From NeRFLiX to NeRFLiX++: A General NeRF-Agnostic Restorer Paradigm

Kun Zhou, Wenbo Li, Nianjuan Jiang, Xiaoguang Han, and Jiangbo Lu,

Abstract—Neural radiance fields (NeRF) have shown great success in novel view synthesis. However, recovering high-quality details from real-world scenes is still challenging for the existing NeRF-based approaches, due to the potential imperfect calibration information and scene representation inaccuracy. Even with high-quality training frames, the synthetic novel views produced by NeRF models still suffer from notable rendering artifacts, such as noise and blur. To address this, we propose NeRFLiX, a general NeRF-agnostic restorer paradigm that learns a degradation-driven inter-viewpoint mixer. Specially, we design a NeRF-style degradation modeling approach and construct large-scale training data, enabling the possibility of effectively removing NeRF-native rendering artifacts for deep neural networks. Moreover, beyond the degradation removal, we propose an inter-viewpoint aggregation framework that fuses highly related high-quality training images, pushing the performance of cutting-edge NeRF models to entirely new levels and producing highly photo-realistic synthetic views. Based on this paradigm, we further present NeRFLiX++ with a stronger two-stage NeRF degradation simulator and a faster inter-viewpoint mixer, achieving superior performance with significantly improved computational efficiency. Notably, NeRFLiX++ is capable of restoring photo-realistic ultra-high-resolution outputs from noisy low-resolution NeRF-rendered views. Extensive experiments demonstrate the excellent restoration ability of NeRFLiX++ on various novel view synthesis benchmarks.

Index Terms—Neural radiance field, degradation simulation, correspondence estimation, deep learning

1 Introduction

Photo-realistic novel view synthesis is a long-standing problem in the fields of computer vision and graphics. Recent years have seen the emergence of learning-based approaches, such as NeRF (Neural Radiance Fields) and its follow-ups, which utilize neural networks to represent 3D scenes and employ various rendering techniques to synthesize novel views. To achieve high-quality rendering, it is essential to design physically-aware systems that optimize multiple factors, including geometry, environment lighting, object materials, and camera poses. However, despite advancements, state-of-the-art NeRF models may still suffer from undesirable rendering artifacts when relying solely on a limited number of input views, as discussed in [20], [23], [33], [63], [79], [80].

Towards high-quality novel view synthesis, we propose NeRFLiX [83] that delivers pioneering efforts to investigate the feasibility of simulating large-scale NeRF-style paired data for training a NeRF-agnostic restorer. The system comprises two primary components: (1) a NeRF-style degradation simulator (NDS) and (2) an inter-viewpoint mixer (IVM). Inspired by practical image restoration approaches [57], [74], NeRFLiX systematically analyzes typical NeRF rendering artifacts and presents three manually designed degradations to simulate NeRF-rendered noises. We take advantage of NDS to generate a substantial amount of simulated training data and further develop a deep restorer, *i.e.*, IVM, to remove NeRF-style artifacts. Consequently, NeRFLiX demonstrates remarkable performance in synthesizing novel views of high fidelity, thereby extending the capabilities of NeRF models to new frontiers. However, there are two perspectives that deserve further investigation: (1) the inadequacy of manual degradation designs in accounting for the dispersion of real NeRF-rendered artifacts, and (2) the difficulty of employing the large inter-viewpoint mixer for processing high-resolution frames.

Hereafter, we extend NeRFLiX to NeRFLiX++ by introducing a two-stage degradation simulation approach, combined with a more efficient guided inter-viewpoint mixer. This refined framework not only achieves superior or comparable performance but also demonstrates significantly improved inference efficiency.

Two-stage degradation simulation. To bridge the domain gap between NeRF-rendered artifacts and simulated ones, we propose a two-stage degradation simulation scheme that consists of a hand-crafted degradation simulator and a deep generative degradation simulator. In the first stage, we utilize a similar degradation pipeline as NeRFLiX, but incorporate more basic degradations (*i.e.*, illumination jetting and brightness compression) to obtain an initially degraded frame. In the second stage, we leverage generative adversarial training to optimize the output from the first stage, making it statistically closer to NeRF-rendered views. However, training a deep generative network for our approach is challenging due to limited samples in the target domain. We observe that conventional pixel-to-pixel supervision actually constrains the diversity of simulated noise. To address this issue, we draw inspiration from Beby-GAN [29]

- *K. Zhou, N. Jiang and J. Lu are with SmartMore Co., Ltd., Shenzhen, China. (Corresponding email: jiangbo.lu@gmail.com)*
- *W. Li is with CSE, The Chinese University of Hong Kong, HongKong, China.*
- *X. Han is with Shenzhen Institute of Big Data, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China.*

Fig. 1: Visualization of restoration results of our proposed NeRFLiX++ for 4K images. It is clear that NeRFLiX++ produces photo-realistic 4K frames from low-resolution and noisy inputs rendered by DVGO [47].

and propose a novel approach that leverages image self-similarity and introduces a weighted top-K buddy loss for adversarial training. Specifically, given a simulated patch, we search for the K most relevant "buddies" (image patches) from the real sample (NeRF-rendered image), which are then used to provide weak supervision. This approach significantly enhances the diversity of generated patterns, resulting in improved degradation modeling. With the two-stage simulator, we are able to construct sizable training pairs and demonstrate that various deep restorers can be trained to effectively eliminate NeRF-style artifacts.

Guided inter-viewpoint mixer. To overcome the efficiency challenges of handling high-resolution frames for NeRFLiX, which incorporates a recurrent aggregation strategy to fuse details from reference views, we propose a more efficient guided inter-viewpoint aggregation scheme in NeRFLiX++. We achieve this by first estimating dense pixel-wise correspondences (optical flow) at a low resolution, based on several considerations. Firstly, the down-sampling operation results in smaller displacements between images, which lowers the difficulty of estimation. Secondly, the distributions of the rendered view and reference views become closer, resulting in more accurate correspondence estimation. Lastly, this approach is computationally more efficient. We then employ a coarse-to-fine guided aggregation by leveraging motion fields predicted at lower scales to aggregate information at higher scales. This strategy eliminates the need for recurrent high-resolution correspondence estimation, largely improving computational efficiency. Compared with NeRFLiX, NeRFLiX++ achieves superior results on benchmark datasets, such as Tanks and Temples and Noisy LLFF Synthetic, while performing on par with the LLFF dataset. Notably, NeRFLiX++ is 9.2× faster in processing scenes of a 1024×1024 size, highlighting its significant efficiency improvements.

In summary, our contributions are threefold:

- **Accurate NeRF Degradation Modeling.** We propose a two-stage degradation modeling scheme that closely approximates the statistical characteristics of real NeRF-rendered artifacts. Through this scheme, we demonstrate the effectiveness of existing deep image/video restorers and our proposed NeRFLiX/NeRFLiX++ in further enhancing the quality of NeRF-rendered views using simulated samples.
- **Efficient inter-viewpoint mixer.** We develop an efficient inter-viewpoint aggregation method that effectively integrates information from multiple viewpoints, enabling fast and accurate processing of ultra-high-resolution frames.
- **High-quality super-resolution.** Given the high efficiency of our accelerated inter-viewpoint aggregation, we demonstrate the potential of NeRFLiX++ to be extended to super-resolution tasks, generating photo-realistic 4K frames from noisy 1K NeRF-rendered views, as illustrated in Fig. 1.

A preliminary version of our work, NeRFLiX [83], has been accepted at the IEEE/CVF Conference on Computer Vision (CVPR) 2023. This extended version presents several key contributions and advancements. First, we address the limitations of hand-crafted degradations by introducing a novel two-stage degradation scheme that better models the complex distribution of NeRF-rendered frames. Second, we systematically analyze the efficiency of the recurrent inter-viewpoint mixer and propose a faster alternative. These improvements result in NeRFLiX++ achieving superior performance with significantly reduced computational costs. Moreover, we demonstrate that NeRFLiX++ can be easily applied to super-resolving photo-realistic 4K images from low-resolution NeRF-rendered views with minimal architecture modifications. The code of NeRFLiX++ will be released at https://redrock303.github.io/nerflix/ to facilitate future research.

2 Related Work

In this section, we review the relevant approaches consisting of NeRF-based novel view synthesis, degradation simulation in low-level version, and inter-frame correspondence estimation.

NeRF-based novel view synthesis. This field has received a lot of attention recently and has been thoroughly investigated. For the first time, Mildenhall *et al.* [39] propose the neural radiance field to implicitly represent static 3D scenes and synthesize novel views from multiple posed images. Inspired by their successes, a lot of NeRF-based models [2], [10], [12], [14], [20], [21], [22], [24], [26], [34], [36], [37], [40], [42], [44], [46], [49], [53], [55], [64], [67], [75], [78] have been proposed. For example, point-NeRF [65] and DS-NeRF [15] incorporate sparse 3D point cloud and depth information for eliminating the geometry ambiguity of NeRFs, achieving more accurate and

efficient 3D point sampling as well as better rendering quality. Plenoxels [17], TensoRF [9], DirectVoxGo [47], FastNeRF [18], Plenoctrees [69], KiloNeRF [45], and Mobilenerf [11], aim to use various advanced technologies to speed up the training or inference phases. Though these methods have achieved great progress, due to the potential issues of inaccurate camera poses, simplified pinhole camera models, and scene representation inaccuracy, they still suffer from rendering artifacts when predicting novel views.

Degradation simulation. Since there are currently no attempts to explore NeRF-style degradation, we overview the real-world image restoration works that are most related to ours. The previous image and video super-resolution approaches [16], [28], [29], [32], [56], [58], [70], [81], [81], [82] typically follow a fixed image degradation type (*e.g.*, blur, bicubic or bilinear down-sampling). Due to the large domain shift between the real-world and simulated degradations, the earlier image restoration methods [28], [30], [73], [81] generally fail to remove complex artifacts of the real-world images. In contrast, BSRGAN [74] designs a practical degradation approach for real-world image super-resolution. In their degradation process, multiple degradations are considered and applied in random orders, largely covering the diversity of real-world degradations. Compared with previous works, BSRGAN achieves much better results quantitatively and qualitatively. Real-ESRGAN [57] develops a second-order degradation process for real-world image super-resolution. Unlike the real-world image and video processing systems that focus on eliminating image and video compression, motion blur, video interlace, and sensor noise, the NeRF-rendering involves different degradation patterns. To the best knowledge, we are the first to investigate NeRF-style degradation removal.

Correspondence estimation. In the existing literature, video restoration methods [3], [7], [50], [54], [71] aim to restore a high-quality frame from multiple low-quality frames. To achieve this goal, cross-frame correspondence estimation is essential to effectively aggregate informative temporal contents. Some works [6], [7], [66], [71] build pixel-level correspondences through optical-flow estimation and perform frame warping for multi-frame compensation. Another line of works [51], [56], [82] tries to use deformable convolution networks (DCNs [13]) for adaptive correspondence estimation and aggregation. More recently, transformer-based video restoration models [5], [31] implement spatial-temporal aggregation through an attention mechanism and achieve promising performance. However, it is still challenging to perform accurate correspondence estimation between frames captured with very distinctive viewpoints.

3 Preliminaries

In this section, we review the general pipeline of NeRF-based novel view synthesis and discuss potential rendering artifacts. As shown in Fig. 2, three main steps are involved in the rendering. (1) Ray shooting. To render the color of a target pixel in a particular view, NeRF utilizes the camera's calibrated parameters π to generate a ray $\mathbf{r}(\mathbf{o}, \mathbf{d})$ through this pixel, where

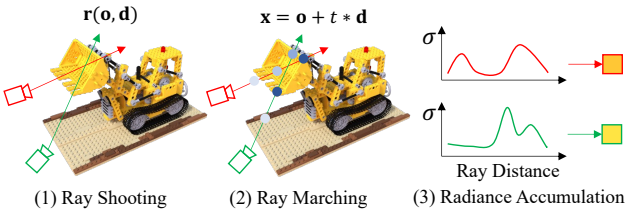


Fig. 2: A general illustration of NeRF-based novel view synthesis pipeline. Three main steps are involved: (1) ray shooting, (2) ray marching, and (3) radiance accumulation.

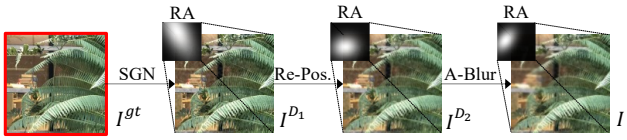


Fig. 3: Overview of our NDS pipeline in NeRFLiX: using our proposed degradations, we process a target view I^{gt} to produce its simulated degraded view I. "SGN", "Re-Pos." and "A-Blur" refer to the splatted Gaussian, re-positioning, anisotropic blur degradations, and "RA" is the region adaptive strategy.

$\mathbf{o}$ and $\mathbf{d}$ are the camera center and the ray direction. (2) Ray marching. A set of 3D points are sampled along the chosen ray as it moves across the 3D scene represented by neural radiance fields. NeRF encodes a 3D scene and predicts the colors and densities of these points. (3) Radiance accumulation. The pixel color is extracted by integrating the predicted radiance features of the sampled 3D points.

Discussion. We see that establishing relationships between 2D photos and the corresponding 3D scene requires camera calibration. Unfortunately, it is very challenging to precisely calibrate camera poses, leading to noisy 3D sampling. Meanwhile, some previous works [23], [61], [68], [72] also raise other concerns, including the non-linear pinhole camera model [23] and shape-radiance ambiguity [76]. Due to these inherent limitations, as discussed in Sec. 1, NeRF models may synthesize unsatisfied novel test views.

4 NeRFLiX

In this work, we present NeRFLiX, a general NeRF-agnostic restorer which employs a degradation-driven inter-viewpoint mixer to enhance novel view images rendered by NeRF models. It is made up of two essential components: a NeRF-style degradation simulator (NDS) and an inter-viewpoint mixer (IVM). As shown in Fig. 4a, during the training phase, we employ the proposed NDS to create large-scale paired training data, which is subsequently used to train an IVM for improving a NeRF-rendered view utilizing two reference images (reference views). In the inference stage, as illustrated in Fig. 4b, IVM is adopted to enhance a rendered view by fusing useful information from the selected most relevant reference views.

4.1 NeRF-Style Degradation Simulator (NDS)

Due to the difficulties in gathering well-posed scenes under various environments and training NeRF models for each scene,

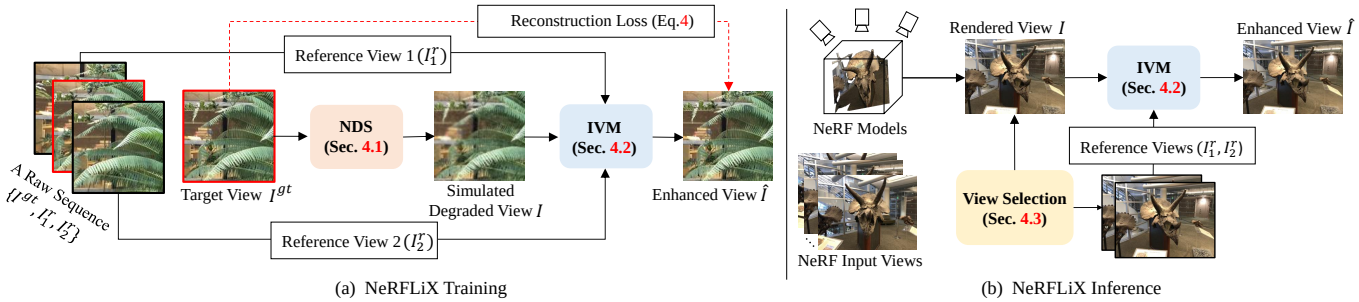


Fig. 4: Illustration of our proposed NeRFLiX. It consists of two essential modules: (1) NeRF degradation simulator that constructs paired training data $\{I, I_1^r, I_2^r | I^{gt}\}$ from a raw sequence $\{I^{gt}, I_1^r, I_2^r\}$, (2) inter-viewpoint mixer trained on this simulated data is capable of restoring high-quality frames from NeRF rendered views.

it is infeasible to directly collect large amounts of *paired* NeRF data for artifact removal. To address this challenge, motivated by BSRGAN [74], we design a general NeRF degradation simulator to produce a sizable training dataset that is visually and statistically comparable to NeRF-rendered images (views).

To begin with, we collect raw data from LLFF-T[1] and Vimeo90K [66] where the adjacent frames are treated as raw sequences. Each raw sequence consists of three images $\{I^{gt}, I_1^r, I_2^r\}$: a target view I^{gt} and its two reference views $\{I_1^r, I_2^r\}$. To construct the paired data from a raw sequence, we use the proposed NDS to degrade I^{gt} and obtain a simulated view I, as shown in Fig. 4(a).

The degradation pipeline is illustrated in Fig 3. We design three types of degradation for processing a target view I^{gt}: splatted Gaussian noise (SGN), re-positioning (Re-Pos.), and anisotropic blur (A-Blur). It should be noted that *there may be other models for such a simulation*, and we only utilize this route to evaluate and justify the feasibility of our idea.

Splatted Gaussian noise. Although additive Gaussian noise is frequently employed in image and video denoising, NeRF rendering noise clearly differs. Rays that hit a 3D point will be re-projected within a nearby 2D area because of noisy camera parameters. As a result, the NeRF-style noise is dispersed over a 2D space. This observation led us to present a splatted Gaussian noise, which is defined as

$$I^{D1} = (I^{gt} + n) \circledast g\,, \tag{1}$$

where n is a 2D Gaussian noise map with the same resolution as I^{gt} and g is an isotropic Gaussian blur kernel.

Re-positioning. We design a re-positioning degradation to simulate ray jittering. We add a random 2D offset $\delta_i, \delta_j \in [-2, 2]$ with probability 0.1 for a pixel at location (i, j) as

$$I^{D2}(i,j) = \begin{cases} I^{D1}(i,j) & \text{if } \ p > 0.1 \\ I^{D1}(i+\delta_i, j+\delta_j) & \text{else } \ p \le 0.1\,, \end{cases} \tag{2}$$

where p is uniformly distributed in $[0, 1]$.

Anisotropic blur. From our observation, NeRF synthetic frames also contain blurry contents. To simulate blur patterns, we use anisotropic Gaussian kernels to blur the target frame.

Neural radiance fields are often supervised with unbalanced training views. As a result, given a novel view, the projected 2D areas have varying degradation levels. Thus, we carry out each of the employed degradations in a spatially variant manner. More specifically, we define a mask M as a two-dimensional oriented anisotropic Gaussian [19] like

$$M(i,j) = G(i - c_i, j - c_j; \sigma_i, \sigma_j, A)\,, \tag{3}$$

where (c_i, c_j) and (σ_i, σ_j) are the means and standard deviations, and A is an orientation angle. After that, we use the mask M to linearly blend the input and output of each degradation, finally achieving region-adaptive degradations.

At last, with our NDS, we can obtain a great number of training pairs, and each paired data consists of two high-quality reference views $\{I_1^r, I_2^r\}$, a simulated degraded view I, and the corresponding target view I^{gt}. Next, we show how the constructed paired data $\{I, I_1^r, I_2^r | I^{gt}\}$ can be used to train our IVM.

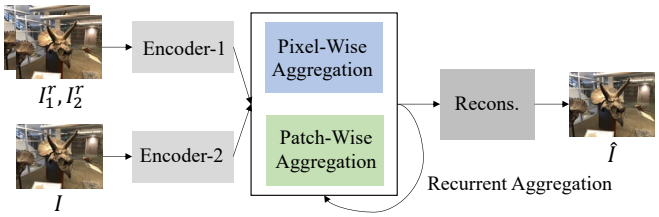


Fig. 5: The framework of our inter-viewpoint mixer in NeRFLiX.

4.2 Inter-viewpoint Mixer (IVM)

Problem formulation. Given a degraded view I produced by our NDS or NeRF models, we aim to extract useful information from its two high-quality reference views $\{I_1^r, I_2^r\}$ and restore an enhanced version $\hat{I}$.

IVM architecture. For multi-frame processing, the existing techniques either use optical flow [6], [54], [71] or deformable convolutions [13], [31], [56] to realize the correspondence estimation and aggregation for *consistent* displacements. In contrast, NeRF-rendered and input views may come from very different angles and locations, making it challenging to perform precise aggregation.

To address this problem, we propose IVM, a hybrid recurrent inter-viewpoint "mixer" that progressively fuses pixel-wise and patch-wise contents from two high-quality reference

1. The training part of LLFF [38].

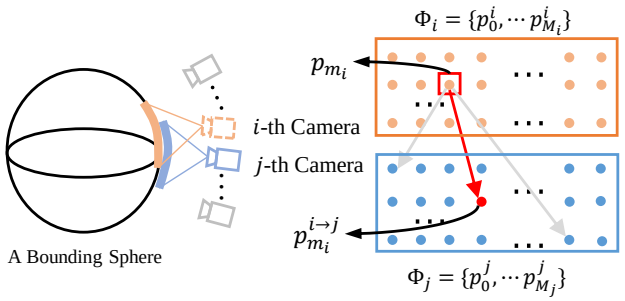


Fig. 6: Illustration of our view selection strategy.

views, achieving more effective inter-viewpoint aggregation. There are three modules, *i.e.*, feature extraction, hybrid inter-viewpoint aggregation, and reconstruction, as shown in Fig. 5. Two convolutional encoders are used in the feature extraction stage to process the degraded view I and two high-quality reference views $\{I_1^r, I_2^r\}$, respectively. We then use inter-viewpoint window-based attention modules and deformable convolutions to achieve recurrent patch-wise and pixel-wise aggregation. Finally, the enhanced view $\hat{I}$ is generated using the reconstruction module under the supervision

$$Loss = |\hat{I} - I^{gt}|, \text{where } \hat{I} = f(I, I_1^r, I_2^r; \theta), \tag{4}$$

where θ is the learnable parameters of IVM. The architecture details are given in our supplementary materials.

4.3 View Selection

In the inference stage, for a NeRF-rendered view I, our IVM produces an enhanced version by aggregating contents from two neighboring high-quality views. Though multiple high-quality views (provided for the training) are available, only a part of them is largely overlapped with I. We only adopt the most pertinent views that are useful for the inter-viewpoint aggregation.

To this end, we develop a view selection strategy to choose two reference views $\{I_1^r, I_2^r\}$ from the input views that are most overlapped with the rendered view I. Specifically, we formulate the view selection problem based on the pinhole camera model. An arbitrary 3D scene can be roughly approximated as a bounding sphere in Fig. 6, and cameras are placed around it to take pictures. When camera-emitted rays hit the sphere, there are a set of intersections. We refer to the 3D point sets as $\Phi_i = \{p_0^i, p_1^i, \cdots, p_{M_i}^i\}$ and $\Phi_j = \{p_0^j, p_1^j, \cdots, p_{M_j}^j\}$ for the i-th and j-th cameras. For m_i-th intersection $p_{m_i}^i \in \Phi_i$ of view i, we search its nearest point in view j with the L2 distance

$$p_{m_i}^{i\to j} = \arg\min_{p\in\Phi_j}(\|p - p_{m_i}^i\|_2^2). \tag{5}$$

Then the matching cost from the i-th view to the j-th view is calculated by

$$C_{i\to j} = \sum_{m_i=0}^{M_i} \|p_{m_i}^i - p_{m_i}^{i\to j}\|_2^2. \tag{6}$$

We finally obtain the mutual matching cost between views i and j

$$C_{i\leftrightarrow j} = C_{i\to j} + C_{j\to i}. \tag{7}$$

In this regard, two reference views $\{I_1^r, I_2^r\}$ are selected at the least mutual matching costs for enhancing the NeRF-rendered view I. Note that we also adopt this strategy to decide the two reference views for the LLFF-T [38] data during the training phase.

5 NeRFLiX++

Based on NeRFLiX, we propose NeRFLiX++ with a two-stage degradation modeling strategy and a guided inter-viewpoint mixer to further improve restoration performance and efficiency.

5.1 Two-stage Degradation Modeling

The proposed two-stage degradation modeling approach comprises a manually designed degradation simulator and a deep generative degradation simulator, as depicted in Fig. 7. In the first stage, we generate initialized degraded frames using multiple hand-crafted degradations, inspired by NeRFLiX, from the selected clean views. In the second stage, the deep generative degradation simulator is employed to refine the first-stage results and generate the final simulated views.

5.1.1 Manual Degradation Simulator

In addition to the three basic degradations used in NeRFLiX, which are splatted Gaussian noise, re-positioning, and anisotropic Gaussian blur, we introduce two supplementary degradation patterns to enhance the realism of our simulation. We apply the same region-adaptive degradation strategy as NeRFLiX for these two additional degradations.

Illumination jetting. To account for the variation in illumination caused by view-dependent shading, we propose a gamma adjustment applied to *both* the target and reference views. The adjustment is defined as

$$y = \text{power}(x, \gamma), \tag{8}$$

where "power" denotes the exponential function and γ is a linear adjustment constant randomly sampled from $[0.95, 1.05]$.

Lightness compression. To simulate structural defects that may occur in NeRF-based rendering, we propose an image compression procedure that degrades the gray-scale density of a target frame. Specifically, we first convert an RGB frame to the LAB color space and compress the L component using the JPEG algorithm at a randomly selected compression level (between 20% and 90%). We then merge the degraded L channel with the raw AB channels and transform them back to the RGB color space.

5.1.2 Deep Generative Degradation Simulator

As noted in Sec. 1, manually designed degradations may not capture the full range of actual NeRF-style artifacts. To address this limitation, we propose a deep generative degradation simulator that refines the results of the manual degradation stage and narrows the gap between the simulated and target domains.

Generative adversarial networks (GANs) have shown remarkable results in image-to-image translation tasks when a large number of training samples are available. However, the

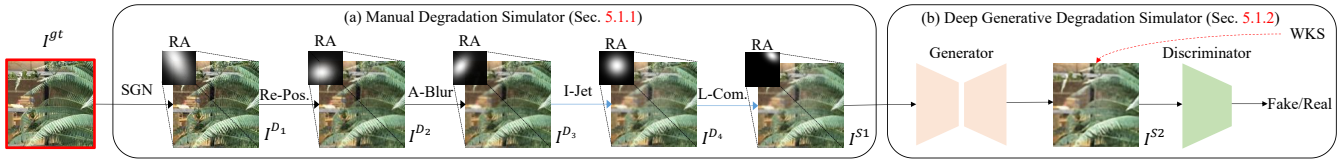


Fig. 7: The pipeline of our two-stage degradation modeling consisted of two sequentially stacked simulators: (a) a manual degradation simulator to get an initialized result I^{S1} from a clean target frame, (b) a deep generative degradation simulator that receives the I^{S1} and outputs the final degraded frame I^{S2} using an adversarial learning scheme. Additionally, we introduce a weighted top-K supervision (WKS) to enhance the degradation diversity of refined views I^{S2}.

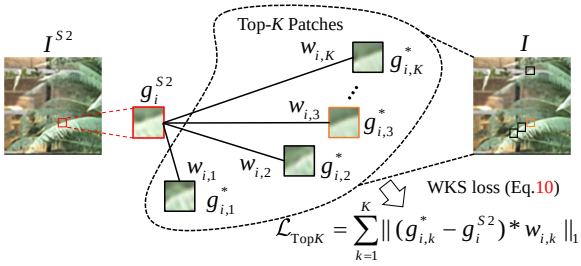


Fig. 8: Illustration of the proposed weighted top-K similarity loss. Where I^{S2} and I are the second-stage degraded frame and the corresponding NeRF-rendered views. $\mathbf{g}^*_{i,\{1,2,\cdots,K\}}$ are the top-K patches from I and $w_{i,\{1,2,\cdots,K\}}$ are the corresponding weighted factors calculated by Eq. (11) for adaptive supervision. Also, we highlight the pre-defined patch $\mathbf{g}_i^{gt}$ that is only ranked third.

scarcity of available data from Neural Radiance Fields (NeRF) poses significant challenges in using GANs to directly fit the underlying degradation distribution. To address this issue, motivated by Beby-GAN [29], we propose a weighted top-K similarity loss, or WKS, as an auxiliary loss function to aid in adversarial training. As shown in Fig. 7, we use a UNet to process the first-stage degraded view I^{S1} to obtain the refined result I^{S2}. In addition to the conventional adversarial and reconstruction losses, we also utilize the WKS to produce results I^{S2} with more diversity.

WKS. Fig. 8 illustrates the weighted top-K supervision. Given the i-th patch of I^{S2}, denoting $\mathbf{g}_i^{S2}$, we use a triple distance function to search for the top-K similar patches $\mathbf{g}^*_{i,\{1,2,\cdots,K\}}$ from the corresponding real rendered view I. The triple distance function is defined as

$$\mathbf{g}^*_{i,\{1,2,\cdots,K\}} = \underset{\mathbf{g}\in\mathcal{G}}{\text{topK}}\ \alpha\left\|\mathbf{g}-\mathbf{g}_i^{S2}\right\|_2^2 + \beta\left\|\mathbf{g}-\mathbf{g}_i^{gt}\right\|_2^2, \tag{9}$$

where $\mathbf{g}_i^{gt}$ is the corresponding real rendered patch, $\mathcal{G}$ is a set of candidate patches generated by unfolding the real-rendered view I, and α,β are two scaling factors to balance the two distance terms. According to the empirical experiments in Beby-GAN [29], we set them to 1 for better evaluation results. After obtaining the top-K similar patches, the proposed WKS is formulated as

$$\mathcal{L}_{\text{TopK}} = \sum_{k=1}^{K}\left\|(\mathbf{g}^*_{i,k}-\mathbf{g}_i^{S2})*w_{i,k}\right\|_1, \tag{10}$$

where $w_{i,k}$ is the k-th normalized weight, calculated as

$$\begin{aligned} d_{i,k} &= -\frac{1}{2}\|\mathbf{g}^*_{i,k}-\mathbf{g}_i^{S2}\|_2^2, \\ w_{i,k} &= \exp(d_{i,k})/\sum_{m=1}^{K}\exp(d_{i,m}). \end{aligned} \tag{11}$$

The $d_{i,k}$ is the scaled negative L2 distance between the predicted patch $\mathbf{g}_i^{S2}$ and one of its k-th most similar patch.

Discussion. Our proposed weighted top-K similarity loss adopts a dynamic strategy to search for multiple pertinent patches from the real rendered frames, enriching the diversity of supervisory signals. This approach encourages the model to find highly similar target patches that have closer degradation degrees than the pre-defined label, resulting in more accurate and effective training. In our experiments, we demonstrate the effectiveness of this design and show that it significantly improves the performance of GANs when limited data from NeRF-rendered frames is available.

5.2 Guided Inter-viewpoint Mixer

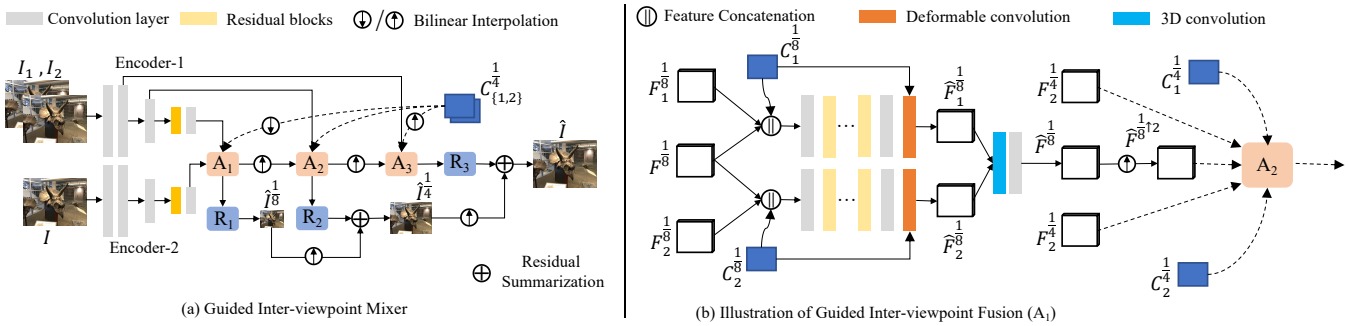
Under the typical NeRF setup, the high-quality input views come for free and they serve as potential reference bases for the restoration of rendered views. To achieve inter-viewpoint mixing, NeRFLiX presents a recurrent aggregation model to handle the distinct viewpoint changes. However, as aforementioned in Sec. 1, it remains impractical to process high-resolution frames due to the high computational expenses.

To overcome this limitation, we propose a guided inter-viewpoint mixer (termed as "G-IVM") with an efficient multi-view fusion module. Fig. 9 depicts the framework architecture of G-IVM. Our approach first utilizes an off-the-shell optical flow model to predict coarse correspondences between a rendered view I and its reference views $\{I_1, I_2\}$[2] at a low resolution. Building upon coarse predictions as guidance, we propose a pyramid neural network to conduct a coarse-to-fine aggregation.

Our guided inter-viewpoint mixer, an extension of the IVM method introduced in NeRFLiX [83], comprises three integral modules: feature extraction, guided inter-viewpoint aggregation, and pyramid reconstruction.

Coarse corresponding estimation. In order to establish coarse correspondences between a given input view I and its reference

2. To provide a more concise description of G-IVM, we omit the upper letter ^{r} in this section and the subsequent ones.



(a) Guided Inter-viewpoint Mixer

(b) Illustration of Guided Inter-viewpoint Fusion (A_1)

Fig. 9: The framework of our guided inter-viewpoint mixer (G-IVM) in NeRFLiX++. (a) The overview of our proposed G-IVM, which consists of three integral modules: (1) two parallel convolutional encoders are employed to extract deep image features from input and reference views, (2) taking the estimated coarse optical flow as guidance, we devise a guided inter-viewpoint aggregation module to progressively fuse pyramid deep image features, and (3) we utilize three reconstruction blocks to gradually restore multi-scale frames in a coarse-to-fine manner. Meanwhile, we adopt a multi-scale supervision scheme to enhance the restoration quality. (b) We outline the detailed structure of the guided inter-viewpoint aggregation. "$A_{\{1,2,3\}}$" means the three pyramid aggregation blocks, and "$R_{\{1,2,3\}}$" refer to the three reconstruction blocks to progressively produce multi-scale frames $\hat{I}^{\{\frac{1}{8},\frac{1}{4}\}}$ and $\hat{I}$.

Method	PSNR (dB)↑	SSIM↑	LPIPS↓
TensoRF [9] (ECCV'22)	26.73	0.839	0.204
TensoRF [9] + NeRFLiX	27.39 (↑ 0.66)	0.867	0.149
TensoRF [9] + NeRFLiX++	27.38 (↑ 0.65)	0.866	0.156
Plenoxels [17] (CVPR'22)	26.29	0.839	0.210
Plenoxels [17] + NeRFLiX	26.90 (↑ 0.61)	0.864	0.156
Plenoxels [17] + NeRFLiX++	26.92 (↑ 0.63)	0.864	0.160
NeRF-mm [61] (ARXIV'21)	22.98	0.655	0.440
NeRF-mm [61] + NeRFLiX	23.38 (↑ 0.40)	0.694	0.360
NeRF-mm [61] + NeRFLiX++	23.40 (↑ 0.42)	0.698	0.354
NeRF [39] (ECCV'20)	26.50	0.811	0.250
NeRF [39] + NeRFLiX	27.26 (↑ 0.76)	0.863	0.159
NeRF [39] + NeRFLiX++	27.25 (↑ 0.75)	0.858	0.170

(a) Quantitative results on LLFF [38] under LLFF-P1.

Method	PSNR (dB)↑	SSIM↑	LPIPS↓
NLF [1] (CVPR'22)	27.46	0.868	0.136
NLF [1] + NeRFLiX	28.19 (↑ 0.73)	0.899	0.093
NLF [1] + NeRFLiX++	28.10 (↑ 0.64)	0.895	0.093
RegNeRF-V3 [41] (CVPR'22)	19.10	0.587	0.373
RegNeRF-V3 [41] + NeRFLiX	19.68 (↑ 0.58)	0.661	0.260
RegNeRF-V3 [41] + NeRFLiX++	19.85 (↑ 0.75)	0.670	0.258
RegNeRF-V6 [41] (CVPR'22)	23.06	0.759	0.242
RegNeRF-V6 [41] + NeRFLiX	23.90(↑ 0.84)	0.815	0.144
RegNeRF-V6 [41] + NeRFLiX++	24.01(↑ 0.95)	0.816	0.152
RegNeRF-V9 [41] (CVPR'22)	24.81	0.818	0.196
RegNeRF-V9 [41] + NeRFLiX	25.68 (↑ 0.87)	0.863	0.114
RegNeRF-V9 [41] + NeRFLiX++	25.76 (↑ 0.95)	0.861	0.124

(b) Quantitative results on LLFF under LLFF-P2. RegNeRF-V3(6,9) takes 3(6,9) input views for training.

Model	Parameters	Inference Resources @ 512×512	Inference Resources @1024×1024	Inference Resources @2048×2048
NeRFLiX	35.2M	917ms / 4.2GB	4005ms / 12.6GB	-
NeRFLiX++	**14.4M**	**109ms / 2.8GB**	**433ms / 7.1GB**	**1588ms / 24.2GB**

(c) Model complexity and inference efficiency comparison between NeRFLiX and NeRFLiX++. The inference resources refer to the GPU memory usage and inference time. We use an NVIDIA RTX 3090 to test three resolutions of 512×512, 1024×1024, and 2048×2048. "-" means the result is unavailable due to out-of-memory.

TABLE 1: Quantitative analysis of our NeRFLiX on LLFF [38]. Best and second best results are highlighted in red and blue.

views $\{I_1, I_2\}$, we utilize a pre-trained SPyNet [43] model to predict the optical flow $C^{\frac{1}{4}}_{\{1,2\}}$ at a down-sampled scale of $\frac{1}{4}$.

Feature extraction. We introduce two convolutional encoders, referred to as "Encoder-1/2", to extract deep pyramid image features $F^{\{\frac{1}{8},\frac{1}{4},\frac{1}{2}\}}$ and $F^{\{\frac{1}{8},\frac{1}{4},\frac{1}{2}\}}_{\{1,2\}}$ from a rendered view I and its two reference views $I_{\{1,2\}}$. Specifically, the pyramid features are at scales of $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$ by applying three convolutions with a stride length of 2.

Guided inter-viewpoint aggregation. Considering the difficulties associated with accurately estimating large displacements between a rendered view I and its reference views $\{I_1, I_2\}$, we present a guided inter-viewpoint aggregation method that operates in a coarse-to-fine manner. Our approach employs the flow-guided deformable convolution (FDCN) technique, leveraging optical flow computed by the SPyNet to facilitate the aggregation of $F^{\frac{1}{8}}$ and its corresponding reference views $F^{\frac{1}{8}}_1, F^{\frac{1}{8}}_2$. The process is formulated as

$$\begin{aligned} C_i^{\frac{1}{8}} &= \frac{1}{2}\text{Bilinear}(C_i^{\frac{1}{4}}, \frac{1}{2}), \\ M_i^{\frac{1}{8}} &= [F^{\frac{1}{8}}, F_i^{\frac{1}{8}}, C_i^{\frac{1}{8}}], \\ \hat{F}_i^{\frac{1}{8}} &= \text{FDCN}(F_i^{\frac{1}{8}}, M_i^{\frac{1}{8}}, C_i^{\frac{1}{8}}), \end{aligned} \tag{12}$$

where $i \in \{1, 2\}$ is the reference index, $\text{Bilinear}(\cdot, s)$ is a bilinear interpolation function (s is the scaling factor), $M_i^{\frac{1}{8}}$ is an offset feature, and $\hat{F}_i^{\frac{1}{8}}$ denotes an aligned feature from the i-th reference

view to the target image. Having obtained the two aggregated features, denoted as $\hat{F}^{\frac{1}{8}}_{\{1,2\}}$, we employ a 3D convolution layer to fuse them with the target-view feature $F^{\frac{1}{8}}$:

$$\hat{F}^{\frac{1}{8}} = \text{Conv3D}(\hat{F}^{\frac{1}{8}}_{\{1,2\}}, F^{\frac{1}{8}}), \tag{13}$$

where $\hat{F}^{\frac{1}{8}}$ is the fused feature.

Moving forward, we proceed with the $\frac{1}{4}$-scale aggregation stage. Instead of utilizing the feature $F^{\frac{1}{4}}$ directly, we opt for the 2× upsampled counterpart $\hat{F}^{\frac{1}{8}\uparrow 2}$ as the target-view feature. This choice is motivated by the presence of potential artifacts in the rendered view I. Given that $\hat{F}^{\frac{1}{8}}$ has already aggregated high-quality details from the reference views, it is deemed more appropriate for conducting correspondence estimation involving $F_i^{\frac{1}{4}}$ and the target view:

$$\begin{aligned} F_i^{\frac{1}{8}\uparrow 2} &= \text{Bilinear}(\hat{F}^{\frac{1}{8}}, 2), \\ M_i^{\frac{1}{4}} &= [F_i^{\frac{1}{8}\uparrow 2}, F_i^{\frac{1}{4}}, C_i^{\frac{1}{4}}], \\ \hat{F}_i^{\frac{1}{4}} &= FDCN(F_i^{\frac{1}{4}}, M_i^{\frac{1}{4}}, C_i^{\frac{1}{4}}). \end{aligned} \tag{14}$$

Afterwards, we take another 3D convolution to mix the two aggregated features $\hat{F}^{\frac{1}{4}}_{\{1,2\}}$:

$$\hat{F}^{\frac{1}{4}} = Conv3D(\hat{F}^{\frac{1}{4}}_{\{1,2\}}, F^{\frac{1}{4}}). \tag{15}$$

Finally, we conduct the third-level aggregation to obtain the fused feature $\hat{F}^{\frac{1}{2}}$, using similar processing steps as in the second stage.

Pyramid reconstruction and multi-scale supervision. We employ pyramid aggregated features $\hat{F}^{\{\frac{1}{8},\frac{1}{4},\frac{1}{2}\}}$ to generate multi-scale outputs. Initially, starting from $\hat{F}^{\frac{1}{8}}$, we employ convolutional blocks to obtain the lowest-scale output $\hat{I}^{\frac{1}{8}}$. Subsequently, we upsample this output and incorporate $\hat{F}^{\frac{1}{4}}$ to learn the image residue at a higher scale, yielding $\hat{I}^{\frac{1}{4}}$. By following this strategy, we ultimately predict the enhanced view $\hat{I}$. To improve reconstruction quality, we incorporate multi-scale supervision during training:

$$\begin{aligned} L_{\{\frac{1}{8},\frac{1}{4}\}} &= \|\hat{I}^{\{\frac{1}{8},\frac{1}{4}\}} - I_{gt}^{\{\frac{1}{8},\frac{1}{4}\}}\|_1; \\ L_f &= \|\hat{I} - I_{gt}\|_1; \\ L &= 0.1 * L_{\{0,1\}} + L_f, \end{aligned} \tag{16}$$

where $I_{gt}, I_{gt}^{\{\frac{1}{8},\frac{1}{4}\}}$ are the full-resolution and down-scaled ground truth views.

6 Experiment

6.1 Implementation Details

Initially, we train the deep generative degradation simulator for 150K iterations. After this, we freeze the weights of both the deep generative degradation simulator and the optical flow model used in G-IVM for the next 300K iterations. Then we jointly train both the deep generative degradation simulator and G-IVM for additional 300K iterations, using a batch size of 16 and a cropped input size of 128 × 128. We use the same data augmentation techniques as NeRFLiX [83], and employ an Adam optimizer and a Cosine annealing learning rate scheme.

6.2 Datasets and Metrics

Following NeRFLiX, we conduct experiments on three popular datasets: LLFF [38], Tanks and Temples [25], and Noisy LLFF Synthetic [39]. The first two benchmarks have eight and five real-world scenes, respectively. Noisy LLFF Synthetic has eight virtual scenes, where we manually apply camera jetting to the precise camera poses to simulate the imperfect in-the-wild calibration.

We evaluate our method using the PSNR (↑), SSIM [59] (↑) and LPIPS [77](↓) metrics, consistent with the evaluation standards of NeRF models.

6.3 Improvements over SOTA NeRFs

We validate the effectiveness of NeRFLiX++ by consistently improving the performance of state-of-the-art NeRF models on diverse datasets. Furthermore, we conduct thorough quantitative and qualitative comparisons between NeRFLiX++ and NeRFLiX, while also assessing their respective inference efficiency.

LLFF. In order to examine the enhancement potential of our NeRFLiX++, we investigate six representative models, including NeRF [39], TensoRF [9], Plenoxels [17], NeRF-mm [61], NLF [1], and RegNeRF [41]. Using rendered views (as well as their reference views) of NeRF approaches as inputs to our model, we aim to further improve the synthesis quality. The quantitative results are provided in Table 1. Under both protocols, NeRFLiX++ exhibits comparable improvements compared to NeRFLiX, elevating the performance of NeRF models to unprecedented levels. For instance, NeRFLiX++ achieves significant improvements of *0.61dB/0.025/0.054* in terms of PSNR/SSIM/LPIPS for the Plenoxels [17] dataset. Notably, NeRFLiX++ demonstrates *2.4× smaller* model capacity and *9.2× faster* processing speed for 1024 × 1024 images compared to NeRFLiX, as shown in Table 1c. Furthermore, NeRFLiX++ achieves remarkable efficiency by processing ultra high-resolution frames of 2048 × 2048 in just 1.5 seconds.

Tanks and Temples. Compared with the LLFF, it has large variations of camera viewpoints. As a result, even recent advanced NeRF models, *e.g.*, TensoRF [9] and DIVeR [62], fail to synthesize high-quality results. As depicted in Table 2a, both NeRFLiX and NeRFLiX++ demonstrate substantial performance improvements across these models. Particularly, NeRFLiX++ exhibits enhanced generalization capabilities, resulting in more significant performance gains. For example, NeRFLiX++ achieves notable improvements of *0.81dB/0.017/0.035* on PSNR/SSIM/LPIPS for the TensoRF [9] model.

Noisy LLFF Synthetic. Apart from in-the-wild benchmarks above, we also demonstrate the enhancement capability of our model on noisy LLFF Synthetic. From the results shown in Table 2b, we see that our NeRFLiX++ yields substantial improvements upon two SOTA NeRF models.

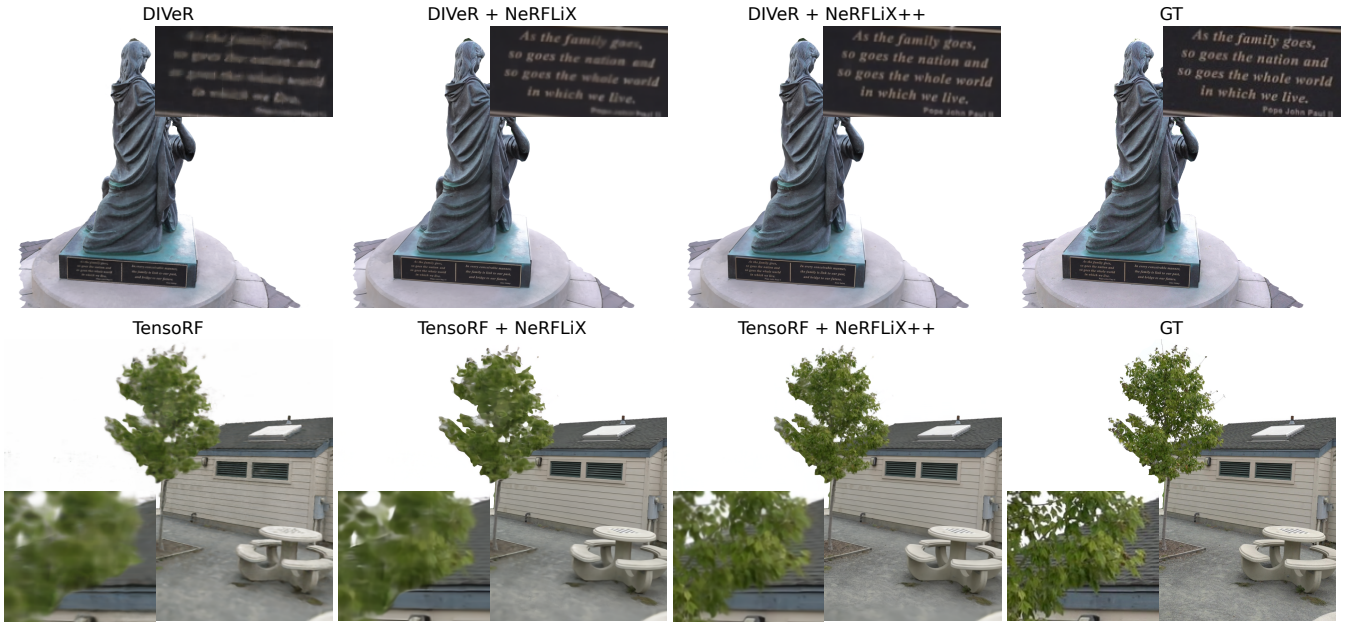


Fig. 10: Qualitative results of NeRFLiX and NeRFLiX++. It is observed that NeRFLiX++ is able to restore richer photo-realistic details than NeRFLiX, showing the superior performance of NeRFLiX++.

Qualitative results. Fig. 10 presents qualitative examples for visual assessment. The results demonstrate that NeRFLiX++ effectively restores clearer image details while significantly reducing NeRF-style artifacts in the rendered images, highlighting the efficacy of our approach.

6.4 Training Acceleration for NeRF Models

In this section, we show how NeRFLiX(++) makes it possible for NeRF models to produce better results even with a 50% reduction in training time. To be more precise, we make use of NeRFLiX and NeRFLiX++ to improve the rendered images of two SOTA NeRF models after training them with half the training period specified in the publications. The enhanced results *outperform* the counterparts with full-time training, as shown in Table 2c. Notably, both NeRFLiX and NeRFLiX++ have reduced the training period for Plenoxels [17] from 24 minutes to 10 minutes while also consistently improving the quality of the rendered images.

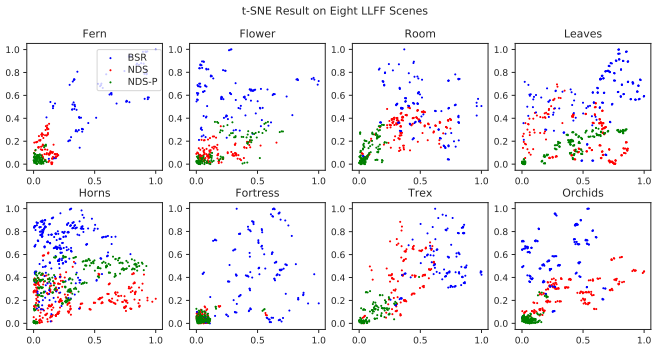


Fig. 11: Quantitative comparison of three degradation models over eight LLFF scenes. We draw the normalized differences between the simulated images of the three degradation methods and the real NeRF-rendered images. Better results are achieved with smaller values. “BSR”, “NDS”, “NDS-P” refer to the degradation models of BSR [74], NeRFLix and NeRFLiX++.

6.5 Ablation Study

In this section, we conduct comprehensive experiments on LLFF [38] under the LLFF-P1 protocol to analyze each of our designs. We use TensoRF [9] as our baseline[3].

Data Simulation quality. We begin by evaluating the simulation quality of our two-stage degradation modeling approach. To assess this, we perform a statistical analysis to measure the similarity between a set of real-rendered frames and multiple sets of simulated frames generated using different degradation techniques. These techniques include BSR [74], the manually designed simulator of NeRFLiX, and our newly proposed NeRFLiX++. To visually represent the comparison, we employ t-SNE [52] to visualize deep image features extracted using Inception-v3 [48]. The results are presented in Fig. 11. Our two-stage simulator produces simulated data that is statistically closest to the real rendered images, outperforming both NeRFLiX and BSR. This finding is further supported by the quantitative analysis provided in Table 3, which demonstrates that our two-stage simulator achieves superior results in terms of PSNR/SSIM compared to NeRFLiX and BSR.

Two-stage degradation modeling. In addition to the previous

3. The TensoRF results (26.70dB/0.838/0.204) that we examined slightly differ from the published results (26.73dB/0.839/0.204).

Method	PSNR (dB)↑	SSIM↑	LPIPS↓
TensoRF [9] (ECCV'22)	28.43	0.920	0.142
TensoRF [9] + NeRFLiX	28.94 (↑ 0.51)	0.930	0.120
TensoRF [9] + NeRFLiX++	29.24 (↑ 0.81)	0.937	0.107
DIVeR [62] (CVPR'22)	28.16	0.913	0.145
DIVeR [62] + NeRFLiX	28.61 (↑ 0.45)	0.924	0.127
DIVeR [62] + NeRFLiX++	28.85 (↑ 0.69)	0.933	0.111

(a) Improvements over TensoRF and DIVeR on Tanks and Temples. Best and second best results are highlighted in red and blue.

Method	PSNR (dB)↑	SSIM↑	LPIPS↓
TensoRF [9] (ECCV'22)	22.83	0.881	0.147
TensoRF [9] + NeRFLiX	24.12 (↑ 1.29)	0.913	0.092
TensoRF [9] + NeRFLiX++	25.39 (↑ 2.56)	0.926	0.085
Plenoxels [17] (CVPR'22)	23.69	0.882	0.127
Plenoxels [17] + NeRFLiX	25.51 (↑ 1.82)	0.920	0.084
TensoRF [9] + NeRFLiX++	26.82 (↑ 3.22)	0.930	0.080

(b) Improvements over TensoRF and Plenoxels on noisy LLFF Synthetic.

Method	PSNR (dB)↑/SSIM↑/LPIPS↓
TensoRF [9](4 hours)	26.73 / 0.839 / 0.204
TensoRF [9](**2 hours**)	26.18 / 0.819 / 0.230
[9](**2 hours**) + NeRFLiX	27.14 / 0.858 / 0.165
[9](**2 hours**) + NeRFLiX++	27.15 / 0.861 / 0.169
Plenoxels [17](24 minutes)	26.29 / 0.839 / 0.210
Plenoxels [17](**10 minutes**)	25.73 / 0.804 / 0.252
[17](**10 minutes**) + NeRFLiX	26.60 / 0.847 / 0.181
[17](**10 minutes**) + NeRFLiX++	26.57 / 0.849 / 0.181

(c) Improvements over TensoRF and Plenoxels trained with half of the recommended iterations on LLFF [38] under LLFF-P1.

TABLE 2: Quantitative evaluation of improvements of NeRFLiX and NeRFLiX++ for various NeRFs.

Metrics	Baseline	D.Models	SwIR	DATSR	EDVR	VST
PSNR	26.70	BSR	26.20↓	25.99↓	26.01↓	25.19↓
		NDS	26.82↑	26.84↑	26.88↑	26.79↑
		NDS-P	26.85↑	26.90↑	26.98↑	26.94↑
SSIM	0.838	BSR	0.834↓	0.826↓	0.819↓	0.705↓
		NDS	0.845↑	0.843↑	0.847↑	0.842↑
		NDS-P	0.847↑	0.847↑	0.850↑	0.849↑

TABLE 3: Quantitative results of utilizing different degradations in the existing image and video processing models including SwIR [32], DATSR [4], EDVR [56] and VSR [35]. We re-train these four models on three simulated datasets produced by BSR [74], NDS [83] and our proposed NDS-P. ↑/↓ indicate the model achieves better/worse performance compared with baseline (TensoRF).

findings that showcase the effectiveness of our two-stage degradation modeling, we conduct a detailed analysis to evaluate the individual components of the degradations, namely the human-crafted simulator and the deep generative simulator. To perform this evaluation, we train four models using different combinations of these degradations. The results, presented in Table 4, highlight the importance of both simulators, emphasizing their necessity in achieving desirable outcomes.

Weighted top-K similarity loss (WKS). We evaluate the

Models	Base	I-J	L-C	S2	PSNR(dB)	SSIM
Model-1	✓				26.92	0.850
Model-2	✓	✓			27.03	0.853
Model-3	✓	✓	✓		27.14	0.858
Model-4	✓	✓	✓	✓	**27.38**	**0.866**

TABLE 4: Performances of different degradations used in our two-stage degradation simulator. "Base" signifies NDS, the manual degradation simulator in NeRFLiX, "I-J" and "L-C" are shorted for illumination jetting and lightness compression schemes, and "S1" refers to our deep generative degradation simulator.

Supervision	L1	$K=1$	$K=3$	$K=5$	$K=10$	$K=20$
PSNR (dB)	27.03	27.21	27.35	27.38	27.38	27.37
SSIM	0.851	0.859	0.865	0.866	0.866	0.866

TABLE 5: Impact of different similarity patch numbers (K) for WKS. Moreover, we also include another model trained with L1 loss to validate the effectiveness of our WKS supervision.

performance of our proposed WKS. For comparison, we train an additional G-IVM model using the conventional L1 loss as supervision. The results in Table 5 demonstrate that this model achieves significantly inferior performance compared to the models trained with our proposed WKS. This outcome emphasizes the effectiveness of WKS for deep degradation training. Furthermore, we investigate the influence of different numbers (K) of similar patches in WKS supervision. We train four additional G-IVM models. As indicated in Table 5, we observe progressive improvements in PSNR values as the number of similar patches increases from $K=1$ to $K=5$, after which the improvements saturate. This behavior is expected since image patches with relatively small similarities contribute less to the overall performance.

Pyramid fusion in G-IVM. In order to utilize multi-scale contextual information from inter-viewpoint frames, we introduce a pyramid-guided aggregation structure to enhance the quality of rendered frames. Table 6a demonstrates that incorporating more aggregation levels consistently improves the final performance. Specifically, our complete model (Model-C) achieves the highest PSNR/SSIM scores, indicating its superiority over other models.

Flow guidance in G-IVM. To address distinct viewpoint changes in high-resolution frames, we introduce the utilization of coarse optical flow for guiding the aggregation process. In order to assess the significance of this strategy, we train an additional model referred to as "NG-IVM" under the same experimental setup, but without utilizing optical flow guidance. The results presented in Table 6b clearly indicate that our guided inter-viewpoint mixer outperforms the NG-IVM model by a substantial margin, highlighting the effectiveness of our design.

Aligning targets in G-IVM. In the l-th level fusion, we deviate from existing approaches [7], [27], [56], [82] that treat F^l as the target feature. Instead, apart from the first-level alignment ($l=0$), we propose using the previously aggregated feature $\hat{F}^{l-1}(l>0)$. To validate the effectiveness of this design choice, we compare these two strategies, and the results are presented in

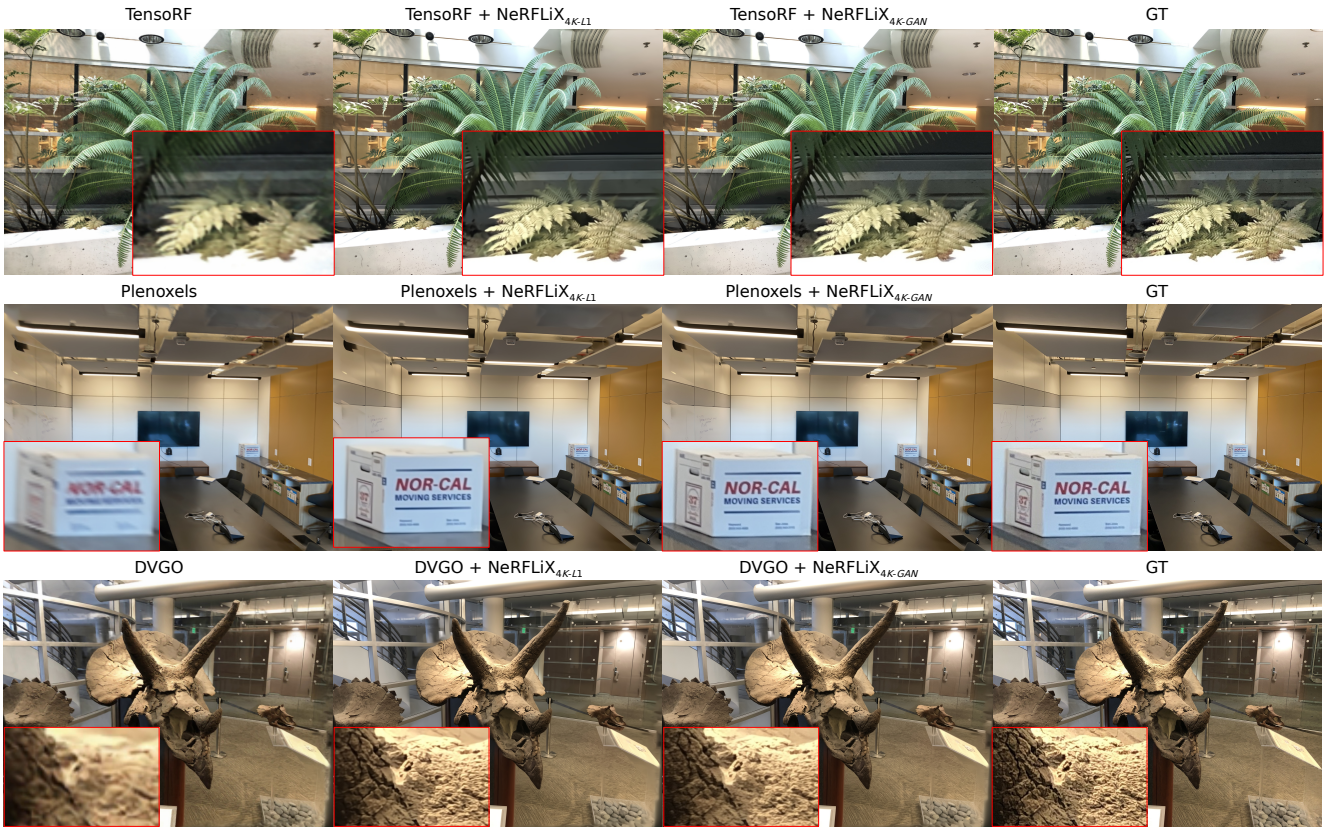


Fig. 12: Qualitative results of restoring 4K images from noisy 1K frames produced by TensoRF, Plenoxels, DVGO. It is clear that NeRFLix++$_{4K}$ is capable of fusing high-quality reference views to generate natural image textures.

Models	L_1	L_2	L_3	PSNR(dB)	SSIM
Model-A	✓			26.90	0.849
Model-B	✓	✓		27.11	0.856
Model-C	✓	✓	✓	**27.38**	**0.866**

(a) Impact of pyramid fusion. $L_{\{1,2,3\}}$ refers to different levels.

Setting	PSNR	SSIM
w/o Flow	27.21	0.860
w/ Flow	**27.38**	**0.866**

(b) Impact of flow guidance.

Target	PSNR	SSIM
F^l	26.97	0.853
$\hat{F}^{l-1}$	**27.38**	**0.866**

(c) Impact of aligning targets.

TABLE 6: Experimental analyses to understand the roles of pyramid fusion, flow guidance and different aligning targets and present the quantitative results.

Table 6c. Our fusion strategy outperforms the existing solution in terms of PSNR, SSIM, and LPIPS, indicating that our design is better suited for NeRF-agnostic restoration tasks.

Correspondence estimation sizes. In Sec. 1, we discuss the potential benefits of utilizing coarse correspondence estimation at a low resolution (downscaled by ×4). In this section, we investigate the impact of building inter-viewpoint correspondences at different sizes ($\times 1, \times 2, \times 8, \times 16$). To this end, in addition to the default settings (referred to as "NeRFLiX++$_{\times 4}$"), we train four additional NeRFLiX++ models (NeRFLiX++$_{\times 1,\times 2,\times 8,\times 16}$) and quantitatively evaluate their accuracy to support our claim. As shown in Table 7, NeRFLiX++$_{\times 4}$ achieves the best results in terms of PSNR and SSIM. It is worth noting that larger downscaling ratios (*e.g.*, ×16) may hinder performance due to their inability to provide accurate guidance for high-resolution aggregations.

Scales	×1	×2	×4	×8	×16
PSNR (dB)	27.25	27.28	**27.38**	27.21	27.17
SSIM	0.863	0.863	**0.866**	0.862	0.860

TABLE 7: Impacts of performing coarse correspondence estimation at different image sizes.

7 NeRFLiX++ for 4K images

In addition to the common challenges encountered in low-resolution novel view rendering, such as artifacts and blurriness, rendering high-resolution images, *i.e.*, 4K resolution, using existing NeRF models poses significant computational resource requirements. Even with highly optimized data structures like

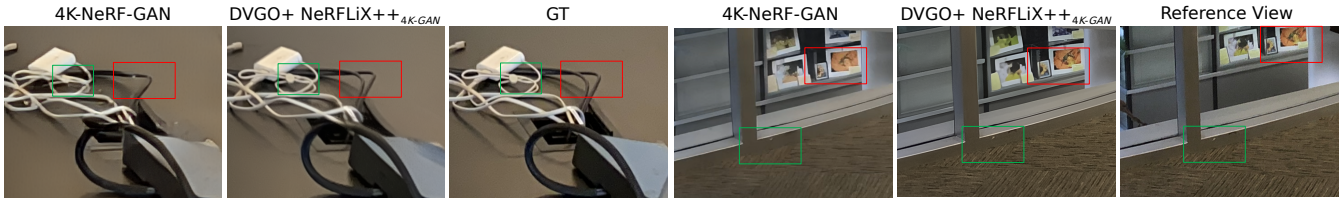


Fig. 13: Visual comparisons of 4K-NeRF and our NeRFLiX++$_{4K-GAN}$. While 4K-NeRF-GAN generates incorrect structures, our approach produces photo-realistic image details that closely resemble the ground truth or reference views (for novel viewpoints).

Baseline	M1	M2	M3	M4	M5	PSNR↑	SSIM↑	LPIPS↓
TensoRF	✓					25.09	0.873	0.43
TensoRF		✓				25.54	0.886	0.42
TensoRF			✓			25.54	0.885	0.40
TensoRF				✓		**25.89**	**0.896**	0.35
TensoRF					✓	25.45	0.885	**0.18**
Baseline	M1	M2	M3	M4	M5	PSNR↑	SSIM↑	LPIPS↓
Plenoxels	✓					24.81	0.873	0.47
Plenoxels		✓				25.22	0.885	0.42
Plenoxels			✓			25.25	0.884	0.42
Plenoxels				✓		**25.50**	**0.895**	0.35
Plenoxels					✓	25.84	0.889	**0.19**
Baseline	M1	M2	M3	M4	M5	PSNR↑	SSIM↑	LPIPS↓
DVGO	✓					25.20	0.869	0.50
DVGO		✓				25.86	0.888	0.43
DVGO			✓			25.86	0.888	0.42
DVGO				✓		**26.21**	**0.900**	0.35
DVGO					✓	25.84	0.889	**0.19**

TABLE 8: Quantitative performance of several up-scaling, image and video restoration methods, including bicubic interpolation (referred to as "Bi"), NeRFLiX ∘ Bi, NeRFLiX++ ∘ Bi, NeRFLiX++$_{4K-L1}$ (supervised by L1), and NeRFLiX++$_{4K-GAN}$ (with an adversarial loss), to enhance the results of three representative NeRF baselines (TensoRF [9], Plenoxels [17], and DVGO [47]). Here, the symbol ∘ denotes that the two methods were performed sequentially. We denote the results of these methods as "M1-5" in our analysis.

tensor decomposition employed in TensoRF [9], training TensoRF models for 2K and 4K images on an NVIDIA RTX 3090 remains impractical due to limitations in GPU memory.

In this section, we investigate the potential of utilizing NeRFLiX++ to super-resolve and enhance low-resolution images generated by different NeRF models, thereby producing high-quality 4K results. We first define the problem and then discuss the modifications made to the G-IVM model. Subsequently, we perform quantitative and qualitative analyses to assess the effectiveness of NeRFLiX++ for 4K images.

7.1 Problem Formulation

Given a low-resolution (1K) target frame I generated by NeRF models and its two 4K reference views $\{I_1, I_2\}$, NeRFLiX++$_{4K}$ aims to restore a 4K output with photo-realistic details.

7.2 Framework

The framework of our proposed G-IVM is depicted in Fig. 9. To enable the restoration of 4K images, we make minimal modifications to the G-IVM framework.

Encoder. To accommodate the resolution difference between the input frame $I \in \mathcal{R}^{H\times W}$ and its reference views $\{I_1, I_2\} \in \mathcal{R}^{4H\times 4W}$, we introduce the following adjustments. For encoder-1, we incorporate two convolutional layers with a stride of 2, resulting in down-sized reference features $F^{\{\frac{1}{4},\frac{1}{2},1\}}_{\{1,2\}} \in \mathcal{R}^{\{H\times W, 2H\times 2W, 4H\times 4W\}}$. Encoder-2 does not involve any down-sampling. Consequently, the two lowest-resolution reference features $F^{\frac{1}{4}}_{\{1,2\}} \in \mathcal{R}^{H\times W}$ match the spatial resolution of the input feature $F \in \mathcal{R}^{H\times W}$.

We train two NeRFLiX++$_{4K}$ models, one using L1 loss (NeRFLiX++$_{4K-L1}$) and the other using a combination of L1 and GAN losses (NeRFLiX++$_{4K-GAN}$).

Implementation details. Compared to the original NeRFLiX++, when utilizing a 1K rendered frame as input, we substitute the two 1K reference frames with their 4K counterparts (obtained from LLFF-T[4]), while keeping other training details unchanged.

For samples obtained from the Vimeo dataset, we initially down-sample the input frame by a factor of ×4, thereby establishing the same setup as LLFF-T. In other words, this configuration involves a low-resolution input view and two high-resolution reference views.

7.3 Improvements over NeRFs for 4K Images

To evaluate the effectiveness of NeRFLiX++$_{4K}$, we conduct experiments using different restoration methods (M1-M5) to generate 4K images from low-resolution inputs produced by three state-of-the-art NeRF models: TensoRF [9], Plenoxels [17], and DVGO [47]. The results in Table 8 demonstrate that all models involving NeRFLiX and NeRFLiX++ (M2-M5) outperform simple bicubic up-sampling (M1), indicating the restoration capability of NeRFLiX and NeRFLiX++. Particularly, NeRFLiX++$_{4K-L1}$ and NeRFLiX++$_{4K-GAN}$ achieve the best performance in terms of PSNR, SSIM, and LPIPS. Furthermore, Fig. 12 visually demonstrates that NeRFLiX++$_{4K-L1}$ produces high-quality 4K frames with clearer textures and reduced rendering artifacts. Meanwhile, NeRFLiX++$_{4K-GAN}$ generates more high-frequency details and sharper edges, resulting in visually appealing results.

4. LLFF-T provides images at a 4K resolution

Method	4K-NeRF$_{L1}$	NeRFLiX++$_{4K-L1}$	4K-NeRF$_{GAN}$	NeRFLiX++$_{4K-GAN}$
PSNR	25.44	**26.21**	24.71	25.84
SSIM	0.883	**0.900**	0.871	0.889
LPIPS	0.41	0.35	0.24	**0.19**

(a) Quantitative comparisons between 4K-NeRF [60] and our method on 4K NeRF-rendered images.

Method	NeRF-SR	NeRF-SR + NeRFLiX++$_{4K-L1}$
PSNR (dB)	27.21	**29.19**
SSIM	0.852	**0.908**
LPIPS	0.09	**0.06**

(b) Quantitative improvements over NeRF-SR.

TABLE 9: Quantitative evaluation of NeRFLiX++$_{4K}$ by comparing it with 4K-NeRF [60] and NeRF-SR [53]. To differentiate between models trained with L1 and GAN supervision, we used subscripts $_{L1}$ and $_{GAN}$, respectively.

Comparison to 4K-NeRF. We also compare NeRFLiX++$_{4K}$ with 4K-NeRF [60][5]. The results in Table 9a show that NeRFLiX++$_{4K-L1}$ achieves a significant improvement of 0.77dB in PSNR over 4K-NeRF$_{L1}$. Furthermore, NeRFLiX++$_{4K-GAN}$ surpasses 4K-NeRF in terms of perceptual quality. Fig. 13 visually demonstrates that 4K-NeRF$_{GAN}$ fails to reconstruct subtle image structures, while NeRFLiX++$_{4K-GAN}$ effectively restores natural image contents from noisy 1K photos, resulting in superior visual enhancement results.

Comparison to NeRF-SR. We also compare NeRFLiX++$_{4K-L1}$ with NeRF-SR [53]. NeRF-SR is a two-stage novel view synthesis approach. In the first stage, they propose a super-sampling NeRF model to generate super-resolved novel views from low-resolution training photos. Then, they utilize a refinement module to enhance the first-stage results. To ensure a fair comparison, we utilize our NeRFLiX++$_{4K-L1}$ model to enhance their first-stage results and quantitatively compare them with their refined results. Table 9b suggests that NeRFLiX++$_{4K-L1}$ significantly outperforms NeRF-SR, highlighting the effectiveness of our method.

In addition to its superior performance, unlike 4K-NeRF and NeRF-SR models that require re-training for new scenes, NeRFLiX++$_{4K}$ offers the advantage of being NeRF-agnostic and scene-agnostic. This characteristic allows for quick and efficient deployment of NeRFLiX++4K in various scenarios.

Comparison to existing image and video restorers. Additionally, we compare our NeRFLiX++$_{4K}$ model with state-of-the-art image and video super-resolution approaches, such as SwIR [32], RealESRGAN [57], and RealBasicVSR [8]. Using TensoRF [9] as the baseline, we utilize these models to generate enhanced high-resolution images and present the detailed results in Table 10. Although these models produce promising restoration outcomes for generally real-world images, they all exhibit inferior performance than our NeRFLiX++$_{4K}$ model, which manifests NeRFLiX++'s excellent restoration capability for NeRF-rendered photos.

5. For a fair comparison, we report the enhanced results using DVGO [47] as our baseline, as 4K-NeRF also employs DVGO as its baseline.

Method	SwIR	RealESRGAN	RealBasicVSR	NeRFLiX++$_{4K}$
PSNR (dB)	23.91	24.22	23.97	**25.45**
SSIM	0.847	0.860	0.851	**0.885**
LPIPS	0.300	0.299	0.311	**0.184**

TABLE 10: To ensure a fair and objective evaluation, we compare NeRFLiX++$_{4K}$ against various representative general image and video restoration models that are trained using adversarial loss. We assess their abilities of up-sampling and enhancing the rendered views (at a low resolution) of TensoRF [9].

4K video demo. We have prepared a video demonstration showcasing the enhancement capabilities of our proposed NeRFLiX++ method, which can be viewed at https://www.youtube.com/watch?v=YiXvgQXiWII[6]. It comprises three parts. And the first two segments of the video highlight that while TensoRF and Plenoxels struggle to generate satisfactory 1K novel views, NeRFLiX++ is capable of restoring ultra-high-resolution outputs from these low-resolution noisy views. Notably, NeRFLiX++ even recovers recognizable characters and sharper textures at 4K resolutions. The last segment shows NeRFLiX++ can be used to significantly improve the visual quality of various NeRF models (*i.e.* TensoRF [9], Plenoxels [17], RegNeRF [41], NLF [1], DIVeR [62], NeRF-mm [61], etc.).

8 Conclusion

We introduce NeRFLiX, a general NeRF-agnostic paradigm for high-quality restoration of neural view synthesis. We systematically analyze the NeRF rendering pipeline and introduce the concept of NeRF-style degradations. Towards eliminating NeRF-style artifacts, we present a novel NeRF-style degradation simulator and construct a large-scale simulated dataset. Through training state-of-the-art deep neural networks on the simulated dataset, we successfully remove NeRF artifacts. Additionally, we propose an inter-viewpoint mixer to restore missing details in NeRF-rendered frames by aggregating multi-view frames. Extensive experiments validate the effectiveness of NeRFLiX.

To further enhance the restoration capability and inference efficiency of NeRFLiX, we present NeRFLiX++. It improves upon NeRFLiX by incorporating better degradation modeling and faster inter-viewpoint aggregation techniques. NeRFLiX++ enables realistic 4K view synthesis ability and achieves superior quantitative and qualitative performance, as demonstrated in our extensive experiments.

Appendix

Raw data collection. We collect raw sequences from Vimeo90K [66] and LLFF-T [38]. In total, Vimeo90K contains 64612 7-frame training clips with a 448×256 resolution. Three frames (two reference views and one target view) are selected from a raw sequence of Vimeo90K in a random order. Apart from the inherent displacements within the selected views, we add random global offsets to the two reference views, largely

6. We recommend watching it in 4K resolution for the best view.

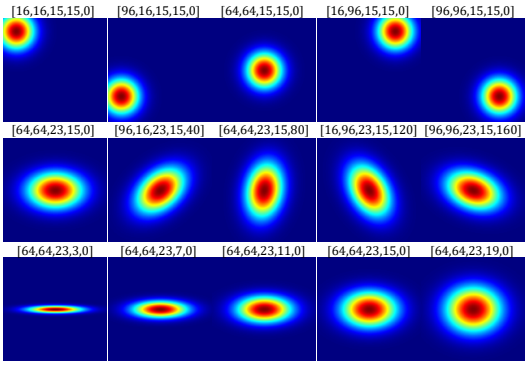


Fig. 14: We give some visualized region-adaptive masks. The parameters refer to the values of $[c_i, c_j; \sigma_i, \sigma_j, A]$ in Eq. (3).

Settings	10%	50%	100%	PSNR (dB)	SSIM
LLFF-T				26.28	0.837
LLFF-T+	✓			26.71	0.840
LLFF-T+		✓		27.08	0.856
LLFF-T+			✓	**27.39**	**0.867**
TensoRF (Base)	-	-	-	26.70	0.838

TABLE 11: Quantitative results of different training data sizes. First, we train an IVM model only using the LLFF-T. Then, we gradually increase the simulated pairs (10%, 50%, 100%) from Vimeo90K [66] to train another three IVM models.

enriching the variety of inter-viewpoint changes. On the other hand, we also use the training split of the LLFF dataset, which consists of 8 different forward-facing scenes with 20-62 high-quality input views. Following previous work, we drop the eighth view and use it for evaluation. To construct a training pair from LLFF-T, we randomly select a frame as the target view and then use the proposed view selection algorithm (Sec. 4.3) to choose two reference views that are most overlapped with the target view.

Hyper-parameter setup. In Eq. (1), the 2D Gaussian noise map n is generated with a zero mean and a standard deviation ranging from 0.01 to 0.05. The isotropic blur kernel g has a size of 5×5. We employ a Gaussian blur kernel to produce blurry contents by randomly selecting kernel sizes (3-7), angles (0-180), and standard deviations (0.2-1.2). Last, in order to obtain a region-adaptive blending map M in Eq. (3), we use random means ($c_i, c_j \in (-16, 144)$), standard deviations ($\sigma_i \in (13, 25), \sigma_j \in (0, 24)$), and orientation angles ($A \in (0, 180)$). Additionally, we visualize some generated masks using different hyper-parameter combinations ($[c_i, c_j; \sigma_i, \sigma_j, A]$) in Fig. 14.

Training data size. We investigate the influence of training data size. Under the same training and testing setups, we train several models using different training data sizes. As illustrated in Table 11, we can observe that the final performance is positively correlated with the number of training pairs. Also, we notice the IVM trained with only LLFF-T data or additional few simulated pairs (10% of the Vimeo90K) fails to enhance the TensoRF-rendered results, *i.e.*, there is no obvious improvement compared to TensoRF [9]. This experiment demonstrates the importance of sizable training pairs for training a NeRF restorer.

In Sec. 4.2, we briefly describe the framework architecture of our inter-viewpoint mixer (IVM). Here we provide more details. As illustrated in Fig. 15a, there are two convolutional modules ("Encoder 1/2") to extract features of the degraded view I and its two reference views $\{I^r_1, I^r_2\}$, respectively. Then, we develop a hybrid recurrent aggregation module that iteratively performs pixel-wise and patch-wise fusion. At last, a reconstruction module is implemented by a sequence of residual blocks (40 blocks) to output the enhanced view $\hat{I}$. The default channel size is 128.

Feature extraction. Given a rendered view I and its two reference views $I^r_{1,2}$, we aim to utilize the two encoders to extract deep image features $\mathbf{f}$ and $\mathbf{f}^r_{1,2}$, respectively. As detailed in Fig. 15a, the two encoders share an identical structure. A convolutional layer is first adopted to convert an RGB frame to a high-dimensional feature. Then we further extract the deep image feature using 5 stacked residual blocks followed by another convolution layer.

Hybrid recurrent aggregation. As depicted in Fig. 15a, we employ three hybrid recurrent aggregation blocks (termed "Hybrid-R1(2,3)") to progressively fuse the inter-viewpoint information from the image features ($\mathbf{f}$ and $\mathbf{f}^r_{\{1,2\}}$). Next, we take the first iteration as an example to illustrate our aggregation scheme.

Pixel-wise aggregation. As shown in Fig. 15b, we first merge the target view feature $\mathbf{f}$ and one of the reference features $\mathbf{f}^r_{\{1,2\}}$ by channel concatenation. Then we use a convolutional layer to reduce the channel dimension and five residual blocks followed by another convolutional layer to obtain a fused deep feature. Later on, the fused feature and the reference feature are further aggregated through a deformable convolution. And the other reference image follows the same processing pipeline. In this case, we finally obtain two features after the pixel-wise aggregation.

Patch-wise aggregation. We adopt a window-based attention mechanism [35] to accomplish patch-wise aggregation. In detail, the pixel-wisely fused features are first divided into several 3D slices through a 3D patch partition layer. Then, we obtain 3D tokens via a linear embedding operation and aggregate patch-wise information using a video Swin transformer block. Finally, 3D patches are regrouped into a 3D feature map.

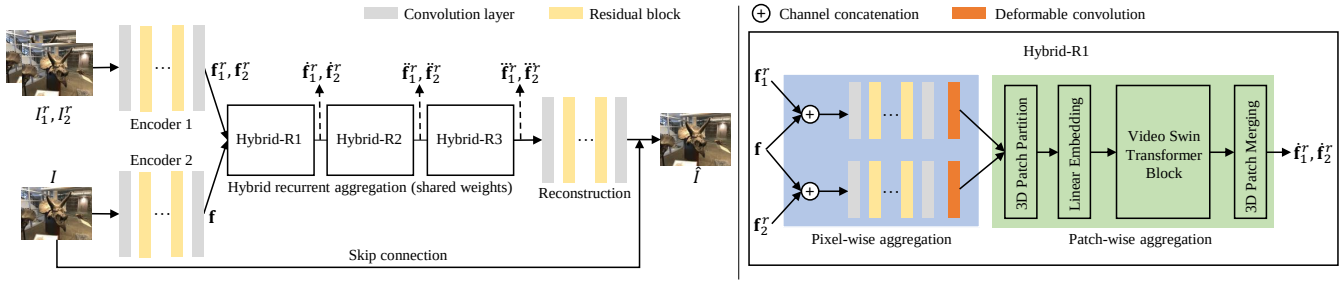
In the next iteration, we split the 3D feature map into two "reference" features $\dot{\mathbf{f}}^r_{\{1,2\}}$ and repeat the pixel-wise and patch-wise aggregation. Note that, the weights of pixel-wise and patch-wise modules are shared across all iterations to reduce the model complexity.

References

[1] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 8, 13

[2] 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*, pages 5855–5864, 2021. 2

[3] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv*, 2021. 3



(a) IVM (b) Detailed structures of pixel-wise and patch-wise aggregation

Fig. 15: The detailed framework architecture of our proposed IVM.

[4] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European conference on computer vision*, 2022. 10

[5] Mingden Cao, Yanbo Fan, Yong Zhang, Jue Wang, and Yujiu Yang. Vdtr: Video deblurring with transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 3

[6] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3, 4

[7] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022. 3, 10

[8] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 13

[9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 7, 8, 9, 10, 12, 13, 14

[10] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2

[11] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. *arXiv preprint arXiv:2208.00277*, 2022. 3

[12] Forrester Cole, Kyle Genova, Avneesh Sud, Daniel Vlasic, and Zhoutong Zhang. Differentiable surface rendering via non-differentiable sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6088–6097, 2021. 2

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 3, 4

[14] Chenxi Lola Deng and Enzo Tartaglione. Compressing explicit voxel grid representations: fast nerfs become also small. *arXiv preprint arXiv:2210.12782*, 2022. 2

[15] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[16] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3

[17] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 3, 7, 8, 9, 10, 12, 13

[18] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 3

[19] J-M Geusebroek, Arnold WM Smeulders, and Joost Van De Weijer. Fast anisotropic gauss filtering. *IEEE transactions on image processing*, 12(8):938–943, 2003. 4

[20] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18409–18418, 2022. 1, 2

[21] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. Efficientnerf: Efficient neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12902–12911, 2022. 2

[22] Jeffrey Ichnowski*, Yahav Avigal*, Justin Kerr, and Ken Goldberg. Dex-NeRF: Using a neural radiance field to grasp transparent objects. In *Conference on Robot Learning (CoRL)*, 2020. 2

[23] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021. 1, 3

[24] 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*, pages 18365–18375, 2022. 2

[25] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 8

[26] Andreas Kurz, Thomas Neff, Zhaoyang Lv, Michael Zollhöfer, and Markus Steinberger. Adanerf: Adaptive sampling for real-time rendering of neural radiance fields. In *European Conference on Computer Vision*, pages 254–270. Springer, 2022. 2

[27] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*, pages 335–351. Springer, 2020. 10

[28] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. *Advances in Neural Information Processing Systems*, 33:20343–20355, 2020. 3

[29] Wenbo Li, Kun Zhou, Lu Qi, Liying Lu, and Jiangbo Lu. Best-buddy gans for highly detailed image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1412–1420, 2022. 1, 3, 6

[30] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3867–3876, 2019. 3

[31] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 3, 4

[32] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 3, 10, 13

[33] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 1

[34] Haotong Lin, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields with learned depth-guided sampling. *arXiv preprint arXiv:2112.01517*, 2021. 2

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 10, 14

[36] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[37] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. 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*, pages 16190–16199, 2022. 2

[38] 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. 4, 5, 7, 8, 9, 10, 13

[39] 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. 2, 7, 8

[40] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 2

[41] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 8, 13

[42] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[43] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4161–4170, 2017. 7

[44] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. Derf: Decomposed radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14153–14161, 2021. 2

[45] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 3

[46] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8269–8279, 2022. 2

[47] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5459–5469, June 2022. 2, 3, 12, 13

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 9

[49] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2

[50] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3

[51] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020. 3

[52] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 9

[53] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6445–6454, 2022. 2, 13

[54] Longguang Wang, Yulan Guo, Zaiping Lin, Xinpu Deng, and Wei An. Learning for video super-resolution through hr optical flow estimation. In *Asian Conference on Computer Vision*, pages 514–529. Springer, 2018. 3, 4

[55] 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*, pages 4690–4699, 2021. 2

[56] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3, 4, 10

[57] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 1, 3, 13

[58] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[59] 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. 8

[60] Zhongshu Wang, Lingzhi Li, Zhen Shen, Li Shen, and Liefeng Bo. 4k-nerf: High fidelity neural radiance fields at ultra high resolutions. *arXiv preprint arXiv:2212.04701*, 2022. 13

[61] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 3, 7, 8, 13

[62] Liwen Wu, Jae Yong Lee, Anand Bhattad, Yu-Xiong Wang, and David Forsyth. Diver: Real-time and accurate neural radiance fields with deterministic integration for volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16200–16209, 2022. 8, 10, 13

[63] Zijin Wu, Xingyi Li, Juewen Peng, Hao Lu, Zhiguo Cao, and Weicai Zhong. Dof-nerf: Depth-of-field meets neural radiance fields. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1718–1729, 2022. 1

[64] Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Hao Su. Neutex: Neural texture mapping for volumetric neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7128, 2021. 2

[65] 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*, pages 5438–5448, 2022. 2

[66] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 127(8):1106–1125, 2019. 3, 4, 13, 14

[67] Guo-Wei Yang, Wen-Yang Zhou, Hao-Yang Peng, Dun Liang, Tai-Jiang Mu, and Shi-Min Hu. Recursive-nerf: An efficient and dynamically growing nerf. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 2

[68] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 3

[69] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 3

[70] Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy.

Path-restore: Learning network path selection for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[71] Songhyun Yu, Bumjun Park, Junwoo Park, and Jechang Jeong. Joint learning of blind video denoising and optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 500–501, 2020. 3, 4

[72] Jiahui Zhang, Fangneng Zhan, Rongliang Wu, Yingchen Yu, Wenqing Zhang, Bai Song, Xiaoqin Zhang, and Shijian Lu. Vmrf: View matching neural radiance fields. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6579–6587, 2022. 3

[73] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020. 3

[74] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1, 3, 4, 9, 10

[75] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 2

[76] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 3

[77] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 8

[78] Wenyuan Zhang, Ruofan Xing, Yunfan Zeng, Yu-Shen Liu, Kanle Shi, and Zhizhong Han. Fast learning radiance fields by shooting much fewer rays. *arXiv preprint arXiv:2208.06821*, 2022. 2

[79] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 1

[80] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18643–18652, 2022. 1

[81] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 3

[82] Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, and Jiangbo Lu. Revisiting temporal alignment for video restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6053–6062, 2022. 3, 10

[83] Kun Zhou, Wenbo Li, Yi Wang, Tao Hu, Nianjuan Jiang, Xiaoguang Han, and Jiangbo Lu. Nerflix: High-quality neural view synthesis by learning a degradation-driven inter-viewpoint mixer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12363–12374, June 2023. 1, 2, 6, 8, 10