More details are provided in Appendix F.

(a) GT (b) Reconstructed view

Figure 8: Mask-based unseen view reconstruction compared with ground truth.

5 Limitation

Similar as most NeRF based methods, our proposed mask based optimization can not render images with high quality when the target view is far from source views as 3D correspondence relationship is hard to utilize in this case. In addition, relighting in novel view synthesis is a challenging problem that needs further investigation.

6 Conclusion

In this paper, we address the challenging sparse view synthesis problem and propose MaskNeRF, an efficient plug-in component for NeRF-based methods in the sparse view setting. With correspondence built among pixels based on depths, we propose that Hard-Mask locates the pixels with 3D consistency, rather than treating all pixels equally in the training objective. Moreover, we observe that the pixels with 3D consistency are usually accompanied with smaller loss decreasing rate. Therefore, we propose to extract DH-Mask based on the trend of training loss without using the depth information. Experiment results show that our proposed methods significantly improve the performance of representative NeRF methods with sparse view settings and could bring larger performance improvement than previous depth-based methods. These promising results suggest that mask-based NeRF is an important direction to render images with both correct geometry and fine-grained details. A potential future direction is how to sample pixels with different losses during the optimization of NeRF.

ETHICS STATEMENT

We do not aware of any potential ethical concerns regarding our work.

REPRODUCIBILITY STATEMENT

We provide a copy of our code in the supplementary material to ensure reproducibility on all search spaces. Our experimental setting is stated in Section 4, and more details are described in the Appendix C and D.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864, 2021.

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *arXiv preprint arXiv:2103.15595*, 2021.

Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): 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.

François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6260–6269, 2022.

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3828–3838, 2019.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin 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.

Ajay Jain, Matthew Tancik, and Pieter 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.

Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12949–12958, 2021.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 406–413, 2014.

Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18365–18375, 2022.

James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984.

Mijeong Kim, Seonguk Seo, and Bohyung 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.

Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee 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.

Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. *arXiv preprint arXiv:2107.13421*, 2021.

Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping 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.

Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995.

Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pp. 405–421. Springer, 2020.

Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. *arXiv preprint arXiv:2112.00724*, 2021.

Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021.

Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14104–14113, 2021.

Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. *arXiv preprint arXiv:2112.03288*, 2021.

Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15872–15882, 2022.

Alex Trevithick and Bo 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.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas 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.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5610–5619, 2021.

Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich 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.

Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4877–4886, 2020.

Lin Yen-Chen. Nerf-pytorch. `https://github.com/yenchenlin/nerf-pytorch/`, 2020.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo 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.

Jian Zhang, Yuanqing Zhang, Huan Fu, Xiaowei Zhou, Bowen Cai, Jinchi Huang, Rongfei Jia, Binqiang Zhao, and Xing Tang. Ray priors through reprojection: Improving neural radiance fields for novel view extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18376–18386, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Kaichen Zhou, Lanqing Hong, Changhao Chen, Hang Xu, Chaoqiang Ye, Qingyong Hu, and Zhenguo Li. Devnet: Self-supervised monocular depth learning via density volume construction. *arXiv preprint arXiv:2209.06351*, 2022.

APPENDIX

A PROOF FOR PROPOSITION 1

Definition A.1 (Consistency of Estimated Appearance) *The consistency of estimated appearance refers to the predicted color difference between pixel $(i,j) \in \mathcal{M}$ and pixel $(m,n) := \mathcal{T}((i,j))$ (in the left/right view of Fig. 3) should be smaller than a threshold value ϵ_c, i.e.:*

$$||\hat{C}_\theta(\mathbf{r}_{ij}) - \hat{C}_\theta(\mathbf{r}_{mn})||_2^2 \leq \epsilon_c, \tag{10}$$

where $\hat{C}_\theta(\mathbf{r}_{ij})$ and $\hat{C}_\theta(\mathbf{r}_{mn})$ are predicted color of pixel (i,j) and (m,n).

Proposition 2 (Appearance Consistency Regularization) *Directly minimizing appearance consistency in Definition A.1 leads to trivial solution $\hat{C}_\theta(\mathbf{r}_{ij}) = \hat{C}_\theta(\mathbf{r}_{mn}) = 0$. Focusing on minimizing the errors between predict color values and their ground truth for pixels included by the Hard-Mask as in Eqn. (7) would help to emphasize the appearance consistency:*

$$||\hat{C}_\theta(\mathbf{r}_{ij}) - C(\mathbf{r}_{ij})||_2^2 + ||\hat{C}_\theta(\mathbf{r}_{mn}) - C(\mathbf{r}_{mn})||_2^2 \geq \frac{1}{4}||\hat{C}_\theta(\mathbf{r}_{ij}) - \hat{C}_\theta(\mathbf{r}_{mn})||_2^2 - \epsilon_c/2.$$

Proof:

$$\begin{aligned}
&2||\hat{C}_\theta(\mathbf{r}_{ij}) - C(\mathbf{r}_{ij})||_2^2 + 2||\hat{C}_\theta(\mathbf{r}_{mn}) - C(\mathbf{r}_{mn})||_2^2 \\
\geq &\|(\hat{C}_\theta(\mathbf{r}_{ij}) - \hat{C}_\theta(\mathbf{r}_{mn})) + (C(\mathbf{r}_{mn}) - C(\mathbf{r}_{ij}))\|_2^2 \\
\geq &\frac{1}{2}\|\hat{C}_\theta(\mathbf{r}_{ij}) - \hat{C}_\theta(\mathbf{r}_{mn})\|_2^2 - \|C(\mathbf{r}_{mn}) - C(\mathbf{r}_{ij})\|_2^2
\end{aligned}$$

The first inequality follows from the fact that for two vectors $\mathbf{a}, \mathbf{b}$,

$$2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2 - \|\mathbf{a}+\mathbf{b}\|_2^2 = \|\mathbf{a}-\mathbf{b}\|_2^2 \geq 0.$$

The second inequality is due to the fact that

$$2\|\mathbf{a}+\mathbf{b}\|_2^2 - (\|\mathbf{a}\|_2^2 - 2\|\mathbf{b}\|_2^2) = \|\mathbf{a}+2\mathbf{b}\|_2^2 \geq 0.$$

B GEOMETRY CONSISTENCY

Definition B.1 (Geometry Consistency) *The geometry consistency refers to the depth difference between the depth of pixel $(m,n) \in \mathcal{M}$ in right camera view and the depth generated by warping its corresponding pixel $(i,j) := \mathcal{T}((m,n))$ from left camera to right camera should be smaller than a threshold value ϵ_s, i.e.:*

$$||s^r_{mn} - s^{'}_{mn}||_2^2 \leq \epsilon_s, \tag{11}$$

where s^r_{mn} is the depth for pixel (m,n) and $s^{'}_{mn}$ is the projected depth from left camera pixel (i,j).

Definition B.2 (Consistency of Estimated Geometry) *The consistency of estimated geometry refers to the predicted depth difference between the depth of pixel $(m,n) \in \mathcal{M}$ in right camera view and the predicted depth generated by warping its corresponding pixel $(i,j) := \mathcal{T}((m,n))$ from left camera to right camera should be smaller than a threshold value ϵ_s, i.e.:*

$$||\hat{s}_\theta(\mathbf{r}_{mn}) - \hat{s}^{'}_\theta(\mathbf{r}_{mn})||_2^2 \leq \epsilon_s, \tag{12}$$

where $\hat{s}_\theta(\mathbf{r}_{mn})$ is the predicted depth for pixel (m,n) and $\hat{s}^{'}_\theta(\mathbf{r}_{mn})$ is the projected depth from left camera pixel (i,j).

Proposition 3 (Geometry Consistency Regularization) *Similar to Appearance Consistency Regularization, focusing on optimizing the error between predicted depth value and its ground truth for pixels included by Hard-Mask as in Eqn. (7) would help to emphasize the geometry consistency:*

$$||\hat{s}_\theta(\mathbf{r}_{mn}) - s^r_{mn}||_2^2 + ||\hat{s}^{'}_\theta(\mathbf{r}_{mn}) - s^{'}_{mn}||_2^2 \geq \frac{1}{4}||\hat{s}_\theta(\mathbf{r}_{mn}) - \hat{s}^{'}_\theta(\mathbf{r}_{mn})||_2^2 - \epsilon_s/2, \tag{13}$$

Proof:

$$
\begin{aligned}
&2||\hat{s}_\theta(\mathbf{r}_{mn}) - s^r_{mn}||_2^2 + 2||\hat{s}^{'}_\theta(\mathbf{r}_{mn}) - s^{'}_{mn}||_2^2 \\
\geq &\|(\hat{s}_\theta(\mathbf{r}_{mn}) - \hat{s}^{'}_\theta(\mathbf{r}_{mn})) + (s^r_{mn} - s^{'}_{mn})\|_2^2 \\
\geq &\frac{1}{2}\|\hat{s}_\theta(\mathbf{r}_{mn}) - \hat{s}^{'}_\theta(\mathbf{r}_{mn})\|_2^2 - \|s^r_{mn} - s^{'}_{mn}\|_2^2
\end{aligned}
$$

The first inequality follows from the fact that for two vectors $\mathbf{a}, \mathbf{b}$,

$$2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2 - \|\mathbf{a}+\mathbf{b}\|_2^2 = \|\mathbf{a}-\mathbf{b}\|_2^2 \geq 0.$$

The second inequality is due to the fact that

$$2\|\mathbf{a}+\mathbf{b}\|_2^2 - (\|\mathbf{a}\|_2^2 - 2\|\mathbf{b}\|_2^2) = \|\mathbf{a}+2\mathbf{b}\|_2^2 \geq 0.$$

C Preliminary Study

By utilizing homography warping relationship, we locate pixels satisfying 3D correspondence relationship. Based on the masked pixels among training views, we find the respective 3D points and randomly sample different portions (30%, 60%, 100%) of 3D points for the purpose of emphasizing the 3D correspondence. We conduct each experiment using 4 random seeds and report the mean results.

D Implementation Details

All our models are trained on the NVIDIA Tesla V100 Volta GPU cards. The NeRF based models are implemented based on the code from (Yen-Chen, 2020). For IBRNet, we pre-train the model on the DTU data for 120,000 iterations and then conduct the per-scene optimization for 20,000 iterations. For MVSNet, we follow the released code and checkpoint to pre-train and finetune the models. For Hard-Mask introduced in Sec. 3.3, we generate the mask information for each training image based on the correspondence among pixels in all training views. For Degenerated Hard-Mask introduced in Sec. 3.3, we generate the mask information for each training image based on the trend of training loss.

E Solutions to Avoid Degenerate Results in NeRF

As mentioned in Sec. 3.3, NeRF is prone to a catastrophic failure at the initialization stage in which MLP emits negative values before the ReLU activation. To address this issue, Mip-NeRF (Barron et al., 2021) proposed to use a softplus function to yield a stable optimization. However, we observe that NeRF overfits to training views by using the softplus function in the sparse view setting. One possible reason could be that the predicted alpha value of sampled points should be sparse and dropping small values with ReLU activation could effectively improve the generalization ability. Based on the above consideration, we instead propose to modify the initialization of bias parameters in the MLP to guarantee both stable optimization and good generalization ability. As shown in Tab. 4, our proposed initilization effectively improve the performance of NeRF and avoid the degenerate results when compared with SoftPlus activation and the original NeRF setting.

F Additional Supervision on Unseen Views using Hard-Mask

With the the proposed Hard-Mask, we can locate pixels satisfying 3D correspondence relationship between unseen views and seen views. This can further allow us reconstruct the color and depth information in unseen views with the help of the homography warping relationship between unseen views and seen views. One example of the reconstructed image is shown in Fig. 8(b). Although the reconstructed image labels (Fig. 8 (b)) miss some information in the unseen area of training views, the overall geometry and sharp details are clear when compared with the ground truth label (Fig. 8 (a)). As shown in Tab. 5, it could serve as an additional supervision for the optimization process and improve the performance, e.g., 1.06 absolute PSNR improvement on the DTU data set.

Table 4: Performance (PSNR, SSIM and LPIPS) comparison between SoftPlus and our proposed stable initialization to avoid degenerate results in NeRF on the DTU data set with 3 training views as input. ↑ means the larger is better; ↓ means the smaller is better.

Method	Real Data (DTU)		
	PSNR↑	SSIM↑	LPIPS↓
ReLU	11.40	0.50	0.49
SoftPlus	14.26	0.68	0.45
Stable Initialization	16.91	0.73	0.41

Table 5: Performance (PSNR, SSIM and LPIPS) comparison of applying Mask-Based data augmentation on the DTU data set with 3 training views as input. For performance (PSNR, SSIM and LPIPS) comparison, ↑ means the larger is better; ↓ means the smaller is better.

Additional Supervision	Real Data (DTU)		
	PSNR↑	SSIM↑	LPIPS↓
✗	19.44	**0.80**	**0.36**
✓	**20.50**	**0.80**	0.38

G Comparison with Segmentation Mask

For scenes with simple background like NeRF synthetic data set, segmentation method can be used to segment the object and background. We compare the performance between the Mask R-CNN He et al. (2017) segmentation method and our proposed mask methods on the NeRF synthetic data set. As shown in Tab. 6, our proposed DH-Mask and H-Mask outperform the segmentation method by emphasizing the learning of 3D consistency. For DTU and LLFF data sets with complex scenes, segmentation method cannot be used to effectively select pixels satisfying the 3D correspondence relationship.

Table 6: Performance (PSNR, SSIM and LPIPS) comparison between our proposed masks and segmentation mask on the NeRF Synthetic data set with 3 training views as input.

Method	Real Data (DTU)		
	PSNR↑	SSIM↑	LPIPS↓
Segmentation	19.92	0.85	0.33
DH-Mask	20.71	0.86	0.31
H-Mask	20.94	0.87	0.29