

PANeRF: Pseudo-view Augmentation for Improved Neural Radiance Fields Based on Few-shot Inputs

Young Chun Ahn

Seokhwan Jang

Sungheon Park

Ji-Yeon Kim

Nahyup Kang

Samsung Advanced Institute of Technology (SAIT)

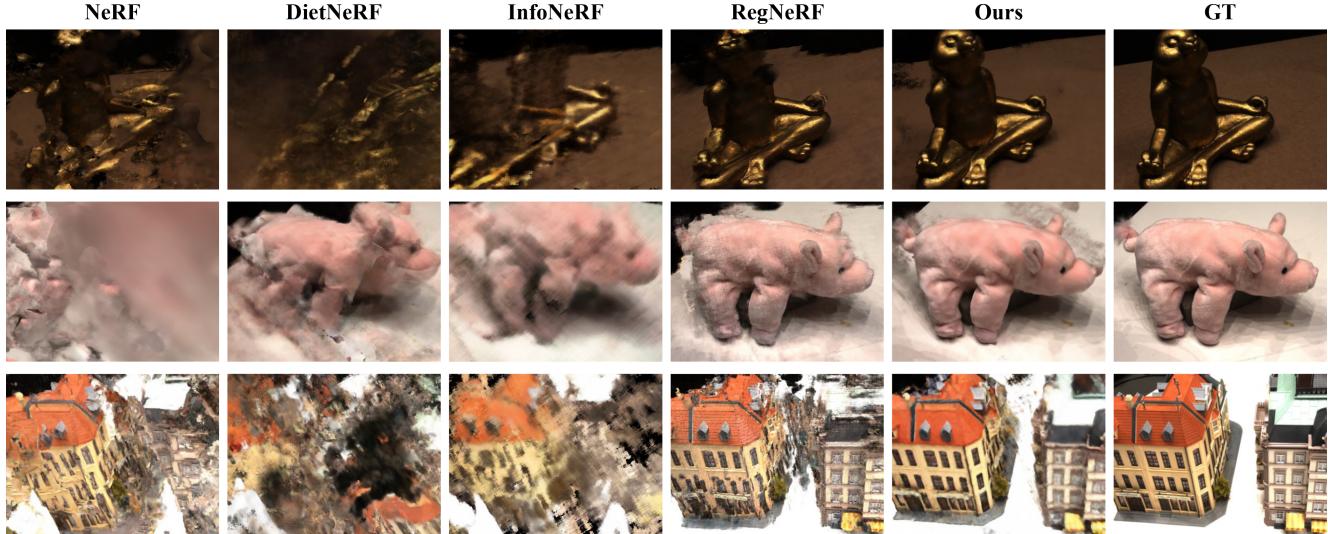


Figure 1. View synthesis based on the DTU with a 3-view setting. We demonstrate a qualitative comparison of pseudo-view augmentation neural radiance fields (PANeRF) with other few-shot methods of *Scan21*, *Scan103*, and *Scan110* scenes based on the DTU dataset. Although other methods experience inaccurate geometry and appearance, our approach yields high-quality rendering results with minimal artifacts.

Abstract

The method of neural radiance fields (NeRF) has been developed in recent years, and this technology has promising applications for synthesizing novel views of complex scenes. However, NeRF requires dense input views, typically numbering in the hundreds, for generating high-quality images. With a decrease in the number of input views, the rendering quality of NeRF for unseen viewpoints tends to degenerate drastically. To overcome this challenge, we propose pseudo-view augmentation of NeRF, a scheme that expands a sufficient amount of data by considering the geometry of few-shot inputs. We first initialized the NeRF network by leveraging the expanded pseudo-views, which efficiently minimizes uncertainty when rendering unseen views. Subsequently, we fine-tuned the network

by utilizing sparse-view inputs containing precise geometry and color information. Through experiments under various settings, we verified that our model faithfully synthesizes novel-view images of superior quality and outperforms existing methods for multi-view datasets.

1. Introduction

Although neural radiance fields (NeRF) [15] have remarkably advanced the field of neural rendering, synthesis of photo-realistic views with sparse inputs via this technology has remained a major challenge. Moreover, in real-world applications, such as AR/VR and autonomous driving, data sparsity is a major technological hurdle. When considering the use of NeRF in practical applications,

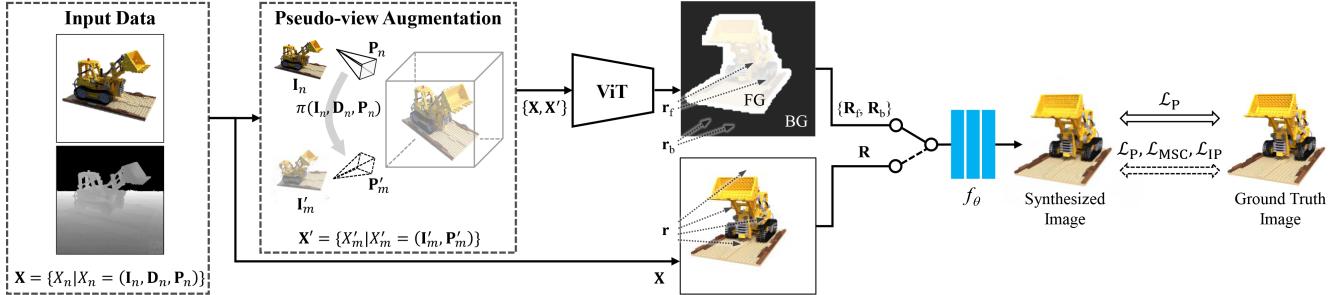


Figure 2. Overall pipeline for training PANeRF. In pseudo-view augmentation, π denotes warping operation. \mathbf{R}_f and \mathbf{R}_b denote a set of 3D points on rays, \mathbf{r}_f and \mathbf{r}_b , in the foreground and background, respectively. Note that we train our model through a two-step learning process with different losses.

achieving a reliable performance despite this limitation is essential.

To overcome the aforementioned limitation, various methods [6, 16, 27, 34] employ a pre-trained model. In particular, these methods utilize the input-image features to construct a feature pyramid, add a ray transformer for density estimation, extract semantic information, or maximize the predicted log-likelihood. These methods yield an adequate rendering performance. However, their results feature inaccurate geometries and floating artifacts.

Another approach [9] proposes regularization techniques with the use of a pre-trained model. Although this strategy enhances the performance in multi-view rendering based on adjacent sparse views such as narrow-baseline imagery, the details tend to be lost for unseen viewpoints.

In addition, certain recent works [4, 31] have leveraged depth maps. In one of these studies [4], depth information was utilized for direct supervision at the rendering view. In another study [31], training images were augmented through homography warping.

In accordance with this line of research, we employed a view augmentation scheme to resolve the aforementioned challenge resulting from sparse data. Similar to our approach, SinNeRF [31] performs data augmentation using an input image and its depth map; however, the synthesized data are utilized solely via a patch-wise manner.

Specifically, although certain inaccurate regions existed owing to warping artifacts, we fully utilized the augmented images for training. To minimize the effects of warping inaccuracy, we propose computation of saliency maps through DINO-ViT [3]. Utilizing the saliency map, we divided training images into foreground and background regions and calculated the loss for each region separately. As a result of the advantageous features of the saliency map, all the pseudo views could be considered reliable for calculating the training loss.

Additionally, we propose a two-step learning process. First, the network was initially trained based on all the augmented images to attenuate the uncertainty corresponding

to the novel views. Subsequently, the network was fine-tuned through novel regularization methods solely based on the ground-truth images. Owing to this strategy, our method yields a superior image quality despite a 1-view setting, which is considered an extreme case.

In this paper, we propose *PANeRF* for synthesizing novel views based on few-shot data. The main contributions are summarized as follows:

- Our pseudo-view augmentation (PA) scheme aggregates geometric accuracy and semantic details in rendering at novel viewpoints.
- We propose novel regularization methods, namely the multi-level semantic consistency (MSC) and information potential (IP). The former method maintains semantic consistency at local and global levels, and the latter minimizes uncertainty along the rays.
- We verified the robustness of our method through numerous experiments under various conditions and achieved high-quality view synthesis superior to that rendered by existing few-shot methods.

2. Related Work

Novel View Synthesis. For novel-view synthesis, certain explicit representations of 3D vision using voxels have been reported [12, 22], meshes [21, 24], point clouds [29], or multiplane images [5, 14, 23, 25]. Although these methods allow for rendering of images based on the reconstructed 3D scenes, optimizing these schemes is generally a challenging task owing to their properties of discontinuity. Moreover, various studies [11, 17, 32] report implicit representation in which geometry and appearance can be directly learned without an explicit 3D model. In this context, neural radiance fields (NeRF) [15] is a representative work that performs photo-realistic novel-view synthesis. In addition, various schemes have been introduced to improve the performance of NeRF. For example, Mip-NeRF [1] and its extension, namely, Mip-NeRF 360 [2], reduce aliasing by trac-

ing a cone instead of a ray. Furthermore, Ref-NeRF [26] achieves improved rendering quality by replacing the parameterization of view-dependent radiance with the representation of reflected radiance.

Few-shot View Synthesis. Various methods have been suggested in the field of study to overcome limitations resulting from insufficient input data. In particular, certain approaches employ well-defined regularization techniques with or without prior models. In this perspective, Pixel-NeRF [34] utilizes a feature volume based on a convolutional neural network (CNN) encoder to train NeRF. IBR-Net [27] extracts density features by accumulating image features, colors, and viewing directions of adjacent views, and transmits these features through a ray transformer to estimate a more accurate density. DietNeRF [6] employs a vision transformer to extract semantic features for sustaining global semantic information at novel viewpoints. This scheme leverages CLIP-ViT [18], which has been trained on enormous 2D image and text data, to compensate for information scarcity. Furthermore, RegNeRF [16] regularizes the geometry and color of patches rendered from novel views through a pre-trained normalizing-flow model. Its geometry regularization is rendered as the patch-wise depth smoothness from unseen views. In addition, color regularization maximizes the log-likelihood of the predicted color of image patches while training by utilizing the normalized flow model.

InfoNeRF [9] introduces a prior-free model that utilizes the ray entropy among seen and unseen poses. This scheme proposes a regularization technique through ray entropy minimization. The ray entropy is computed based on the density distribution by using the volume densities of the rays. Although this strategy improves the image quality of novel-view synthesis by minimizing the entropy among rays sampled in the seen and unseen views, artifacts such as blurring and cloud effects are usually rendered into the synthesized image.

Recently, a few studies have introduced the use of depth supervision for novel-view synthesis. In this regard, DSNeRF [4] applies geometric constraint as a form of direct depth supervision; it applies the supervision using sparse 3D points obtained from structure-from-motion (SFM) such as COLMAP [19]. Moreover, depth supervision can be performed by projecting the 3D points with the corresponding camera parameters. The most recently proposed strategy in this regard, SinNeRF [31], addresses more extreme conditions wherein neural radiance fields are trained solely using a single view. This strategy augments the geometry-based pseudo labels near the reference view via forward image warping based on depth supervision. However, on account of the incorrect warping, the geometry and texture-guidance loss are designed in a small patch.

3. Method

Figure 2 illustrates an overview of our proposed method. We propose a two-step learning process comprising network initialization and fine-tuning. For the network initialization, we introduce PA, a technique for expanding novel views from given input data to their surroundings via forward image warping based on depth information. In addition, we extracted their saliency maps to minimize artifacts resulting from warping inaccuracy during the training. Following the initialization, the network was fine-tuned using only the few-shot input images. For this purpose, we employed novel regularization methods. MSC was applied to maintain semantic attributes, and IP was adopted to minimize uncertainty in novel viewpoints. The relevant details are described in Sec. 3.2, Sec. 3.3, and Sec. 3.4, respectively.

3.1. Preliminary

NeRF [15] represents a 3D scene with a neural implicit function using a multi-layer perceptron (MLP). In this approach, a 3D position, $\mathbf{x} = (x, y, z)$, and a viewing direction, $\mathbf{d} = (\theta, \phi)$, are fed as inputs to the MLP network, and the volume density σ and color $\mathbf{c} = (r, g, b)$ at the 3D point are predicted accordingly. By integrating colors and densities along the ray based on the volume rendering [13], the RGB color at the target pixel is rendered. In practice, a subset of points on a ray is sampled and fed into the network to produce the rendered RGB color, as expressed in (1):

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

where \mathbf{r} denotes a ray, N indicates the total number of sample points, and δ_i represents the distance between the i^{th} and $(i+1)^{\text{th}}$ samples. $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$ symbolizes the transmittance accumulated along ray until the i^{th} sample point. The NeRF model is optimized by minimizing the photometric loss between the ground truth and synthesized images, which can be expressed as follows:

$$\mathcal{L}_P = \sum_{\mathbf{r} \in \mathcal{R}} \left\| \hat{\mathbf{C}}(\mathbf{r}) - C(\mathbf{r}) \right\|^2, \quad (2)$$

where \mathcal{R} denotes a set of training rays and C is a pixel color of a training image.

3.2. Pseudo-view Augmentation

To address the scarcity of viewpoints, pseudo-view images are generated through homography warping, as depicted in Fig. 2. Referring to the poses of few-shot inputs, we initially rotated the camera pose in the x , y , and z axes in the range of $\pm\alpha^\circ$, respectively. Subsequently, warping is performed ranging from the reference camera pose \mathbf{P} to the

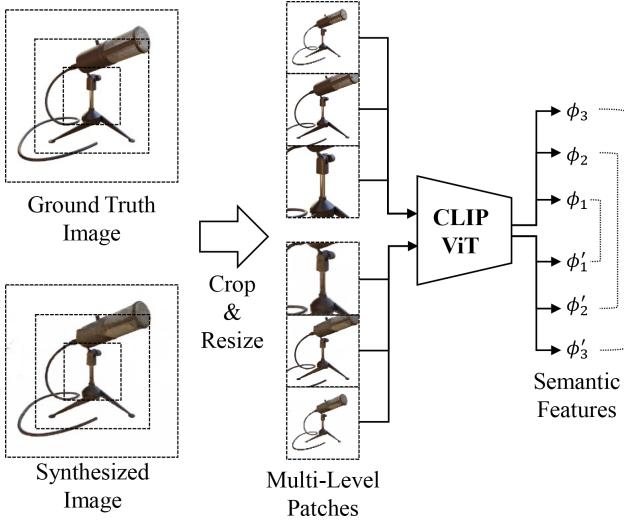


Figure 3. Illustration of measuring the MSC between the ground-truth and synthesized images in the case of $L = 3$. Unlike the approach assumed in another study [6], we investigated the consistency at local levels as well as global levels.

transformed pose \mathbf{P}' . Accordingly, the transformed pixel position \mathbf{p}' can be obtained through (3) as follows:

$$\mathbf{p}' = \mathbf{K} \mathbf{T}_{\mathbf{P} \rightarrow \mathbf{P}'} D(\mathbf{p}) \mathbf{K}^{-1} \mathbf{p}, \quad (3)$$

where \mathbf{K} , $\mathbf{T}_{\mathbf{P} \rightarrow \mathbf{P}'}$, and D denote a camera intrinsic matrix, transformation matrix from \mathbf{P} to \mathbf{P}' , and depth map, respectively. Notably, the forward-warping scheme with scattering operation is adopted in our implementation. Specifically, a source pixel value is distributed to the target position's neighboring pixels via the scattering operation.

Moreover, the quality of the image obtained using the forward-warping scheme primarily depends on the accuracy of the depth map. To minimize this dependency, in this study, we adopted the reliable depth map estimated from the pre-trained NeRF [15]. To further attenuate the warping effect, we generated saliency maps by employing DINO-ViT [3], and these maps were leveraged to separately supervise pixel values of the foreground and background.

3.3. Multi-level Semantic Consistency

We explored semantic features from the pre-trained CLIP-ViT [18] to ensure consistent semantic information for synthesizing novel views. In particular, DietNeRF [6] extracts features using CLIP-ViT [18]; however, the features are employed at a global level. In contrast, we propose the MSC that can maintain semantic attributes at both local and global levels.

To extract the embedding features at each level, we obtained patches of three different sizes by cropping an image around its center as illustrated in Fig. 3. The patch-wise

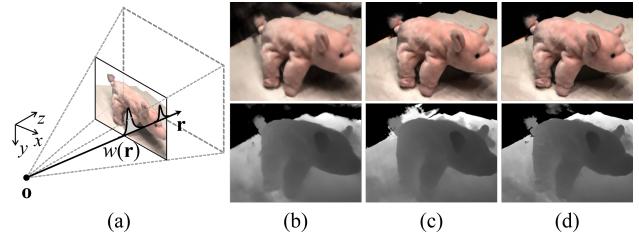


Figure 4. (a) Illustration of ray \mathbf{r} , camera position \mathbf{o} , and its weight distribution $w(\mathbf{r})$. The synthesized images and their depth maps obtained (b) after the initialization, (c) without IP, (d) with IP, respectively.

images were scaled and subsequently fed into the CLIP-ViT [18]. Thereafter, the cosine similarities of the embedding features obtained from the same region were calculated, and summed up as the MSC loss. This loss is expressed as

$$\mathcal{L}_{\text{MSC}} = \sum_{l=1}^L S(\phi_l(\hat{\mathbf{I}}), \phi_l(\mathbf{I})), \quad (4)$$

where $\hat{\mathbf{I}}$ and \mathbf{I} denote the synthesized and the ground-truth images, respectively. Notably, L indicates the number of levels, and ϕ_l denotes embedding features at level l . $S(\cdot)$ symbolizes a function of cosine similarity between two embedding features.

3.4. Information Potential

In addition to the MSC, we introduce a new regularization method that suppresses the uncertainty in transmittance along the rays, which results from scarce data. We observed in an experiment that a weight distribution on a ray, which is represented by multiplying the transmittance and opacity at sample point as in (1), tends to be concentrated on the surfaces of a scene, as illustrated in Fig. 4(a).

Considering that the Shannon entropy [20] decreases as the probability density function sharpens, the entropy of the weight distribution will be minimized when the NeRF model is optimized on the scene. Based on this observation, we adopted IP which is derived from the Rényi quadratic entropy [30]. It is a convex function, which is more suitable for gradient-based optimization than the Shannon entropy, as proven for 3D-image reconstruction [7], [10]. The IP with respect to the weights can be expressed by discarding the negative logarithm from the quadratic entropy as follows:

$$\mathcal{L}_{\text{IP}} = -\frac{1}{|\mathcal{R}|} \sum_{\mathbf{r} \in \mathcal{R}} \sum_{i=1}^N \tilde{w}_i(\mathbf{r})^2, \quad (5)$$

where

Algorithm 1 Training PANeRF

```

1: Input: Input data  $\mathbf{X} = \{X_n | X_n = (\mathbf{I}_n, \mathbf{D}_n, \mathbf{P}_n)\}$ ,  

   loss weights  $\lambda_1, \lambda_2$ , learning rate  $\eta$   

2: Output: Radiance field  $f_\theta$  with the optimum  $\theta$   

3: Generate  $\mathbf{X}' = \{X'_m | X'_m = (\mathbf{I}'_m, \mathbf{P}'_m)\}$  via  

   Pseudo-view Augmentation  

4: Initialize  $\theta = \theta_0$   

5: for  $\xi \subset \{\mathbf{X}, \mathbf{X}'\}$  do Initialization  

6:   Sample rays  $\mathbf{r} \in \mathcal{R}$   

7:   Calculate  $\mathcal{L}_P$  using Eq. (2)  

8:    $\mathcal{L}_{total} \leftarrow \mathcal{L}_P$   

9:    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{total}$   

10: end for  

11: for  $\xi \subset \mathbf{X}$  do Fine-tuning  

12:   Sample rays  $\mathbf{r} \in \mathcal{R}$   

13:   Calculate losses using Eqs. (2), (4), and (5)  

14:    $\mathcal{L}_{total} \leftarrow \mathcal{L}_P + \lambda_1 \mathcal{L}_{MSC} + \lambda_2 \mathcal{L}_{IP}$   

15:    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{total}$   

16: end for

```

$$\tilde{w}_i(\mathbf{r}) = w_i(\mathbf{r}) / \sum_{j=1}^N w_j(\mathbf{r}) \quad (6)$$

and

$$w_i(\mathbf{r}) = T_i(1 - \exp(-\sigma_i \delta_i)). \quad (7)$$

Herein, \mathcal{R} denotes a set of rays. Notably, in (6) and (7), $w_i(\mathbf{r})$ represents a weight at the i^{th} point that is sampled on ray \mathbf{r} . Moreover, IP observably increases when the weight distribution is concentrated on specific bins, and therefore, \mathcal{L}_{IP} is multiplied with -1 to maximize its value during the training.

As the IP is designed to complement the aforementioned tendency, the cloudy artifacts are attenuated. Furthermore, the boundary of the object is rendered clearer with the application of our regularizer IP during training, as indicated by the depth maps presented in Fig. 4.

3.5. Overall Objective

The total loss for training our model is given by

$$\mathcal{L}_{total} = \mathcal{L}_P + \lambda_1 \mathcal{L}_{MSC} + \lambda_2 \mathcal{L}_{IP}, \quad (8)$$

where λ_1 and λ_2 denote parameters for balancing the data term, acting as two regularizers. Algorithm 1 summarizes the entire training process.

4. Experiments

4.1. Datasets and Evaluations

Datasets. We utilized the Realistic Synthetic 360° [15] dataset, which is one of the most commonly used bench-

marks for novel-view synthesis. This dataset includes 8 synthetic scenes generated by natural non-Lambertian materials with complex geometries. Each scene contains 400 images, which are captured using various camera poses, covering from the upper hemisphere to the entire sphere. In addition, the DTU dataset [8] was employed for real-world benchmark. This dataset comprises multi-view stereo images, including those of various real-world objects, captured under calibrated camera environments. Each scene contains 49 images captured through different camera poses.

Evaluation Protocols. To evaluate our results on the Realistic Synthetic 360°, we followed the evaluation protocol of InfoNeRF [9]. Our model was trained with 4 out of 100 training images, and their depth maps were used by training [33] for each scene. Subsequently, we evaluate the model with 200 testing images. In particular, 5 different sets of randomly selected 4 viewpoints were utilized to train our model, and the average performance was reported as the final result. Moreover, we evaluated our results on the DTU dataset. For a fair comparison with prior works [16, 31], we adopted their protocols. Firstly, to compare with RegNeRF [16], we performed experiments on 15 scenes. Among 49 images of each scene, we trained our model with 3 and 6 images and evaluated the synthesized results using 25 images along with object masks. Depth maps for the PA were obtained by training [33]. Secondly, we compared our method with SinNeRF [31]. Experiments were performed on 19 scenes. Considering that [31] aims to train using a single view, we selected camera ID 2 as the reference view for training and tested our model with 10 images captured at views in close proximity of the reference view. For this comparison, depth maps provided by [31] were leveraged.

Evaluation Metrics. To evaluate our approach, the average peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [28], and learned perceptual image patch similarity (LPIPS) [35] were measured. In addition, we report the geometric mean of MSE, $\sqrt{1 - SSIM}$, and LPIPS, as conducted in a study [16].

4.2. Implementation Details

We implemented our method using the PyTorch version of the NeRF [33]. Our model was trained for 10k iterations for the network initialization and 40k iterations for fine-tuning. For the PA, the range of the rotation angles α was set to 30°, along with an interval of 5° between views. Thus, the number of images generated by referring to a source view was determined by $2,196 = \{(30 \times 2)/5 + 1\}^3 - 1$. We implemented the forward-warping scheme with the linear-scattering operation based on CUDA. In all of the experiments, the proposed model was trained and evaluated on a single NVIDIA A100 GPU.

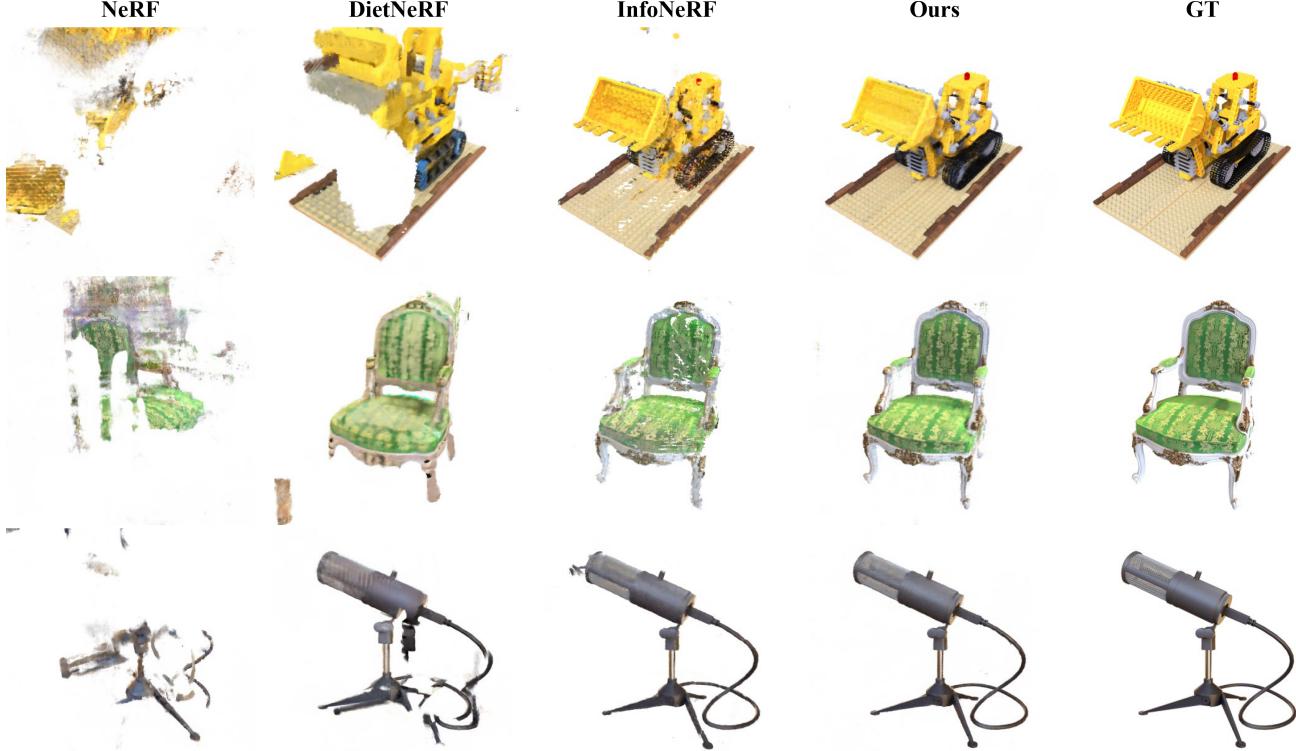


Figure 5. View synthesis on the Realistic Synthetic 360° with a 4-view setting. As revealed in comparison of our model to other few-shot NeRF methods, PANeRF yields a more realistic rendering. Our approach exhibits accurate scene geometry in *Lego* and *Chair*, generating superior details, as observed in *Mic*.

Table 1. Quantitative results on the Realistic Synthetic 360° with 4-view reconstruction. Notably we referred to Tab. 1 reported in InfoNeRF [9] for comparison with other models. The optimal results are marked in bold, and the secondary results are underlined.

Model	PSNR(\uparrow)	SSIM(\uparrow)	LPIPS(\downarrow)
NeRF [15]	15.93	0.780	0.320
PixelNeRF [34]	16.09	0.738	0.390
DietNeRF [6]	16.06	0.793	0.306
InfoNeRF [9]	<u>18.65</u>	<u>0.811</u>	<u>0.230</u>
Ours	22.13	0.839	0.156

5. Results

5.1. Realistic Synthetic 360

On the Realistic Synthetic 360° dataset, we compare our method with InfoNeRF [9], whose NeRF model is trained using 4 input views. Figure 5 depicts the rendering images on the 3 different scenes, respectively. The experiments of the prior works, NeRF [15], DietNeRF [6], and InfoNeRF [9], are reproduced by following the evaluation protocol reported in the literature [9]. NeRF tends to degenerate when trained with sparse views. PixelNeRF and DietNeRF

achieve relatively superior results via prior models, but the improvement is insignificant. InfoNeRF incorporates prior-free model and yield better results. Nevertheless, it does not yield competitive results. Our method exhibits accurate geometry and appearance representation in the synthesized results, whereas certain artifacts such as the cloud effect are clearly observed in the images rendered by the previous works. Furthermore, the quantitative evaluation indicates a significant improvement on the performance compared to that of the existing methods, as indicated in Tab. 1.

5.2. DTU

In addition to the synthetic data, we compared our method with prior works [16, 31] on the DTU dataset. Figure 1 illustrates the qualitative results on the 3 different scenes with the model trained through a 3-view setting. For the qualitative comparison, we reproduced the results of NeRF [15], DietNeRF [6], and InfoNeRF [9] by following the evaluation protocol reported in the study [16]. Notably, we referred to RegNeRF [16] for their qualitative results. Our method is superior to the other works in all aspects such as color, detail, and geometry, as shown in Fig. 1. In particular, as expressed in detail in the figure, the face of the statue and the tail of the pig doll are clearly expressed,

Table 2. Quantitative results on the DTU dataset with 3, 6, and 9 views. Notably we referred to Tab. 1 reported in [16] for the previous methods except for those of InfoNeRF [9] whose results are reproduced.

Model	PSNR(\uparrow)			SSIM(\uparrow)			LPIPS(\downarrow)			Average(\downarrow)		
	3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view
Mip-NeRF [1]	8.68	16.54	23.58	0.571	0.741	0.879	0.353	0.198	0.092	0.323	0.148	0.056
DietNeRF [6]	11.85	20.63	23.83	0.633	0.778	0.823	0.314	0.201	0.173	0.243	0.101	0.068
InfoNeRF [9]	14.13	19.84	21.78	0.626	0.713	0.741	0.314	0.265	0.246	0.213	0.128	0.108
RegNeRF [16]	18.89	22.20	24.93	0.745	0.841	0.884	0.190	0.117	0.089	0.112	0.071	0.047
Ours	24.16	26.60	-	0.829	0.856	-	0.163	<u>0.149</u>	-	0.072	0.056	-

Table 3. Quantitative results on the DTU dataset with a 1-view setting. Notably we referred to Tab. 2 reported in SinNeRF [31] for comparison with other models.

Model	PSNR(\uparrow)	SSIM(\uparrow)	LPIPS(\downarrow)
DSNeRF [4]	12.17	0.41	0.6493
DietNeRF [6]	12.84	0.44	0.6469
PixelNeRF [34]	12.06	0.42	0.6471
SinNeRF [31]	<u>16.52</u>	<u>0.56</u>	<u>0.5250</u>
Ours	16.61	0.59	0.4001

respectively. In addition, the structure of the buildings is well restored and its color is correctly generated. This tendency is readily observed in the quantitative analysis as well, as demonstrated in Tab. 2. Our method outperforms the other methods in terms of all of the metrics, considering the 3-view results. Moreover, a significant difference exists among the evaluation metrics except for LPIPS at 6 input views. Notably, the performance of the proposed method on PSNR exceeded those of the other methods using 9 input views despite the fact that our model was trained using 6 views. Both the qualitative and quantitative results validate that our method based on the view extension functions optimally.

Furthermore, although our model was trained using a single image, it exhibited comparable rendering quality. Figure 6 demonstrates the corresponding results. Considering the single view (right) for training, we reveal the view-synthesis results (left) from an unseen viewpoint (middle). Our method yields the results corresponding to the correct geometry and fine details, achieving superior quantitative results. Especially significant difference in results is observed in terms of the LPIPS, as listed in Tab. 3.

5.3. Analysis

Effect of the Pseudo-view Augmentation. As discussed in Sec. 3.2, our method trains the NeRF by expanding the insufficient views through PA. In Fig. 7, the results suggest that PA can yield a reliable initial point in terms of optimization. By comparing the results presented in (c) with (a) and (b) in Fig. 7, we discover that the PA facilitates prediction of the correct geometry and semantic information. Moreover,

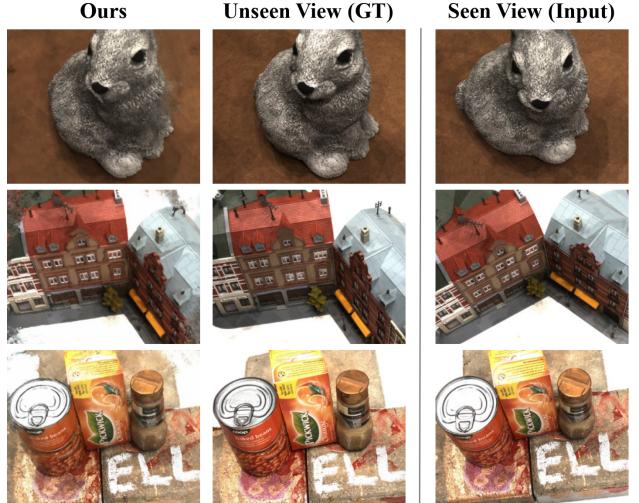


Figure 6. View synthesis on the DTU dataset with a single view.

Table 4. Ablation study for the proposed methods on the *Lego* scene with a 4-view setting.

PA	PANeRF			PSNR(\uparrow)	SSIM(\uparrow)	LPIPS(\downarrow)
	SC	MSC	IP			
			✓	20.07	0.781	0.190
			✓	20.25	0.792	0.178
✓				18.75	0.750	0.271
✓	✓			22.47	0.833	0.149
✓		✓		<u>22.58</u>	<u>0.836</u>	<u>0.146</u>
✓	✓	✓	✓	22.82	0.839	0.144

the proposed regularization methods along with PA can successfully minimize uncertainties; therefore, the high-quality of synthesized images can be synthesized. These results obtained via quantitative analysis validate the impact of the PA scheme, as indicated in Tab. 4.

Benefit of the Regularizers. The proposed regularizers of MSC and IP improved the performance and contributed to the synthesis of fine details. The results in Fig. 7 validate that the application of the MSC generates more details than the SC. In addition, IP on top of PA and MSC, supports the proposed method in achieving the best performance.

Robustness of Our Approach. We performed the ex-

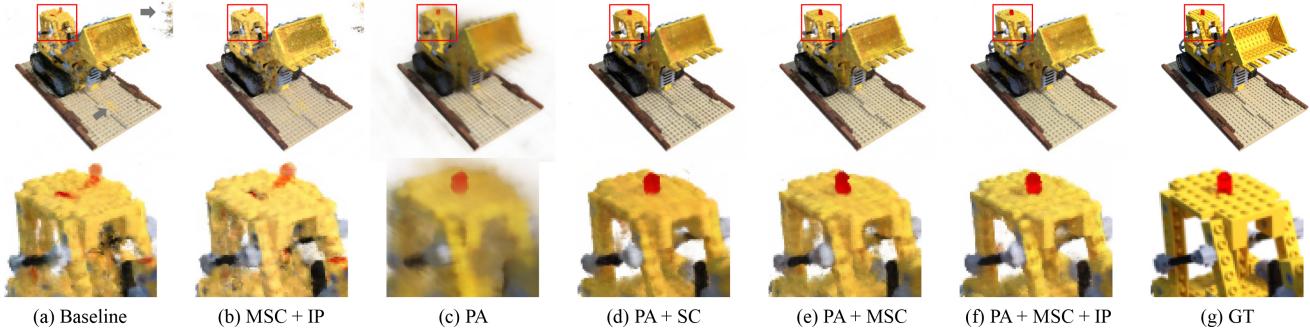


Figure 7. Results of our ablation study on the *Lego* scene of the Realistic Synthetic 360°. The second row presents the enlarged images of the regions marked as a red box in the first row. (a) NeRF [15] trained with a 4-view setting as our baseline. (b) Proposed regularization methods. (c) Psuedo-view augmentation only. (d) Add Semantic Consistency (SC) loss proposed in [6]. (e) MSC loss. (f) Our proposed method. (g) Ground truth.

Table 5. Quantitative results of the *Lego* scene on the Realistic Synthetic 360° with the narrow-baseline training images.

Model	PSNR(\uparrow)	SSIM(\uparrow)	LPIPS(\downarrow)
InfoNeRF [9]	18.41	0.775	0.199
Ours	21.50	0.813	0.156

periments in various environments. For the DTU dataset, our approach outperforms other methods even with 3 input views, as demonstrated in Tab. 2. Furthermore, even under the more challenging setting, our model is capable of achieving the high-quality view synthesis, as depicted in Fig. 6. Furthermore, we addressed the narrow-baseline setting, which is considered a difficult task. We obtained only a few viewpoints with the reported narrow view ranges in the setting. As a result, our model trained under this setting outperforms InfoNeRF in all metrics, as listed in Tab. 5. The synthesized results are provided in the supplementary material.

6. Conclusion

We proposed a model capable of synthesizing high-quality novel views with only sparse inputs. Through the PA scheme, we extended unseen views around the reference viewpoints for an improved learning of the geometry and appearance. Additionally, by introducing two novel regularization methods, the MSC and IP, our model could synthesize superior images featuring precise structures and fine details. The MSC maintained semantic consistency at the local and global levels, and the IP attenuated the uncertainty resulting from data sparsity. We validated the performance of the proposed method through qualitative and quantitative analysis, respectively, demonstrating that our approach outperformed the previous methods on synthetic and real-world datasets.

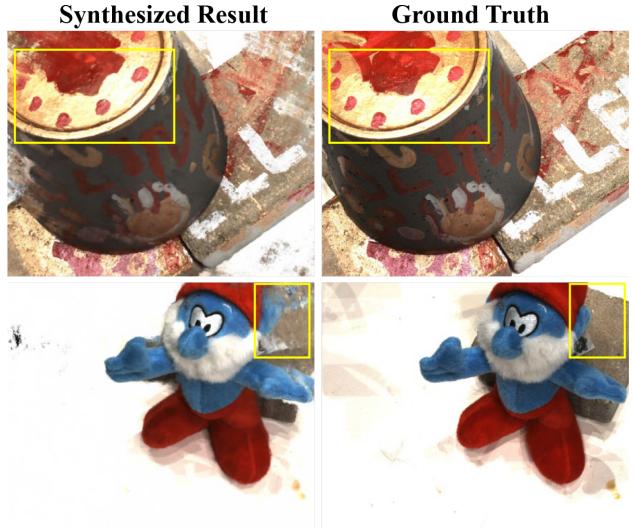


Figure 8. Limitations of our approach. The synthesized images are obtained from training our model on the DTU with 1-view.

Limitations and Future Work. In the real-world dataset, the lighting effect varied according to the camera poses, as observed in the DTU dataset. As we utilized few-shot inputs, the synthesized images could not represent the exact appearance reflecting the lighting condition. Therefore, our model could not produce an accurate appearance in a view with light reflection, as illustrated in the first view in Fig. 8. In addition, our model was limited in generating semantic properties of the occluded regions when trained on an extreme setting, such as a single-input view. Although we could extend the views through the PA scheme, a small amount of artifact was manifested in a part of the boundary, as depicted in the second view in Fig. 8. We shall address these limitations in future research.

References

- [1] 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 *ICCV*, 2021.
- [2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.
- [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.
- [4] Kangle Deng, Andrew Liu, Junyan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022.
- [5] John Flynn, Michael Broxton, Paul E. Debevec, Matthew DuVall, Graham Fyffe, Ryan S. Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *CVPR*, 2019.
- [6] Ajay Jain, Matthew Tancik, and P. Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021.
- [7] Seokhwan Jang, Seungeon Kim, Mina Kim, Kihong Son, Kyoung-Yong Lee, and Jong Beom Ra. Head motion correction based on filtered backprojection in helical ct scanning. *IEEE Transactions on Medical Imaging*, 2020.
- [8] Rasmus Ramsbøl Jensen, A. Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014.
- [9] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*, 2022.
- [10] Seungeon Kim, Yongjin Chang, and Jong Beom Ra. Cardiac image reconstruction via nonlinear motion correction based on partial angle reconstructed images. *IEEE Transactions on Medical Imaging*, 2017.
- [11] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. In *NeurIPS*, 2019.
- [12] Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. Neural volumes. *ACM TOG*, 2019.
- [13] Nelson L. Max. Optical models for direct volume rendering. *TVCG*, 1995.
- [14] 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 TOG*, 2019.
- [15] 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 *ECCV*, 2020.
- [16] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022.
- [17] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020.
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [19] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.
- [20] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:623–656, 2001.
- [21] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020.
- [22] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019.
- [23] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *CVPR*, 2019.
- [24] Worldsheet: Wrapping the World in a 3D Sheet for View Synthesis from a Single Image. Deepvoxels: Learning persistent 3d feature embeddings. In *ICCV*, 2021.
- [25] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020.
- [26] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd E. Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *CVPR*, 2022.
- [27] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibnet: Learning multi-view image-based rendering. In *CVPR*, 2021.
- [28] Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.
- [29] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, 2020.
- [30] Dongxin Xu and Deniz Erdogmuş. Renyi's entropy, divergence and their nonparametric estimators. In *Information Theoretic Learning*, 2010.
- [31] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *ECCV*, 2022.
- [32] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020.
- [33] Lin Yen-Chen. Nerf-pytorch. <https://github.com/yenchenlin/nerf-pytorch/>, 2020.

- [34] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021.
- [35] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.