Enhancing Multimedia Applications by Removing Dynamic Objects in Neural Radiance Fields

XianBen Yang[1,2], Tao Wang[1,2] *, He Liu[1,2], Yi Jin[1,2], Congyan Lang[1,2], and Yidong Li[1,2]

[1] Key Laboratory of Big Data & Artificial Intelligence in Transportation (Beijing Jiaotong University), Ministry of Education, China
[2] School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China
{yxb_2023,twang,liuhe1996,yjin,cylang,ydli}@bjtu.edu.cn

Abstract. Neural Radiance Fields (NeRF) are at the forefront of view synthesis technology, renowned for their versatility and ease of implementation across various applications. However, their integration into multimedia environments faces challenges: objects occlude background information during motion, which usually compromises the quality of reconstructions. In this paper, we present a novel framework to exclude dynamic interference from NeRF scenes, enhancing the practicability in multimedia applications. Our method leverages perceptual optimization, informed by image quality assessment (IQA), and employs text-guided 2D image inpainting to address view synthesis inaccuracies. Furthermore, we propose a new and challenging dataset of real-world scenes to address the lack of evaluation ground truth for dynamic object inpainting in scenes. Experimental results show that our method significantly outperforms existing methods in terms of appearance metrics for the task of removing dynamic objects from scenes.

Keywords: Novel View Synthesis · NeRF · IQA · Image Inpainting

1 Introduction

Dense scene reconstruction in photorealistic image synthesis has long been a core challenge in computer vision . Traditional *structure from motion* (SfM) methods [36,37] necessitate estimating the camera pose in a pre-processing step. Unfortunately, they often fall short in practical scenarios, particularly in hand-held video settings [19,27]. Recently, the introduction of *neural radiance fields* (NeRF) marks a significant breakthrough in this field [22]. Because of its ability to accurately model complex 3D scenes based on intricate micro-details extracted from 2D images, the current application of NeRF extends beyond its original task of novel view synthesis.

In multimedia applications such as augmented reality [28,24] and the estate sales website [39] where the removal of interference objects is imperative, various NeRF edition methods have been explored. In some pioneer studies [48,51],

* Corresponding author

Fig. 1: Image showcases the ground truth of dynamic objects at the top, the inpainting of the image using LaMa [41] in the middle, and the inpainting using our method at the bottom. Our method removes moving objects and real inpaints scene details, such as laptops or posters.

trained scenes are edited with user-guided transformations that effectively separate objects from the background. More recently, some researchers [47,23] have successfully addressed the task of removing objects from scenes through the use of *artificial intelligence generated content* (AIGC) [41] technology. However, a notable limitation of these approaches lies in their underlying assumption that scene interference remains static. When directly applied to dynamic scenes, these techniques often lead to flawed reconstructions. Furthermore, the images inpainted by their generative models tend to omit crucial scene details, leading to a loss of critical information, as illustrated in Fig. 1. It therefore demands research attention on how to remove dynamic objects from the scene, while generating indoor scenes realistically.

In this paper, we propose a novel neural radiance field framework in response to the challenges posed by dynamic objects in scene reconstructions. Our framework begins with the evaluation of input image sequences through a *convolutional neural network* (CNN) and *contrastive language image pre-training* (CLIP) [29], assigning quality scores based on the image contents. This scoring adaptively selects images with relatively complete scene information across three dimensions: technology, aesthetics and perception, thereby reducing the negative impact of dynamic interference. Subsequently, the dynamic objects in the selected images are automatically detected and removed, and then the empty regions are inpainted using a stable diffusion model [35]. To further improve the quality of the reconstructed scene, we generate some geometric priors, including monocular depth and optical flow, from the RGB images, and integrate them into the optimization of the neural radiance field. In this way, the influence of

interference from moving objects in the scene is remarkably alleviated, as shown in Fig 1.

Our method is firstly evaluated on the TUM dataset [38], a public benchmark that contains dynamic interference in the scene. Secondly, we propose a new dataset to evaluate the removal of moving objects in indoor scenes. This dataset comprises six indoor scenes, each presenting three different scenarios: one with no dynamic object, one with a single dynamic object, and one with multiple dynamic objects. We increase the challenge of scene reconstruction by adding dynamic objects. Our method is validated on the TUM dataset and our proposed dataset, and shows that we outperform existing approaches on standard metrics of image quality. To enhance reproducibility and facilitate further research, we have opened our new dataset is publicly available: https://drive.google.com/file/d/1z8BBco8HziwwcDsPdq4kVBtfLldqMjwJ/view?usp=sharing.

In summary, our work contribution in three-fold:

- We propose a method specifically designed to address the inpaint of new view synthesis in scenes with dynamic elements;
- We introduce a novel view selection mechanism utilizing CLIP for optimizing views;
- We present a new dataset for evaluating the removal of dynamic elements and inpainting in diverse scenes.

2 Related work

2.1 Scene Object Removal

Reconstructing scenes from images of dynamic objects presents significant challenges, mainly when local details exhibit subtle differences. The prevalent approach employs COLMAP [36,37] for multi-view 3D reconstruction, encompassing image calibration and depth image rendering. However, reconstruction efforts can falter in scenarios characterized by insufficient overlap, limited texture, or forward camera motion—conditions that render the scene non-calibrable and, thus, incompatible with NeRF-based methodologies.

To address the challenge of reconstruction failure attributed to dynamic objects, *image quality assessment* (IQA) techniques [42,4] is instrumental. IQA technology facilitates the automatic evaluation and identification of dynamic objects contributing to image blurriness and ghosting. Furthermore, the CLIP model [29], renowned for its general perceptual abilities derived from extensive image and text pre-training, offers another dimension to IQA. By integrating the CLIP model [44] with NIMA [42], we harness both textual and visual cues for a more comprehensive image quality evaluation. This hybrid approach aims to minimize the impact of dynamic elements on image quality assessment, thus enhancing the robustness of the evaluation process.

The presence of dynamic objects can obscure scene information, adversely affecting the quality of the final 3D reconstruction. Image inpainting [35,41] emerges as a potent strategy for eliminating dynamic objects and supplementing

the elimination of regions that create holes. Traditional inpainting approaches [6,26] leverage static information from adjacent frames to restore segments obscured by dynamic objects, thereby enabling the integration of newly synthesized frames as keyframes in the reconstruction process. However, these methods are constrained by their reliance on the contextual data from preceding and succeeding frames.

Recent advancements have introduced the use of masks [18] or Fourier convolutions [41] to address missing regions within a single view caused by the removal of dynamic objects. Such techniques have shown promise in the restoration of static scenes. Nonetheless, ensuring temporal consistency across frames remains a challenge when employing masks or Fourier convolutions, leading to potential discrepancies in consecutive frame inpaint.

In response to these limitations, we propose the adoption of a diffusion model [35] guided by textual descriptions for 3D generation. This approach aims to diminish ambiguity and enhance the precision of reconstructions by utilizing textual cues to guide the inpainting process. By integrating text guidance, our method seeks to improve coherence and accuracy across consecutive frames, mitigating the adverse effects of inconsistencies in frame inpaint.

2.2 Novel View Synthesis

In recent developments, Mildenhall et al. [22] have markedly propelled the field of dense surface reconstruction forward by adopting NeRF. The utility of NeRF in temporal optimization techniques is highlighted by its ability to refine methodologies within limited evaluation periods, showcasing robust performance in scenarios marked by variable camera exposure.

Although numerous systems [20,32,45] have been devised for such unbounded scenarios, many require omnidirectional input, or specific drone or satellite imagery, often benchmarked against real-world monocular videos. Current methodologies [5,3] typically employ strategies that compress the background into a compact representation, optimize an environmental map to depict the background, and rely on multi-view inputs for multiple observations. Conversely, some techniques [12,17] aim to simultaneously and progressively estimate both the camera pose and the radiance field. To surmount specific NeRF limitations, the integration of depth constraints into the training of neural brightness fields has been proposed [7,16], improving geometric fidelity and meeting the demands of dense view synthesis through the use of structural-motion-sparse depth or direct sensor depth.

To boost NeRF training efficiency, recent studies have explored more explicit representations, such as voxel-like structures [9,40], tensor factorization techniques [8], light field representations [2,1], and hybrid models combining hash voxels with MLPs [25]. The integration of TensoRF [8] has been investigated to enhance efficiency further. With the rise of AIGC, generative models [41,35,7] can now synthesize new object views by sampling from latent spaces. The potential of NeRF [47,23] for editing applications appears promising, addressing the classic challenge of object removal without adequately filling in the background.

Current NeRF implementations are limited by minimal editing capabilities and require extensive datasets with RGB images, camera poses, and depth information. In contrast, our approach only requires RGB images, reducing both complexity and resource demands.

3 Method

Our dynamic neural radiance field framework is established upon [21], which primarily seeks to bolster reconstruction robustness through the dual optimization of camera poses and radiance fields. We have expanded upon this foundation by removing dynamic objects and incorporating depth and optical flow data as supplementary guidance for the NeRF model. We use the extension module to perform calculations on the input images and select those with high image quality scores for the output of view selection. As shown in Fig. 2, the proposed dynamic neural radiance field process is organized into three main modules:

- **View selection:** Initiating the process, this module adaptively assesses image sequences for technical and aesthetic integrity, as detailed in Section 3.1. It ensures the selection of high-quality images, which can well meet the training needs of NeRF.
- **RGB inpainting:** This module processes the selected high-quality views to address and restore occlusions caused by dynamic objects, as detailed in Section 3.2. This module applies image inpainting guided by mask and text information to produce complete images that correspond to real-world images.
- **Geometric prior generation:** This module builds upon the inpainted images and infuses them to produce precise geometric priors that include flow and depth information, as detailed in Section 3.3. This module improves the quality of generated scenes, reinforcing our approach to creating visually coherent and authentic reconstructions.

3.1 View selection

Among NeRF-based approaches, multi-view 3D reconstruction and image calibration frequently leverage COLMAP [36,37]. Nevertheless, the regions near dynamic object boundaries often exhibit incompleteness or distortion, which undermines the accuracy and reliability of edge features, as illustrated in Fig 3. These low image-quality views may impact the ability of COLMAP to accurately reconstruct the scene, thereby increasing the risk of reconstruction failure. These fail-to-calibrate views cannot meet the requirements of NeRF-based methods, which need to be removed manually.

To mitigate such risks and prevent the need for time-consuming manual curation of images, our method harnesses the NIMA [42] and CLIP [44] models to evaluate image quality comprehensively. Utilizing CLIP, we perform perceptual

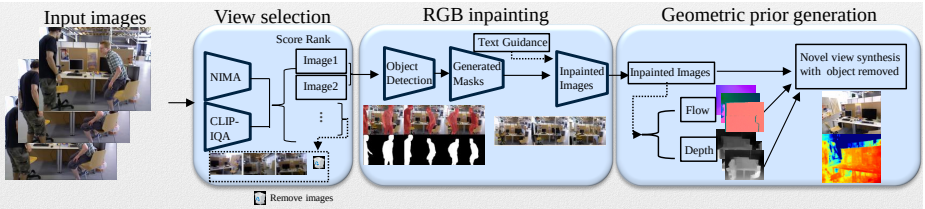

Fig. 2: Our framework adaptively processes N images using a view selection module that employs NIMA and CLIP models to score image quality. This module effectively screens out images adversely affected by dynamic objects. The RGB inpainting module then masks and refines these objects through text-guided inpainting, enhancing image integrity. Subsequently, the geometric prior generation module utilizes flow and depth priors from the inpainted images. This is used to refine the NeRF model, optimizing scene reconstruction. Ultimately, this process enables the rendering of novel views from scenes where dynamic objects have been removed.

evaluations by computing the cosine similarity between the feature representations of an image and a textual prompt (e.g., "Good photo"). Let $x \in \mathbb{R}^C$ and $t \in \mathbb{R}^C$ represent the feature vectors of the image and the prompt, respectively, where C denotes the channel number. The cosine similarity score $s \in [0, 1]$ is given by:

$$s_i = \frac{x \odot t_i}{\|x\| \cdot \|t_i\|}, \quad i \in \{1, 2\} \ , \tag{1}$$

where $\odot$ denotes the dot product and $\|\cdot\|$ indicates the ℓ_2 norm, facilitating vector normalization to unit length. This process yields two scores from which the final CLIP score, $S_{CLIP} \in [0, 1]$, is derived using a softmax function:

$$S_{CLIP} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}} \ . \tag{2}$$

Simultaneously, the NIMA model evaluates images for aesthetic and technical qualities, each being assigned scores within the range of $[1, 10]$, denoted as $S_{\text{NIMA\_Aesth}}$ for aesthetic quality and $S_{\text{NIMA\_Tech}}$ for technical quality. These scores are derived from the probabilities $p_{i,\text{aesthetic}}$ and $p_{i,\text{technical}}$, which indicate the likelihood of the image receiving an aesthetic or technical rating of i. Here, i signifies a discrete rating score that ranges from 1 to N, with N customarily set to 10, representing the spectrum of possible quality scores an image might achieve. The calculation of these scores employs the expected values of their respective distributions. The formulation is as follows:

$$S_{\text{NIMA\_Aesth}} = \sum_{i=1}^{N} p_{i,\text{aesthetic}} \cdot i \ , \tag{3}$$

$$S_{\text{NIMA\_Tech}} = \sum_{i=1}^{N} p_{i,\text{technical}} \cdot i \ . \tag{4}$$

Fig. 3: Image quality is divided into low (top) and high (bottom). The overall scene of the high-quality score image clearly and effectively retains the local details of the scene. In contrast, images with low-quality scores suffer severely from dynamic interference, showing noticeable artifacts, blurring, and ghosting.

The NIMA model systematically quantifies images' aesthetic and technical merits through this method, comprehensively evaluating image quality. Finally, we use a holistic image quality assessment strategy that integrates these models. The overall quality score, S_{Total}, combines the individual assessments:

$$S_{\text{Total}} = w_1 \cdot S_{\text{CLIP}} + w_2 \cdot S_{\text{NIMA\_Tech}} + w_3 \cdot S_{\text{NIMA\_Aesth}} \ . \tag{5}$$

We calibrate the weights w_1, w_2 and w_3 in the image quality assessment framework. This calibration is designed to preemptively filter out images that threaten the integrity of the reconstruction process, thus safeguarding against the pitfalls that commonly impede NeRF-based reconstructions. By ensuring $w_1+w_2+w_3 = 1$, our approach maintains a balanced assessment, prioritizing technical quality while still considering perceptual and aesthetic evaluations. This strategy adaptively enhances the efficiency and success rate of NeRF-based 3D reconstructions by automatically excluding problematic images before they can impact the reconstruction workflow.

3.2 RGB inpainting

The image inpainting method we employ is similar to other 2D image inpainting [41,15,50] techniques, which typically require distinct masks for each frame to identify the areas in need of inpaint. However, manually annotating each frame with a 2D mask is time-consuming and error-prone. We introduce an automated methodology for detecting and segmenting dynamic objects within image sequences to overcome this obstacle. This process requires only one execution per sequence, streamlining the mask creation workflow. By employing a 2D object detection algorithm, such as YOLO[13], we can effectively process images to capitalize on its advanced segmentation capabilities. This enables the extraction of precise target object details, which are subsequently transformed into accurate mask representations. This step prevents the inclusion of extraneous background details in the mask, resulting in an efficient solution that improves accuracy, as shown in Fig. 4.

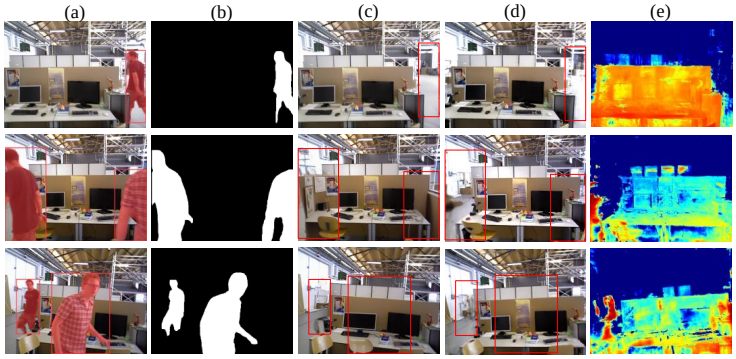

Fig. 4: Results of our inpaint method on TUM dataset. (a) and (b) Automatic object segmentation generates masks. (c) We use a stable diffusion model [35] inpainting images. (d) and (e) We utilize optical flow and depth information as geometric guidance to inpainting detail defects.

Leveraging a stable diffusion model [35], we adeptly remove dynamic objects from scenes, integrating high-precision masks and precise textual descriptions to maintain the fidelity and detail of the original image. This process ensures spatial and semantic coherence, enabling image inpainting that predicts the natural occupants of the space from which objects have been removed. The combination of automated adjusted mask objects inpainting efforts accurately, while accompanying textual descriptions guide the model towards semantically appropriate inpainting. We advance digital image processing by producing visually compelling images that seamlessly integrate into inpainted areas without distorting the surrounding context. This integration ensures the production of images that exhibit high fidelity and maintain the richness of detail synonymous with the original visuals, as depicted in Fig. 4.

3.3 Geometric prior generation

We adopt the original NeRF[22], formulating the scene representation through an MLP, F_θ, which is configured to predict the RGB color values $c = [r, g, b]$ and spatial coordinates x, y, z, across two distinct viewing directions. The predicted color for pixel r, $\widehat{P}(r)$, is obtained by volume rendering along its associated ray:

$$\hat{P}(r) = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) c_i, \tag{6}$$

$$T_i = \exp\left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right), \tag{7}$$

where δ_i represents the distance between two consecutive sampling points, N is the number of samples along the ray, and T_i represents the cumulative transmittance along the ray. $\hat{P}(\mathbf{r})$is its predicted color. The NeRF loss operates on

training images as:

$$\mathcal{L}_{RGB} = \left\| \hat{P}(r) - P(r) \right\|_2^2, \tag{8}$$

where $P(\mathbf{r})$ is the input RGB value for pixel $\mathbf{r}$. The parameters of the MLP, Θ, are optimized to minimize this loss. We leverage optical flow and depth information for geometric guidance to enhance scene reconstruction quality. We follow [34,10] to exploit the depth map generated by DPT [30,31], adding loss to relate the sample color to its ray origin distance through depth mapping, using translation and scale-invariant losses for depth supervision:

$$\mathcal{L}_d = \left| \hat{\mathbf{S}}^* - \mathbf{S}^* \right|, \tag{9}$$

where $\hat{\mathbf{S}}^*$ and $\mathbf{S}^*$ are per-frame normalized depth. Moreover, we use inpainting RGB images to compute the optical flow [43] between consecutive frames ($I_{k\to k+1}$, where $k \in \{1, 2, \ldots, P-1\}$). We improve geometric and pose accuracy within scenes by adapting [21,43] optical flow prediction.

$$\mathcal{L}_{Flow} = \left\| \hat{\mathcal{I}}_{k\to k+1} + \mathcal{I}_{k\to k+1} \right\|_1, \tag{10}$$

where $\hat{\mathcal{I}}_{k\to k+1}$ and $\mathcal{I}_{k\to k+1}$ denotes the predicted optical flow and actual optical flow from frame k to frame $k+1$. These optical flows by converting pixel coordinates and depth to 3D points. This process is utilized to calculate the pose and geometry of the scene. To optimize the output results as shown in Fig. 4.

3.4 Implementation details

In terms of image quality evaluation, it is considered that the final scores of the NIMA model and the CLIP-IQA model are different. We ensure that NIMA scores are similar to those of CLIP-IQA through normalization:

$$S_{\text{NIMA\_Aesth normalization}} = \frac{S_{\text{NIMA\_Aesth}} - 1}{9}, \tag{11}$$

$$S_{\text{NIMA\_Tech Standardization}} = \frac{S_{\text{NIMA\_Tech}} - 1}{9}. \tag{12}$$

We initialize the values of w_1, w_2, and w_3 as $\frac{1}{3}$. Regarding mask refinement, the mask obtained from the bounding box for object detection contains most of the background. Therefore, we use the YOLOv8 [13] segmentation model to obtain position encoding, which is then transformed to generate a more accurate mask. This mask refinement removes the background of the 2D bounding box. We use a stable diffusion model [35] to inpaint RGB images and combine it with text information to reduce the risk of inpainting failure. This is achieved by utilizing the inpainted RGB images to generate depth maps and optical flow images with text guidance. We optimize NeRF by providing geometric guidance,

using inpainted RGB images to generate depth and optical flow maps. This approach enhances NeRF optimization and adaptively improves the quality of 3D scene reconstruction.

Throughout the optimization process, all parameters are refined using the Adam optimizer [14]. Key parameters such as learning rates, loss weights, and the initial resolution of TensoRF [8] are maintained constant to prevent overfitting of the radiance fields. Optimization in the geometric prior generation module is conducted with an initial learning rate of $l_{\text{rotations}} = 0.005$ and $l_{\text{translations}} = 0.0005$, with the initial TensoRF resolution set to 64^3. We weight the terms in the loss function with $\lambda_{\text{flow}} = \lambda_{\text{rgb}} = 1$, and $\lambda_{\text{depth}} = 0.1$.

4 Experiments

4.1 Datasets

The TUM dataset [38], originating from the Kinect depth camera, has been foundational in our preliminary assessments, featuring six sequences characterized by high dynamic range and two with lesser dynamics. However, a considerable limitation of the TUM dataset becomes apparent during post-image inpainting: the TUM dataset lacks comparative ground truth data, making it difficult to accurately verify the quality of reconstructions. Therefore, we refer [49] to the practice of diffusion models to create synthetic datasets. We use diffusion models to remove dynamic objects from each scene in the TUM dataset. Then, we use the synthetic dataset as ground truth to test the effect of dynamic object removal in the scene.

We have compiled a dynamic object dataset to address the critical need for such comparative ground truth. This dataset is distinct from the TUM dataset. It is crafted to reflect real-world scenarios, thus offering an extensive framework for evaluating dynamic object removal methodologies. It comprises six scenarios, each designed to present three distinct conditions: one baseline scenario devoid of objects (establishing the ground truth for comparison) and two scenarios introducing either single or multiple objects for removal. This deliberate structure is essential for thoroughly examining system performance across scenarios of diverse complexity, influenced by factors such as background texture, object size, and geometric complexity within the scene.

Our *dynamic object removal* (DOR) dataset includes 18 videos captured using an iPhone 15, featuring frame counts ranging from 99 to 419 and a resolution of 1280x720 pixels. Specifically, our collection is designed to enhance the analysis of complex visual inpainting, such as shadows, reflections, and the generation of novel viewpoints. This work provides available resources and sets a new standard for the realistic reconstruction of scenes, as detailed in Fig. 5.

4.2 Metrics

To assess the effectiveness of dynamic object removal and scene restoration, we select test images at intervals of every ten frames for a detailed comparative analysis against the corresponding ground truth images within the dataset.

Fig. 5: Representative samples of our DOR dataset. Each indoor scene of our DOR dataset is divided into three scenarios: the first scenario depicts an actual situation without dynamic objects; the second scenario contains a single dynamic object in the same scene; the third scenario features multiple dynamic objects in the same scene. This setup effectively simulates the complexity of real-life scenarios by incorporating dynamic objects.

We utilize the *peak signal-to-noise ratio* (PSNR)[11], *structural similarity index* (SSIM)[46], and *learning perceptual imagery patch similarity* (LPIPS)[52] metrics to evaluate the quality of the synthetic views. It is important to note that the scope of our evaluation encompasses all synthetic and corresponding real views, not just the inpainted regions, thereby providing a comprehensive assessment of the scene reconstruction process.

4.3 Comparison With Baselines

We compare two state-of-the-art Nerf-based object removal methods [47,23] and a 3D scene completion method [33] on the TUM and DOR datasets. PixelSynth [33] is engineered for scene outpainting, facilitating the completion of scenes from single or multiple frames to synthesize novel viewpoints. This method is predicated on generative models for indoor scenes and obviates the need for optimization in the testing phase. Both RO-NeRF [47] and SPin-NeRF [23] are tailored to excise masked objects in static scenes, necessitating image rectification via LaMa[41] before new viewpoint synthesis through NeRF[22]. Tab. 1 provides a detailed comparison of these method evaluation results. These methods primarily employ generative models for single-image inpainting, relying solely on the information contained within the image content. In contrast, our approach utilizes image content and integrates textual guidance. This dual-input strategy adaptively mitigates the risk of inpainting failures and fosters the generation of more lifelike images, consequently elevating the quality of the training data. This improvement is illustrated in Fig. 7.

Compared with other methods, the scenes generated by our method are more consistent with real scenes, as shown in Fig. 6. Combining multi-view information with advanced 2D inpainting techniques, our method surpasses existing baselines in novel view synthesis, especially in achieving more realistic scene reconstruction effects. Due to the limitations of the TUM dataset, each scenario only provides a limited set of sequences and no ground truth for detailed evaluation, so a synthetic dataset (as an alternative to the ground truth) was used for testing.

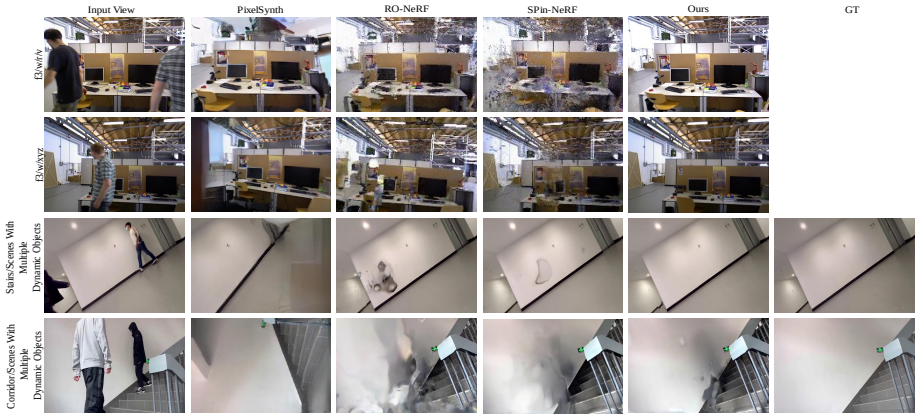

Fig. 6: Qualitative comparison to baseline. Our method significantly improves over 3D scene completion baselines on the TUM and DOR datasets for removing dynamic objects from scenes using 3D inpainting. (In the TUM dataset, each scene is only one set of sequences and, therefore, no ground truth.)

Table 1: Comparison with baseline methods on TUM dataset and DOR dataset. The best results are shown in bold.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓
TUM [38]	PixelSynth [33]	16.32	0.54	0.42
	SPIn-NeRF [23]	19.72	0.73	0.26
	RONerf [47]	20.65	0.77	0.33
	Ours	**22.33**	**0.87**	**0.20**
DOR	PixelSynth [33]	14.36	0.33	0.49
	SPIn-NeRF [23]	18.96	0.67	0.29
	RONerf [47]	19.87	0.66	0.32
	Ours	**20.54**	**0.76**	**0.22**

4.4 Ablation Study

NeRF-based ablation. We explore a variety of methodologies for training baseline NeRF models, each tailored to enhance the quality and realism of the synthesized views. The **Inpainted NeRF** employs a model trained on all inpainted images. The **Inpainted NeRF + IQA** introduces a model trained with a view selection of inpainted images, the goal of which is to minimize interference from dynamic objects. The **Inpainted NeRF + IQA + Depth** further incorporates depth maps along with the view selection inpainted images, offering richer data for training. These methodological variations are uniformly applied across all views under evaluation, with detailed comparisons presented in Tab. 2. We reduce dynamic interference through view selection and improve the effect of generated scenes by adding depth information for geometric guidance.

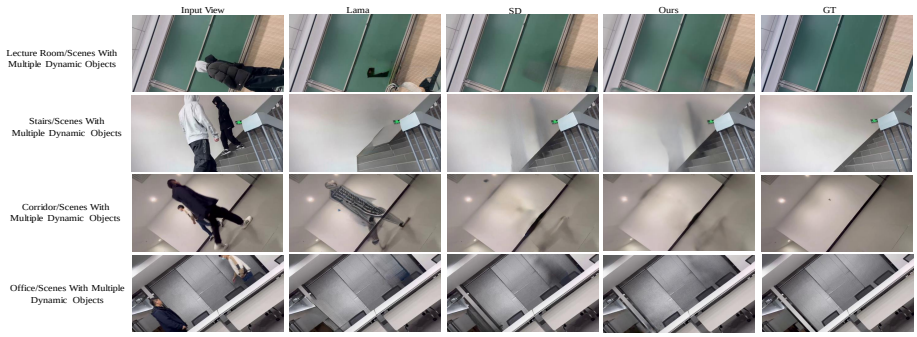

Fig. 7: This comparative visualization demonstrates the superior efficacy of our method over other image inpainting techniques. By incorporating progressive optimization[21] within the SD, our approach adaptively diminishes artifacts and guarantees stable convergence.

Table 2: **Ablation Studies.** We report PSNR, SSIM, and LPIPS for twelve scenes on the DOR dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
Inpainted NeRF	N/A	N/A	N/A
Inpainted NeRF + IQA	9.32	0.47	0.49
Inpainted NeRF + IQA + Depth	16.43	0.64	0.41
Ours	**20.54**	**0.76**	**0.22**

Image inpainting ablation. We assess the effectiveness of our image inpainting strategies by comparing our methodology with several leading-edge image inpainting techniques[35,41] on specific datasets, as outlined in Section 3. The outcomes of this evaluation are depicted in Fig. 7. We engage in detailed comparative analyses against two forefront techniques: LaMa [41] and SD [35], utilizing their official implementations and the networks trained by the original authors for a fair comparison. The results of these comparisons are systematically presented in Tab. 3. The findings from our evaluations reveal that our method surpasses established benchmarks, underscoring the efficacy of our proposed image inpainting technique. Our approach produces scene reconstructions that are more closely aligned with the GT and demonstrate enhanced realism. This highlights the superior capability of our method in refining image inpainting processes to achieve more realistic and precise restorations.

Table 3: **Ablation on image inpainting methods.** We report PSNR, SSIM, and LPIPS for twelve scenes on the DOR dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
LaMa [41]	19.24	0.71	0.25
SD [35]	20.23	0.73	0.24
Ours	**20.54**	**0.76**	**0.22**

Table 4: **Ablation on view selection methods.** We report PSNR, SSIM, and LPIPS for twelve scenes on the DOR dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
All views	18.76	0.64	0.46
1/5th	18.85	0.69	0.39
Single view	N/A	N/A	N/A
Ours	**20.54**	**0.76**	**0.22**

View selection ablation. We evaluate the efficacy of the view selection mechanism proposed in Section 3.1. This is achieved by benchmarking our methodology against various view selection strategies within the dataset, with the comparative results detailed in Tab. 4. Our method selects two views with higher scores every five to train the NeRF model based on the image quality score. This contrasts with the baseline strategy, which trains the NeRF model using a selected subset of views at regular intervals. **All views** approach utilizes all the inpainting views when training the NeRF model. In the **1/5th** approach, one frame is selected every five frames to train the NeRF model. In the **Single view** approach, only a single middle frame is used to train the NeRF model. The results of our experiments indicate that our method surpasses these baseline approaches, demonstrating that our proposed view selection strategy effectively identifies an optimal set of views for incorporation into the NeRF optimization process. This highlights the superior ability of our approach to enhance the quality and precision of the rendered scenes.

5 Conclusion

We present a novel framework for training NeRF that effectively addresses the challenge of realistically removing dynamic objects from rendered scenes. Our approach combines 2D inpainting techniques by integrating an automated evaluation mechanism based on CLIP. This mechanism is adept at selecting high-quality, single-view inpaintings with reduced interference from dynamic objects. Through rigorous experimental validation, we have confirmed that our method markedly improves the synthesis of new 3D viewpoints in scene painting over existing approaches. Additionally, we have developed and provided a dataset specifically curated to evaluate the performance of our framework, thereby establishing a new benchmark for future research in the field. This endeavor contributes significantly to advancing methodologies for enhancing the realism and fidelity of digitally rendered environments.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (Nos. 62076021 and 62376020).

References

1. Attal, B., Huang, J.B., Richardt, C., Zollhoefer, M., Kopf, J., O'Toole, M., Kim, C.: Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In: CVPR (2023) 4
2. Attal, B., Huang, J.B., Zollhoefer, M., Kopf, J., Kim, C.: Learning neural light fields with ray-space embedding networks. In: CVPR (2022) 4
3. Author1, F., Author2, F.: Symmnerf: Learning to explore symmetry prior for single-view view synthesis. In: ACCV (1). pp. 228–244. Asian Conference on Computer Vision (2022) 4
4. Author1, F., Author2, F.: Teacher-guided learning for blind image quality assessment. In: ACCV (3). pp. 206–222. Asian Conference on Computer Vision (2022) 3
5. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: CVPR. pp. 1–7 (2022) 4
6. Bescos, B., Fácil, J.M., Civera, J., Neira, J.: Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2258–2264. IEEE (2018) 4
7. Bešić, B., Valada, A.: Dynamic object removal and spatio-temporal rgb-d inpainting via geometry-aware adversarial learning. IEEE Transactions on Intelligent Vehicles **7**(2), 2 (2022) 4
8. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: ECCV (2022) 4, 10
9. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: CVPR. p. 3 (2022) 4
10. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: ICCV (2021) 9
11. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: International Conference on Pattern Recognition (ICPR). p. 6. IEEE (2010) 11
12. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J.: Self-calibrating neural radiance fields. In: ICCV (2021) 4
13. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (2023) 7, 9
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015) 10
15. Li, Z., Lu, C.Z., Qin, J., Guo, C.L., Cheng, M.M.: Towards an end-to-end framework for flow-guided video inpainting. In: CVPR. pp. 5–7 (2022) 7
16. Liao, M., Lu, F., Zhou, D., Zhang, S., Li, W., Yang, R.: Dvi: Depth guided video inpainting for autonomous driving. In: ECCV. p. 2 (2020) 4
17. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: ICCV (2021) 4
18. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: European Conference on Computer Vision (ECCV). p. 2 (2018) 4
19. Luo, W., Schwing, A.G.: Real-time 3d reconstruction of complex scenes from a single hand-held camera. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 714–730 (2018) 1
20. Martín-Martín, R., Méndez-González, A.P., Morán, F., Barreto, A.: Mip-nerf: Real-time neural radiance field rendering using multiscale image pyramids. In: arXiv preprint arXiv:2107.00701 (2021) 4

21. Meuleman, A., Liu, Y.L., Gao, C., Huang, J.B., Kim, C., Kim, M.H., Kopf, J.: Progressively optimized local radiance fields for robust view synthesis. In: CVPR (2023) 5, 9, 13
22. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 1, 4, 8, 11
23. Mirzaei, A., Aumentado-Armstrong, T., Derpanis, K.G., Kelly, J., Brubaker, M.A., Gilitschenski, I., Levinshtein, A.: SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In: CVPR (2023) 2, 4, 11, 12
24. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: Lens: Localization enhanced by nerf synthesis. In: Conference on Robot Learning. p. 1 (2022) 1
25. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. TOG (2022) 4
26. Newson, A., Almansa, A., Fradet, M., Pérez, P., Bascou, G.: Video inpainting of complex scenes. SIAM Journal on Imaging Sciences **7**(4), 1993–2019 (2014) 4
27. Nguyen, D.T., Meilhac, C.: Real-time monocular dense mapping on aerial robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 768–774 (2018) 1
28. Ozturkcan, S.: Service innovation: Using augmented reality in the ikea place app. Journal of Information Technology Teaching Cases **11**(1), 1 (2021) 1
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 2, 3
30. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ArXiv preprint (2021) 9
31. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer (2020) 9
32. Reiser, H.P., Aittala, M., Durand, F.: Regnerf: Plenoptic neural radiance fields from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1793–1802 (2021) 4
33. Rockwell, C., Fouhey, D.F., Johnson, J.: Pixelsynth: Generating a 3d-consistent experience from a single image. In: ICCV (2021) 11, 12
34. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: CVPR. pp. 2–3 (2022) 9
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2, 3, 4, 8, 9, 13
36. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016) 1, 3, 5
37. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV (2016) 1, 3, 5
38. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems (2012) 3, 10, 12
39. Sulaiman, M.Z., Abdul Aziz, M.N., Abu Bakar, M.H., Halili, N.A., Azuddin, M.A.: Matterport: Virtual tour as a new marketing approach in real estate business during pandemic covid-19. In: International Conference of Innovation in Media and Visual Design (IMDES). p. 1 (2020) 1

40. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: CVPR. p. 3 (2022) 4
41. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021) 2, 3, 4, 7, 11, 13
42. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE TRANSACTIONS ON IMAGE PROCESSING (2018) 3, 5
43. Teed, Z., Deng, J.: Raft: Recurrent all pairs field transforms for optical flow (2020) 9
44. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023) 3, 5
45. Wang, S., et al.: Ie-nerf: Inpainting enhanced neural radiance fields in the wild. arXiv preprint arXiv:2407.10695 (2024) 4
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 6 (2004) 11
47. Weder, S., Garcia-Hernando, G., Monszpart, A., Pollefeys, M., Brostow, G.J., Firman, M., Vicente, S.: Removing objects from neural radiance fields. In: CVPR (2023) 2, 4, 11, 12
48. Wu, Q., Liu, X., Chen, Y., Li, K., Zheng, C., Cai, J., Zheng, J.: Object-compositional neural implicit surfaces. In: ECCV. pp. 1–2 (2022) 1
49. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models (2023) 10
50. Xu, R., Li, X., Zhou, B., Loy, C.C.: Deep flow-guided video inpainting. In: CVPR. pp. 2–5 (2019) 7
51. Yang, B., Zhang, Y., Xu, Y., Li, Y., Zhou, H., Bao, H., Zhang, G., Cui, Z.: Learning object-compositional neural radiance field for editable scene rendering. In: ICCV. pp. 1, 2, 6, 7, 8 (October 2021) 1
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 11