NeRF-VPT: Learning Novel View Representations with Neural Radiance Fields via View Prompt Tuning

Linsheng Chen[1*], **Guangrun Wang**[2*],
Liuchun Yuan[1], **Keze Wang**[1†], **Ken Deng**[1], **Philip H.S. Torr**[2]

[1]Sun Yat-sen University
[2]University of Oxford
{kezewang, wanggrun,ylc0003}@gmail.com,{chenlsh36,dengk26}@mail2.sysu.edu.cn,philip.torr@eng.ox.ac.uk

Abstract

Neural Radiance Fields (NeRF) have garnered remarkable success in novel view synthesis. Nonetheless, the task of generating high-quality images for novel views persists as a critical challenge. While the existing efforts have exhibited commendable progress, capturing intricate details, enhancing textures, and achieving superior Peak Signal-to-Noise Ratio (PSNR) metrics warrant further focused attention and advancement. In this work, we propose NeRF-VPT, an innovative method for novel view synthesis to address these challenges. Our proposed NeRF-VPT employs a cascading view prompt tuning paradigm, wherein RGB information gained from preceding rendering outcomes serves as instructive visual prompts for subsequent rendering stages, with the aspiration that the prior knowledge embedded in the prompts can facilitate the gradual enhancement of rendered image quality. NeRF-VPT only requires sampling RGB data from previous stage renderings as priors at each training stage, without relying on extra guidance or complex techniques. Thus, our NeRF-VPT is plug-and-play and can be readily integrated into existing methods. By conducting comparative analyses of our NeRF-VPT against several NeRF-based approaches on demanding real-scene benchmarks, such as Realistic Synthetic 360, Real Forward-Facing, Replica dataset, and a user-captured dataset, we substantiate that our NeRF-VPT significantly elevates baseline performance and proficiently generates more high-quality novel view images than all the compared state-of-the-art methods. Furthermore, the cascading learning of NeRF-VPT introduces adaptability to scenarios with sparse inputs, resulting in a significant enhancement of accuracy for sparse-view novel view synthesis. The source code and dataset are available at https://github.com/Freedomcls/NeRF-VPT.

Introduction

In recent years, Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) have exhibited significant promise across a range of disciplines, including novel view synthesis (Tretschk et al. 2021; Tao et al. 2023), 3D reconstruction (Barron et al. 2021; Verbin et al. 2022; Mildenhall et al. 2022), and 3D editing (Kania et al. 2022). NeRF employs a neural network

---

*These authors contributed equally and are co-first authors.

†Corresponding Author.

to encapsulate the color and density attributes inherent to each point within 3D space. This encoding technique enables the synthesis of novel views of a scene, encompassing a comprehensive range of viewpoints. Nonetheless, as visually exemplified in Figure 1., generating high-quality images portraying novel views using NeRF remains challenging.

In recent times, a plethora of outstanding work endeavors have emerged to address this challenge, yielding encouraging results (Barron et al. 2021; Roessle et al. 2022; Varma T et al. 2022; Bautista et al. 2022; Vora et al. 2021; Johari, Lepoittevin, and Fleuret 2022; Tschernezki et al. 2022; Kurz et al. 2022; Arandjelović and Zisserman 2021; Hedman et al. 2021; Yariv et al. 2021; Tancik et al. 2022). While the strides made by current endeavors are considerable, there persists a clear necessity for a more intensified concentration on facets including the precise capture of intricate details (Bautista et al. 2022; Vora et al. 2021; Kurz et al. 2022), refinement of textures (Roessle et al. 2022; Vora et al. 2021; Kurz et al. 2022), and attainment of elevated Peak Signal-to-Noise Ratio (PSNR) metrics (Barron et al. 2021; Roessle et al. 2022; Varma T et al. 2022; Johari, Lepoittevin, and Fleuret 2022; Kurz et al. 2022).

In this paper, we introduce NeRF-VPT, a method that incorporates effective guidance to fully harness the inherent potential of NeRF-based models, resulting in more comprehensive and accurate outcomes. We adopt a cascading view prompt tuning paradigm, where RGB information obtained from preceding rendering iterations acts as informative visual prompts for subsequent rendering stages. The underlying goal is to leverage the prior knowledge embedded in these prompts to progressively enhance the quality of the rendered images. Our method substantially augments the guidance for both the reconstruction and rendering processes within the neural radiance field framework. It excels at iteratively combining comprehensive RGB information and spatial dependencies through a multi-stage refinement process, capitalizing on the benefits of cascading learning strategies (Lin et al. 2017; Ramakrishna et al. 2014).

As illustrated in Figure 1., our proposed NeRF-VPT has been meticulously designed to exhibit modularity and portability, thereby facilitating effortless integration with prior NeRF techniques such as vanilla NeRF, Mip-NeRF (Barron et al. 2021), and TensoRF (Chen et al. 2022a). This integration serves to further augment the performance of the prior
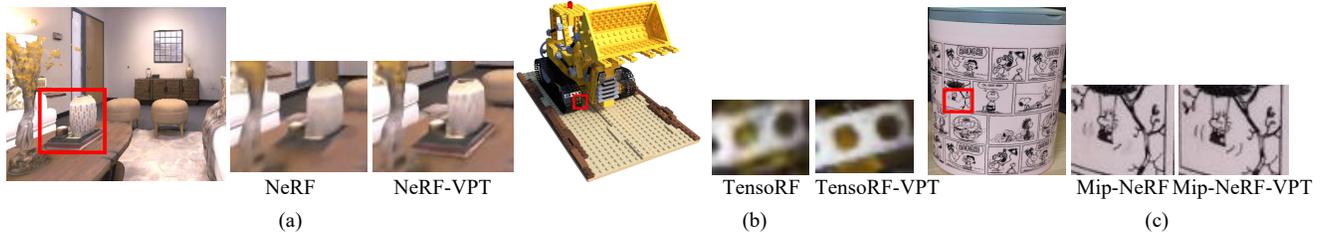
Figure 1: Qualitative comparison between our proposed NeRF-VPT and its corresponding baselines. (a) is on the Replica dataset. (b) is on the Realistic Synthetic 360 dataset. (c) is on a user-captured scene using a mobile phone.

NeRF approaches. The main contributions of our work are summarized as follows:

- We introduce a novel and efficacious framework, referred to as NeRF-VPT, featuring a cascading view prompt tuning designed to iteratively leverage the representation inherent to the NeRF family in the form of visual prompts.
- NeRF-VPT stands out for its inherent simplicity and seamless integration capabilities with pre-existing models, thereby elevating the performance of previous models to achieve heightened levels of excellence. Our method can be plug-and-play, which can be integrated into prior NeRF methods, such as vanilla NeRF, Mip-NeRF, and TensoRF.
- Through qualitative and quantitative experiments, we demonstrate that our method effectively improves the performance of view synthesis.

In summary, our approach offers a promising solution for enhancing the performance of NeRF-based view synthesis. Furthermore, our framework is plug-and-play, capable of augmenting the capabilities of existing NeRF methods.

Related Work

The NeRF architecture learns 3D scene representation through supervised learning from multi-view RGB images. It employs direct rays for pixel rendering and equidistant space sampling. NeRF's success in image synthesis has spurred significant research interest, resulting in various methods aiming to enhance its visual quality.

Priors in Scene Representation. RGB images alone lack comprehensive structural scene information, necessitating the incorporation of additional priors to guide the learning process. Recent advancements emphasize the efficacy of depth information as a valuable prior for refining scene representation (Deng et al. 2022; Roessle et al. 2022). Depth priors offer a suitable optimization direction, particularly during initial training stages, resulting in significant performance enhancements. Other studies (Chen et al. 2022b; Varma T et al. 2022; Bautista et al. 2022) have highlighted NeRF's limitations in adhering to specific spatial constraints, such as planar or the epipolar geometry. To mitigate these shortcomings, these approaches have embedded relative information within the NeRF framework, leading to improved consistency. These methods involve additional loss functions and constraints during optimization. However,

these approaches utilizing prior knowledge have limitations in their broader applicability. They often rely on external networks or necessitate specific data preprocessing steps. In contrast, our proposed method capitalizes exclusively on the network's internal output for prior knowledge, enhancing its potential for transferability and wider utility.

Improvements on Rendering. The rendering process plays a pivotal role in obtaining RGB images for novel views. Previous research endeavors (Zhi et al. 2021; Tian et al. 2021; Tschernezki et al. 2022; Guo et al. 2022) have predominantly concentrated on enhancing integration functions and devising more effective sampling methods. Acknowledging the challenge posed by single straight rays in capturing the influence of neighboring points, (Barron et al. 2021) introduced Mip-NeRF. This innovative approach substitutes a single ray with 3D conical frustums, yielding anti-aliased and visually crisp images in view synthesis. Furthermore, NeRF encounters limitations in terms of rendering's sampling effectiveness, which can restrict the optimization's coverage. To tackle these issues, approaches like those presented by (Kurz et al. 2022) and Arandjelović and Rieger (Arandjelović and Zisserman 2021) incorporate predictive networks to gauge the sampling's significance. This strategy reduces computational complexity in rendering, allowing for more resources to be allocated to expanding networks and enhancing result quality. Although these efforts have made significant strides, there remains a need for greater focus on capturing intricate details, refining textures, and achieving higher PSNR metrics.

Advanced Structure. Excellent endeavors strive to elevate the photorealistic synthesis of novel views through the incorporation of sophisticated architectures, such as Recursive-NeRF (Yang et al. 2022a) and BungeeNeRF (Xiangli et al. 2022). These approaches aim to partition the 3D space into more intricate or finely-grained segments, each corresponding to an individual network. This partitioning allows the scene to be represented across multiple levels of granularity, rendering it well-suited for encompassing diverse and intricate scenarios, such as natural landscapes, urban streets, and indoor environments.

Our approach sets itself apart from Recursive-NeRF and BungeeNeRF through the utilization of an innovative defective visual prompt tuning technique. This technique enables the integration of prior knowledge derived from earlier rendered outcomes, thereby serving as valuable guid-
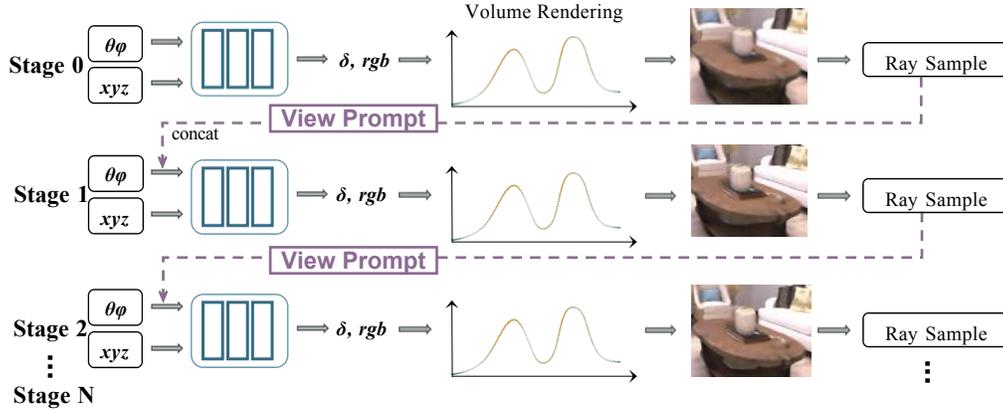
Figure 2: The overview of our proposed framework. In each iteration, the image rendered in the previous round is combined with the encoded direction after ray sampling, where stage 0 is a NeRF that does not rely on additional input. Each stage adopts a new network to learn with the previous RGB information.

ance for the network to gain a deeper understanding of intricate scenes. This, in turn, contributes to an enhancement in the quality of novel view synthesis.

Visual Prompt Tuning. Visual prompt learning (Jia et al. 2022) capitalizes on an extensive corpus of scholarly investigations encompassing domains such as visual question answering (Chappuis et al. 2022), few-shot learning (Alayrac et al. 2022), and generative models (Wu et al. 2022). Through the amalgamation of insights harvested from these scholarly realms, the field of visual prompt learning aspires to engender innovative methodologies for the generation of responses to visual inputs, empowering them to grasp and engender content from visual stimuli more adeptly. Our approach to prompt tuning significantly diverges from prevailing methods in the realm of visual prompt tuning. Specifically, we introduce a novel cascading view prompt tuning paradigm that embeds prior knowledge within our visual prompts. In this paradigm, RGB information derived from antecedent rendering iterations assumes the role of informative visual prompts for subsequent rendering stages.

Methodology

In this section, we commence by laying the groundwork with foundational concepts. We then proceed to the structure of NeRF-VPT. Following this, we elucidate the intricacies of the NeRF-VPT design, underscoring its effectiveness as a versatile and comprehensive framework.

Preliminary

For a point in scene $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{S}^2$, NeRF (Mildenhall et al. 2021) learns a function to query each point color $\mathbf{c}$ and voxel density σ with specific views:

$$(\mathbf{c}, \sigma) = F(\mathbf{x}, \mathbf{d}). \tag{1}$$

Specifically, the $\mathbf{c}$ and σ are output after F_c and F_d. In NeRF, $\mathbf{x}$ and $\mathbf{d}$ will present as position encoding and direction encoding:

$$\gamma(v) = (sin(2^0 \pi v), cos(2^0 \pi v), sin(2^1 \pi v), cos(2^1 \pi v), \\ ..., sin(2^{N-1} \pi v), cos(2^{N-1} \pi v)) \tag{2}$$

Volumetric rendering is proposed to create a 2D image based on F. It imagines each pixel emitting a single ray to 3D space from the image center $\mathbf{o}$, and sampling points along this ray, $\mathbf{x_i} = \mathbf{o} + t_i \mathbf{d}$. Thus, the color of this pixel can be approximated by numerical quadrature:

$$\hat{\mathbf{c}}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(t)\mathbf{c}(t)dt; \quad T(t) = e^{-\int_{t_n}^{t} \sigma(s)ds}, \tag{3}$$

where T handles the occlusion; t_n, t_f denote the near and far boundary of the 3d scene. Limited to computing resources, it adopts a coarse-to-fine sampling. The rendered color of each ray corresponds to the ground-truth value, thus the training is supervised by:

$$\mathcal{L} = \sum_{r \in \mathcal{R}} ||\hat{\mathbf{c}}(\mathbf{r}) - \mathbf{c}(\mathbf{r})||, \tag{4}$$

where $\mathcal{R}$ present all image pixels in training data, and $\mathbf{c}(\mathbf{r})$ is ground-truth color.

Priors within the scene can optimize sampling efficiency, offer an enhanced methodology for characterizing occlusion, and introduce supplementary loss functions in specific conditions. These priors are typically incorporated alongside the position $\mathbf{x}$. Consequently, the mapping function becomes: $(\mathbf{c}, \sigma) = F(\mathbf{x}, d, p)$, where p represents priors. The advancements in rendering primarily involve updates to Eq. (3), aimed at achieving more realistic results. These improvements can be implemented in various ways, rendering it intricate to consolidate them into a single comprehensive equation. Furthermore, these NeRF variants with hierarchical structures adopt a repetitive Eq. (1): $\{(\mathbf{c_i}, \sigma_i) = F_i(\mathbf{x_i}, d) | i \in N\}$. Some methods not only focus on a single improvement but also apply them synchronously.

Structural Design of NeRF-VPT

Most NeRF-based methods optimize a scene from random initialization. However, we believe that the inclusion of prior knowledge encompassing specific scene information could potentially simplify the model's comprehension of the

scene. Hence, our objective is to enable the network to directly learn the posterior distribution related to prior knowledge. To this end, we propose NeRF-VPT, a method that leverages prior knowledge sampled from the RGB space as a view prompt to aid the model in understanding the scene. Moreover, we introduce cascading learning to progressively enhance the quality of synthesized images. Figure 2. illustrates the framework of our approach.

View Prompt with Prior Knowledge

To incorporate scene information as prior knowledge embedded by a view prompt into the network, we conduct the following experiments: Initially, we attempt to extract features from multiple images in the scene and performed feature fusion by Swin-Transformer or another CNN-based network. The fused features are then incorporated into an MLP (Multi-Layer Perceptron) using the SIREN activation function. However, the experimental results show limited improvement, primarily because it is challenging to encapsulate multi-view information within a single feature. Furthermore, we explore the direct provision of ground truth images from the training set to the network during the training process for each viewpoint. This approach can be seen as providing the network with an answer during the learning process. The experimental results demonstrate a significant improvement in the quality of synthesized images during both the training and validation stages. However, this approach also introduces a challenge: we are unable to provide ground truth images for unknown viewpoints during the testing phase as prior knowledge.

In order to address this issue, we first employ NeRF to reconstruct the scene and render images from training, validation, and testing viewpoints. We incorporate this information as prior knowledge in the form as a prompt for the neural radiance field model for retraining, ensuring that the views during training, validation, and testing remain consistent with the NeRF training process. This training phase is fundamentally a guiding procedure facilitated through view prompt tuning (see Figure 2). Compared to directly using ground truth as priors, it can be understood as providing answers that may not be particularly accurate but effectively guide the optimization direction during the learning process. This approach can somewhat reduce the difficulty of learning while still benefiting from the supervision of ground truth. Importantly, during the testing phase, we can also sample images with an equivalent level of detail as prior knowledge in the form of a prompt.

As shown in Eq. (1) and (3), we obtain three essential components for a new view: RGB values, the color, and density of all sampling points in this scene. Among these, RGB values are the most informative, as they encode both the density and color of each ray. Moreover, incorporating RGB values as a prompt brings negligible complexity due to its lower dimensions. Experiments have shown that incorporating pixel values in this way leads to improved performance in novel view synthesis compared to using just the scene color and density information.

View Prompt Tuning with Cascading Learning

Cascade structures have been proven useful in various computer vision applications, such as object detection (Cai and Vasconcelos 2018), super-resolution (Saharia et al. 2022), and pose estimation (Lin et al. 2017). However, few works in the context of NeRF and beyond have made use of their advantages. Previous methods have focused on upgrading each part of NeRF, while NeRF-VPT disentangles the generation problem by refining it through a series of stages, sharing the merit of prompt tuning. One of the cornerstones of our method is to re-utilize the representation of the framework. For clarity, we employ a combination of conventional NeRF and view prompt tuning for illustrative purposes. It is important to highlight that our methods are plug-and-play for seamless integration with other NeRF variants, as demonstrated in our experiments.

We employ multi-steps to learn the representation of 3D scenes. In the first step, we use NeRF to gain knowledge of 3D space, which can generate the prior (value of RGB) as a view prompt for the next stage. As this step does not include any prior, NeRF-VPT also calls it the zero stage (stage 0 shown in Figure 2.). In the second and next steps, we insert the RGB as a view prompt and concatenate it to the direction encoding. The recurrent process can be formulated as:

$$
\begin{aligned}
(\mathbf{c_0}, \sigma_0) &= F_0(PE(\mathbf{x}), DE(\mathbf{d})); \\
(\mathbf{c_i}, \sigma_i) &= F_i(PE(\mathbf{x}), (DE(\mathbf{d}), \mathbf{c(r)_{i-1}}))
\end{aligned}
\tag{5}
$$

where F_0 and F_i note the mapping function in iterations.

We explore different methods to connect each stage, incorporating it into position encoding or directly taking it as raw input. We suppose that the encoding is signal decomposition, enabling NeRF to learn high-frequency information. The observation will significantly affect images during synthesis. Therefore, NeRF-VPT concatenates pixel value after direction encoding as the color is exactly high-frequency data. Moreover, the ablation studies prove the importance of connection with previous stages.

Experiments

The experiments are divided into the following parts: i) quantitatively and qualitatively showing that our method outperforms prior work; ii) verifying the plug-and-play capability; iii) verifying the adaptability to Sparse-View Novel View Synthesis (Wang et al. 2023a,b); iv) conducting a comprehensive ablation study, including depth comparison, the effect of prior knowledge and effect of prior model re-utilize.

Datasets

Realistic Synthetic 360 Dataset. This dataset consists of 4 Lambertian objects with simple geometry. Each object is rendered at 512x512 pixels from viewpoints sampled on the upper hemisphere. This dataset also consists of 8 objects of complicated geometry and realistic non-Lambertian materials. The resolution used for training is 400×400.

Real Forward-Facing Dataset (Mildenhall et al. 2019). This dataset consists of 8 forward-facing scenes captured with a cellphone at a size of 1008x756 pixels. Following the commonly used protocols (Mildenhall et al. 2021), we adopt 504×376 resolution for this dataset.

	Realistic Synthetic 360			Real Forward-Facing		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF	31.01	0.947	0.081	26.50	0.811	0.250
IBRnet	28.14	0.942	0.072	26.73	0.851	0.175
Recursive-NeRF	31.34	0.953	0.052	-	-	-
Mip-NeRF	33.09	0.961	0.043	-	-	-
TensoRF	33.14	0.963	0.044	27.13	0.839	0.164
Ours Stage 1	34.22	0.972	0.047	27.89	0.869	0.161
Ours Stage 6	**34.95**	**0.977**	**0.042**	**28.95**	**0.872**	**0.158**

Table 1: Quantitative comparison to baselines. A quantitative comparison of NeRF-VPT and baseline algorithms on the Realistic Synthetic 360 dataset and Real Forward-Facing dataset. '-' denotes the results are not available. Methods: NeRF (Mildenhall et al. 2021), IBRnet (Wang et al. 2021), Recursive-NeRF (Yang et al. 2022b), Mip-NeRF (Barron et al. 2021), TensoRF (Chen et al. 2022a).

User-captured Dataset. Compiled by capturing footage using a mobile phone, encompassing scenes with complex textures. We adopt 504×376 resolution for training.

Replicate Dataset (Straub et al. 2019). This dataset has high-quality reconstructions of various indoor spaces. Each reconstruction has clean dense geometry, high resolution, and high dynamic range textures, glass, and mirror surface information. The resolution used for training is 320x240.

Metrics. We report the mean of PSNR, structural similarity index (SSIM) (Wang et al. 2004), and the LPIPS (Zhang et al. 2018) perceptual metric for all evaluations.

Experimental Settings. During the training process, we begin by rendering images from a pre-trained NeRF model (or other NeRF-based models) for the training, validation, and testing viewpoints, which serve as prior knowledge for NeRF-VPT. In each subsequent training stage, the model relies solely on its output from the previous iteration as guidance. Loading weights from the previous training round is an option to facilitate faster convergence. Our experiments show that five to six stages on a standard baseline can achieve maximum performance, and training is concluded once this limit is reached. Unless otherwise specified, the following performance comparisons of NeRF-VPT are based on the model iterated to its performance upper limit.

Comparison with State-of-the-Art Methods

We compare our approach to several current state-of-the-art methods: NeRF, IBRnet, Recursive-NeRF, Mip-NeRF, and TensoRF. Neural Radiance Fields (NeRF) is a well-established method that employs Multi-Layer Perceptrons (MLPs) to implicitly represent scenes by utilizing encoded position and direction as input. IBRnet enhances the generalization capability of multi-scene rendering by learning from multi-view images. Mip-NeRF is an enhanced version of NeRF that employs 3D conical frustums instead of single rays, resulting in anti-aliased and visually clear synthesized views. TensoRF utilizes tensor decomposition to parameterize the model as a collection of low-rank tensor components, thereby improving the quality of reconstruction. We report the qualitative and quantitative comparison to baseline algorithms in Table 1. and Figure 3.. The experimental results

	Replica Dataset		User-Captured Dataset	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Mip-NeRF	33.12	0.961	26.73	0.841
Mip-NeRF-VPT	35.63	0.983	28.55	0.855
TensoRF	33.17	0.969	27.01	0.838
TensoRF-VPT	34.62	0.977	28.89	0.853

Table 2: Quantitative comparison of the results obtained by applying view prompt tuning on Mip-NeRF and TensoRF.

indicate that our method achieves outstanding performance in generating high-quality images from novel viewpoints. These metrics are obtained by calculating the average value of multiple scenes in each dataset. We use the related open-source projects to implement these methods. Additionally, in Figure 5., we present a visual comparison between our method and the baseline on real-world datasets.

Plug and Play

Our method is based on the framework of NeRF, essentially trying to improve the learning ability of MLPs for the scene via visual prompt tuning and reusing the model. For NeRF and a series of NeRF-based works, we can incorporate the view prompt of the model to improve the view synthesis quality through iterative methods. To fully verify the generalization ability of our NeRF-VPT, we conducted experiments not only with the NeRF baseline but also integrated our NeRF-VPT into other renowned methods such as Mip-NeRF and TensoRF. Based on Mip-NeRF and TensoRF, we add the view prompt to the network and iteratively learn the model. Our objective is to improve the quality of synthesized images based on these two methods. Figure 4. and Table 2. show our experimental results. We find our method raises NeRF-based model performance entirely to new levels via visual prompt tuning.

Adaptability to Sparse-View Novel View Synthesis

Neural Radiance Fields rely solely on RGB supervision information for comprehending the entire scene and generating images based on the viewing angle during the 3D reconstruction process. Consequently, it exhibits a strong reliance on densely sampled viewing angles. To address this limitation, we propose incorporating prior knowledge during the reconstruction stage to facilitate the network's understanding of the scene. This approach reduces the network's difficulty in comprehending the scene and consequently diminishes its dependence on densely sampled views. We conducted a comparative analysis of image quality between NeRF rendering using the original 180 views and our method utilizing multiple sparse views on the Replica dataset. The results, presented in Table 3., demonstrate that while training NeRF on 180 views and our method on 30, 45, and 60 views, the image quality achieved after training with approximately 45 views is comparable to NeRF training with 180 views. These results indicate that our method is capable of achieving a NeRF-like rendering effect even with sparser viewing angles, effectively emulating the dense viewing angle rendering capability of NeRF. In essence, our

NeRF   Mip-NeRF   TensoRF   NeRF-VPT   NeRF   Mip-NeRF   TensoRF NeRF-VPT   NeRF   Mip-NeRF   TensoRF   NeRF-VPT
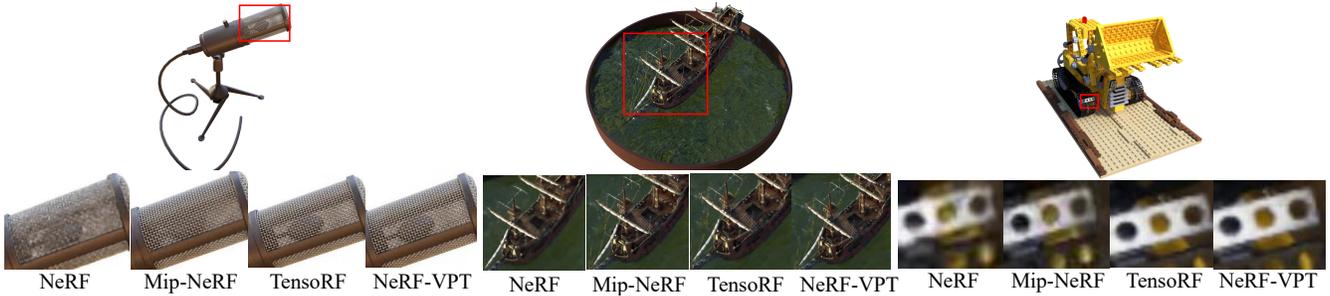
Figure 3: Qualitative results of NeRF-VPT and comparison methods (NeRF, Mip-NeRF, TensoRF) on Realistic Synthetic 360 Dataset. Please zoom in for better viewing.

Ground Truth   Mip-NeRF   Mip-NeRF-VPT   Ground Truth   TensoRF   TensoRF-VPT
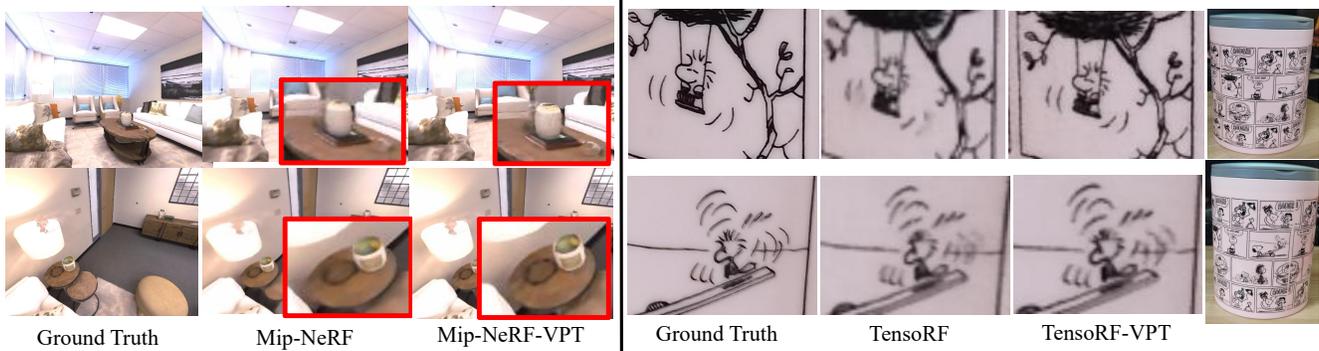
Figure 4: Qualitative evaluation of the improvement over Mip-NeRF and TensoRF on Replica dataset and user-captured dataset. Please zoom in for better viewing.

(1) Ground Truth   (2) NeRF(PSNR:23.88)   (3)NeRF-VPT(PSNR:25.74)

Figure 5: Visual comparisons on real-world datasets.

(1) Ground Truth   (2) NeRF(PSNR:23.21)   (3)NeRF-VPT(PSNR:27.14)

Figure 6: Visual comparisons in sparse view between NeRF and NeRF-VPT on forward-facing datasets.

method enables the achievement of NeRF-like high-quality rendering effects under sparse views. In Figure 6., we also present a visual comparison result of NeRF and NeRF-VPT under sparse views in the forward-facing dataset.

Ablation Study

Depth Comparison. While a notable enhancement in image quality, NeRF-VPT gains limited improvement in depth esti-

	Replica	
	PSNR↑	SSIM↑
NeRF (180view)	31.22	0.912
Ours Stage 1 (30view)	30.31	0.896
Ours Stage 1 (45view)	31.56	0.925
Ours Stage 1 (60view)	32.87	0.950

Table 3: Quantitative comparisons for the sampling views. We compared the reconstruction results of NeRF at 180 views and Ours at different views on the Replica dataset.

mation. As illustrated in Figure 7. and Table 4., the improved image quality does not seem to result in any significant qualitative or quantitative improvements in predicted depth. We suspect the task of image synthesis differs significantly from that of depth prediction, with the former requiring a smooth distribution and the latter requiring a sharp distribution. Nevertheless, we believe that the cascade structure employed by NeRF-VPT has the potential to be extended to other methods in 3D reconstruction and therefore remains a promising avenue for future research. The Replica dataset provides a real depth map, so we make a quantitative analysis of the depth predicted by the reconstruction model on it. The calculation of the depth distance refers to the form of Mean-Squared Loss. We calculated the MSE of the predicted depth map and the predicted depth map according to the ray sampling.

Computational Complexity. We compare the parameter

Ground Truth   Predicted Image   Predicted Depth   Details Patch
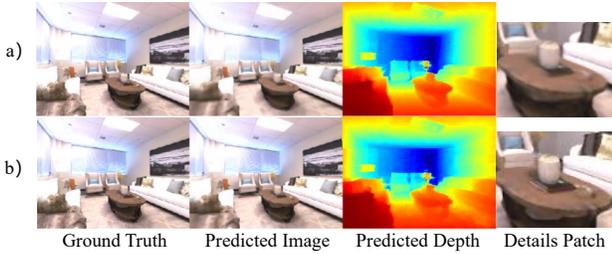
Figure 7: Qualitative comparison of different methods on the depth map. a) NeRF; b) NeRF-VPT. From left to right are the real image, the predicted image, the depth map, and the details of the predicted image. It can be seen from the figure that while improving the quality of the view synthesis, we also maintain the density of the reconstruction.

	PSNR↑	depth distance
NeRF	31.22	0.02
NeRF-VPT	34.95	0.02

Table 4: Quantitative comparisons for depth distance. We compared the depth performed by NeRF and NeRF-VPT on the Replica dataset, and the table shows that the predicted depth can be maintained while improving the image quality.

	Parameter Number	FPS
NeRF	1.1916 M	0.103
Ours Stage 1	1.1924 M	0.108

Table 5: Comparison on complexity with NeRF. FPS is the rendered frame per second.

	View Prompt	PSNR↑	SSIM↑
NeRF w/o prior	No Prior	31.22	0.912
NeRF-VPT	Feats	31.76	0.919
NeRF-VPT dir.	Gaussian Noise	11.58	0.567
NeRF-VPT dir.	Ground Truth	55.83	0.977
NeRF-VPT pos.	Last Stage Images	32.88	0.923
NeRF-VPT dir.	Last Stage Images	33.96	0.936

Table 6: Quantitative comparison of effect of prior knowledge. We show the comparative effects of adding prior knowledge to NeRF in different forms. Here, "Feats" means feature fusion as prior for training and testing; "dir. +" means incorporating prior to the direction; "pos. +" means incorporating prior to the position.

	PSNR↑	SSIM↑	Time(h)
Stage 0	31.22	0.912	8.78
Stage 1	33.96	0.936	2.42
Stage 2	34.76	0.943	1.96
Stage 3	34.99	0.947	1.54
Stage 4	35.19	0.951	1.22
Stage 5	35.32	0.953	0.92
Stage 6	35.39	0.954	0.61

Table 7: The performance of our method at each stage and the corresponding time consumed. Conducting experiments on the replica dataset.

number and the rendered frame per second (FPS) in Table 5.. Our method brings negligible cost to networks, with only 0.0008 M overhead. As we adopt recurrent modules, it is inevitable that NeRF-VPT requires more time for training. Table 7. outlines the training time of NeRF-VPT in each stage for 640 × 480 indoor scenes on Replica, showing our method is cost-effective: From Stage 1 to 6, the PSNR consistently increases, but the training time does not increase proportionally despite the visual prompts becoming more accurate and informative. As the iterations progress, the training time per round significantly decreases, primarily due to the increasing accuracy of the given visual prompts, which reduces the learning difficulty for the network.

Effect of Prior Knowledge. Prior knowledge plays a crucial role in our method, so how and where to incorporate the prior to warrant further study. Hence, we conducted a comparative analysis, which involved the following steps: i) Extracting and fusing features from scene images. We uniformly sampled 20 images from the training set to represent the entire scene and utilized the Swin-Transformer to extract their features. Subsequently, we employed soft-view pooling(Li et al. 2020) to fuse the extracted features. Finally, we integrated these fused features into each Linear Layer within the MLPs. ii) As discussed in the previous section on the method, we also explored the direct integration of ground truth as prior knowledge into the network. However, due to the unavailability of ground truth for unknown viewpoints during testing, we utilized the ground truth from the test set as prior in this set of experiments. Furthermore, we supplemented the results with experiments using Gaussian noise as a substitute for prior knowledge; iii)We also conducted a comparison to evaluate the disparities arising from incorporating prior knowledge in position encoding and direction encoding. The experimental results further validated our hypothesis: the ability of MLPs to predict color is more significantly influenced by the direction information. This can be attributed to the fact that, during the reconstruction of objects, after predicting volume density based on positional information, different direction information can induce diverse color variations for the same position. We did the above comparative experiments on the Replica dataset, and the detailed experimental results are in Table 6.

Effect of Prior Model Re-utilization. Our method enhances the quality of synthesized images by incorporating prior knowledge as guidance during the iterative process. We illustrate this evolution in Table 7, revealing that the performance reaches its peak after the sixth iteration stage.

Conclusion

This study proposes a novel and general framework to improve the performance of view synthesis based on NeRF. We present NeRF-VPT which introduces a new structure with recurrent modules and adopts the output of NeRF as priors. This allows NeRF-VPT to significantly improve the quality of view-dependent appearance. It is portable-friendly and is capable of combining with the existing methods to get state-of-art performance. We believe that this work provides new perspectives to take full advantage of representation.

Acknowledgements

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Arandjelović, R.; and Zisserman, A. 2021. Nerf in detail: Learning to sample for view synthesis. *arXiv preprint arXiv:2106.05264*.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Bautista, M. A.; Guo, P.; Abnar, S.; Talbott, W.; Toshev, A.; Chen, Z.; Dinh, L.; Zhai, S.; Goh, H.; Ulbricht, D.; et al. 2022. Gaudi: A neural architect for immersive 3d scene generation. *arXiv preprint arXiv:2207.13751*.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Chappuis, C.; Zermatten, V.; Lobry, S.; Le Saux, B.; and Tuia, D. 2022. Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1372–1381.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022a. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, 333–350. Springer.

Chen, Z.; Wang, C.; Guo, Y.-C.; and Zhang, S.-H. 2022b. StructNeRF: Neural Radiance Fields for Indoor Scenes with Structural Hints. *arXiv preprint arXiv:2209.05277*.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12882–12891.

Guo, H.; Peng, S.; Lin, H.; Wang, Q.; Zhang, G.; Bao, H.; and Zhou, X. 2022. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5511–5520.

Hedman, P.; Srinivasan, P. P.; Mildenhall, B.; Barron, J. T.; and Debevec, P. 2021. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5875–5884.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18365–18375.

Kania, K.; Yi, K. M.; Kowalski, M.; Trzciński, T.; and Tagliasacchi, A. 2022. Conerf: Controllable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18623–18632.

Kurz, A.; Neff, T.; Lv, Z.; Zollhöfer, M.; and Steinberger, M. 2022. AdaNeRF: Adaptive Sampling for Real-Time Rendering of Neural Radiance Fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, 254–270. Springer.

Li, L.; Zhu, S.; Fu, H.; Tan, P.; and Tai, C.-L. 2020. End-to-End Learning Local Multi-view Descriptors for 3D Point Clouds. *arXiv: Computer Vision and Pattern Recognition*.

Lin, M.; Lin, L.; Liang, X.; Wang, K.; and Cheng, H. 2017. Recurrent 3d pose sequence machines. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 810–819.

Mildenhall, B.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P. P.; and Barron, J. T. 2022. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16190–16199.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Ramakrishna, V.; Munoz, D.; Hebert, M.; Andrew Bagnell, J.; and Sheikh, Y. 2014. Pose machines: Articulated pose estimation via inference machines. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, 33–47. Springer.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12892–12901.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J.; Mur-Artal, R.; Ren, C. Y.; Verma, S.; Clarkson, A.; Mingfei, Y.; Budge, B.; Yan, Y.; Pan, X.; Yon, J.; Zou, Y.; Leon, K.; Carter, N.; Briales, J.; Gillingham, T.; Mueggler, E.; Pesqueira, L.; Savva, M.; Batra, D.; Strasdat, H.; Nardi, R. D.; Goesele, M.; Lovegrove, S.; and Newcombe, R. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv: Computer Vision and Pattern Recognition*.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.

Tao, T.; Gao, L.; Wang, G.; Chen, P.; Hao, D.; Liang, X.; Salzmann, M.; and Yu, K. 2023. LiDAR-NeRF: Novel LiDAR View Synthesis via Neural Radiance Fields. *arXiv preprint arXiv:2304.10406*.

Tian, Y.; Xie, L.; Zhang, X.; Fang, J.; Xu, H.; Huang, W.; Jiao, J.; Tian, Q.; and Ye, Q. 2021. Semantic-Aware Generation for Self-Supervised Visual Representation Learning. *arXiv preprint arXiv:2111.13163*.

Tretschk, E.; Tewari, A.; Golyanik, V.; Zollhöfer, M.; Lassner, C.; and Theobalt, C. 2021. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12959–12970.

Tschernezki, V.; Laina, I.; Larlus, D.; and Vedaldi, A. 2022. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. *arXiv preprint arXiv:2209.03494*.

Varma T, M.; Wang, P.; Chen, X.; Chen, T.; Venugopalan, S.; and Wang, Z. 2022. Is Attention All NeRF Needs? *arXiv e-prints*, arXiv–2207.

Verbin, D.; Hedman, P.; Mildenhall, B.; Zickler, T.; Barron, J. T.; and Srinivasan, P. P. 2022. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5481–5490. IEEE.

Vora, S.; Radwan, N.; Greff, K.; Meyer, H.; Genova, K.; Sajjadi, M. S.; Pot, E.; Tagliasacchi, A.; and Duckworth, D. 2021. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*.

Wang, G.; Chen, Z.; Loy, C. C.; and Liu, Z. 2023a. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *arXiv preprint arXiv:2303.16196*.

Wang, G.; Wang, P.; Chen, Z.; Wang, W.; Loy, C. C.; and Liu, Z. 2023b. PERF: Panoramic Neural Radiance Field from a Single Panorama. *arXiv preprint arXiv:2310.16831*.

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

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, C. H.; Motamed, S.; Srivastava, S.; and De la Torre, F. D. 2022. Generative visual prompt: Unifying distributional control of pre-trained generative models. *Advances in Neural Information Processing Systems*, 35: 22422–22437.

Xiangli, Y.; Xu, L.; Pan, X.; Zhao, N.; Rao, A.; Theobalt, C.; Dai, B.; and Lin, D. 2022. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, 106–122. Springer.

Yang, G.-W.; Zhou, W.-Y.; Peng, H.-Y.; Liang, D.; Mu, T.-J.; and Hu, S.-M. 2022a. Recursive-NeRF: An efficient and dynamically growing NeRF. *IEEE Transactions on Visualization and Computer Graphics*.

Yang, G.-W.; Zhou, W.-Y.; Peng, H.-Y.; Liang, D.; Mu, T.-J.; and Hu, S.-M. 2022b. Recursive-NeRF: An Efficient and Dynamically Growing NeRF. *IEEE Transactions on Visualization and Computer Graphics*, 1–14.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume Rendering of Neural Implicit Surfaces. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 4805–4815. Curran Associates, Inc.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhi, S.; Laidlow, T.; Leutenegger, S.; and Davison, A. J. 2021. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15838–15847.