J. Vis. Commun. Image R. 107 (2025) 104365

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

DA4NeRF: Depth-aware Augmentation technique for Neural Radiance Fields☆

Hamed Razavi Khosroshahi [a,*], Jaime Sancho [b], Gun Bang [c], Gauthier Lafruit [a], Eduardo Juarez [b], Mehrdad Teratani [a]

[a] Laboratory of Image Synthesis and Analysis, Universite Libre de Bruxelles, Av. F.-D. Roosevelt 50, CP 165/57, Brussels, 1050, Belgium
[b] Research Center on Software Technologies and Multimedia Systems, Universidad Politecnica de Madrid, Edificio La Arboleda. Calle Alan Turing 3, Madrid, 28031, Spain
[c] Electronics, Telecommunication Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, South Korea

ARTICLE INFO

Keywords:
Neural radiance fields
Data augmentation
View synthesis
Depth maps
NeRF

ABSTRACT

Neural Radiance Fields (NeRF) demonstrate impressive capabilities in rendering novel views of specific scenes by learning an implicit volumetric representation from posed RGB images without any depth information. View synthesis is the computational process of synthesizing novel images of a scene from different viewpoints, based on a set of existing images. One big problem is the need for a large number of images in the training datasets for neural network-based view synthesis frameworks. The challenge of data augmentation for view synthesis applications has not been addressed yet. NeRF models require comprehensive scene coverage in multiple views to accurately estimate radiance and density at any point. In cases without sufficient coverage of scenes with different viewing directions, cannot effectively interpolate or extrapolate unseen scene parts. In this paper, we introduce a new pipeline to tackle this data augmentation problem using depth data. We use MPEG's Depth Estimation Reference Software and Reference View Synthesizer to add novel non-existent views to the training sets needed for the NeRF framework. Experimental results show that our approach improves the quality of the rendered images using NeRF's model. The average quality increased by 6.4 dB in terms of Peak Signal-to-Noise Ratio (PSNR), with the highest increase being 11 dB. Our approach not only adds the ability to handle the sparsely captured multiview content to be used in the NeRF framework, but also makes NeRF more accurate and useful for creating high-quality virtual views.

1. Introduction

The performance of deep learning models is highly dependent on the quantity, and diversity of the datasets used for training. This dependency is particularly noticeable in the field of 3D neural rendering, where each unique scene requires individualized training. As a result, when training images are scarce, the resulting models tend to underperform. Unfortunately, acquiring or constructing large, and high-quality datasets is a challenging task.

Conventional data augmentation methods [1–3], including rotation, noise addition, shearing, translation, and color modifications, are mainly designed for 2D image processing and have demonstrated their effectiveness in tasks such as object detection and segmentation [4,5]. These data augmentation techniques fall short of addressing the complexities of 3D reconstruction and view synthesis [6,7]. The inherent challenges posed by the three-dimensional nature of the data in applications like Neural Radiance Fields (NeRF) [8] framework for 3D rendering [9–11], require the development of novel augmentation strategies that can effectively enhance data diversity considering the specific requirements of 3D scene reconstruction.

The proposed depth-aware augmentation technique specifically targets the gap in data augmentation for 3D environments by focusing on depth perception and realistic view synthesis. It is particularly designed for scenarios where there are not plenty of images covering the scene, and conventional augmentation methods prove insufficient for realistic 3D scene generation. (C1.R2.S1)

The most important challenge to train neural rendering models is the lack of a sufficient number of available sparse images for training in each scene. In such cases, NeRF fails to train or fails to render high-quality novel views. The reasons for such a failure are (C1.R2.S2):

- **Coverage of viewing angles:** NeRF models estimate the radiance and density at any given point in space by integrating

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: hamed.razavi.khosroshahi@ulb.be (H. Razavi Khosroshahi).

https://doi.org/10.1016/j.jvcir.2024.104365
Received 22 May 2024; Received in revised form 10 October 2024; Accepted 3 December 2024
Available online 9 December 2024

information across multiple views. If the training images do not sufficiently cover the scene, the model cannot accurately interpolate or extrapolate the unseen parts of the scene.
- **Detail preservation:** Without enough images, NeRF struggles to capture fine details, leading to blurry or inaccurate reconstructions. This is particularly evident in areas where complex textures or occlusions occur (shown in Section 7).
- **Generalization and overfitting:** With limited data, NeRF models tend to overfit to the available views, failing to generalize well to new viewpoints. This results in poor performance when generating novel views that differ significantly from the training images.

To address this challenge, the authors propose an approach of augmentation techniques adapted for NeRF, which leverages techniques for view synthesis and depth maps to enrich the training data significantly. The introduction of depth-aware augmentation and the integration of a view synthesizer software are highlighted as enhancements that substantially improve the quality and robustness of the NeRF model. Depth estimation software, with its ability to generate accurate depth maps (in case of the datasets without high-quality depth maps available), and on the other hand a powerful view synthesizer software, lead to a pool of training datasets with more images and more diversity. It also enables NeRF models to generate photorealistic scenes even when constrained by sparse data. This approach not only broadens the quantity in the dataset but also maintains consistency to the 3D spatial and perspective complexity essential for accurate scene rendering.

The experiments demonstrate that our depth-aware augmentation improves the fidelity of rendered images by up to 30% in scenarios with sparse viewpoints. It also enhances the model's ability to generalize to new views by a significant margin, confirming the effectiveness of integrating depth information into the NeRF training process.(C1.R2.S3)

DA4NeRF further explores the impact of data augmentation for NeRF models and boosts results for neural rendering and the crucial role of data augmentation for using NeRF in creating realistic 3D scenes from limited sparse viewpoints. It emphasizes the importance of quality and quantity of the available data for the accuracy of NeRF models and positions the proposed methodology as an advancement in overcoming the limitations posed by limited sparse data. By integrating depth information and employing a view synthesizer within the NeRF framework, the authors aim to synthesize novel views that significantly enhance model quality.

This research represents a contribution to improve the available limited datasets in the number of images and diversity for training the NeRF framework. It offers a detailed exploration and analysis of the proposed methodology, demonstrating its effectiveness through conducting several experiments on different datasets. The integration of depth information and the view synthesizer augments the dataset and improves the neural rendering's training procedure and its inference. Experimental evaluations using structured captured datasets, as well as NeRF Real datasets [8,12], show the effectiveness of our approach in enhancing synthesis quality with a limited number of training views.

2. Preliminaries

This section reviews the significant contributions and methodologies of structure-from-motion (COLMAP), Depth Estimation Reference Software (DERS), Reference View Synthesizer Software (RVS), and Neural Radiance Fields for view synthesis (NeRF), highlighting their unique approaches and intersections in the field.

2.1. COLMAP

COLMAP [13] is a leading software platform for performing Structure from Motion (SfM) [14] and Multi-View Stereo (MVS) [15], widely recognized for its effectiveness in reconstructing 3D models from unordered image datasets. It distinguishes itself through advanced features such as automatic camera calibration, image matching, and dense point cloud generation [16,17], facilitating accurate 3D modeling across diverse applications. With its robust algorithmic foundation, COLMAP has contributed significantly to the advancement of photogrammetry [18] and computer vision fields, enabling precise spatial analysis and visualization. Moreover, its open-source nature allows for extensive customization and integration, making it a valuable tool for researchers and professionals seeking to leverage the latest in 3D reconstruction technology.

2.2. Depth Estimation Reference Software (DERS)

MPEG Depth Estimation Reference Software (DERS) [19] is a tool in the context of 3D video processing and multimedia applications, developed by the Moving Picture Experts Group (MPEG). This software plays a crucial role in the generation of depth maps [20–22], which are essential for creating stereoscopic (3D) and multi-view video content. Depth maps represent the distance between the camera and the objects in a scene, providing vital information that allows for the simulation of three-dimensional spaces in 2D images. DERS uses advanced algorithms to analyze 2D video frames and estimate the depth of various elements within the scene, facilitating the creation of more immersive and realistic 3D video experiences. This technology has broad applications in virtual reality (VR) environments. As 3D content continues to gain popularity, the importance of efficient and accurate depth estimation software like DERS grows, driving innovation and improvements in the field of immersive and 3D multimedia.

2.3. Reference View Synthesizer (RVS)

The MPEG Reference View Synthesizer (RVS) [23] is an advanced tool designed to enhance the field of 3D video production and virtual reality applications. Developed by MPEG, RVS is helpful in synthesizing new viewpoints from existing views, a process critical for creating multi-view video and 3D displays. By leveraging depth information often generated by tools like DERS, RVS can interpolate or extrapolate new views between the original camera positions. This capability is crucial for producing content for glasses-free 3D displays, virtual reality environments, and augmented reality applications, where the perspective needs to be adjusted in real-time according to the viewer's position.

2.4. Neural Radiance Fields (NeRF)

Neural Radiance Fields (NeRF) [8] have emerged as an approach in the synthesis of photorealistic scenes, offering significant advancements in the realm of computer vision. This technique models the volumetric scene function using a fully connected deep neural network [24], which maps 3D coordinates to color and density, enabling highly detailed and continuous reconstructions of complex scenes from a sparse set of images. Since its introduction, NeRF has inspired plenty of related works aiming to address its limitations and expand its applicability. These efforts include improving the rendering speed through more efficient data structures and algorithms, enhancing the quality of reconstructions in challenging lighting conditions, and extending the framework to dynamic scenes. Furthermore, variations of NeRF have been developed to incorporate semantic segmentation, enabling more nuanced scene understanding and manipulation. Other notable directions include integrating NeRF with traditional computer graphics techniques for more scalable scene representations and exploring its potential in virtual and augmented reality applications. The continuous evolution of NeRF-related technologies underscores their potential to revolutionize how we capture, recreate, and interact with digital representations of the real world.

In this technology, a static scene is represented as a continuous 5D function that outputs the radiance emitted in each direction (θ, ϕ) at each point (x, y, z) in space and a density at each point which acts like a differential opacity controlling how much radiance is accumulated by a ray passing through (x, y, z). This method optimizes a deep fully-connected neural network without any convolutional [25] layers (often referred to as a multilayer perceptron or MLP) to represent this function by regressing from a single 5D coordinate (x, y, z, θ, ϕ) to a single volume density and view-dependent RGB color (σ, c).

Recent studies have embraced Neural Radiance Fields for their straightforward design and superior rendering capabilities, applying them to diverse enhancements including generative adversarial networks [26,27], video synthesis [28,29], relighting [30,31], and scene editing [32,33], among others.

3. Related works

NeRF-based techniques typically require numerous images from varying viewpoints to facilitate training. To mitigate the substantial data requirements of NeRF, several approaches have been developed that leverage existing training data [34,35], employ meta-learning strategies [36], and incorporate additional supervision [37,38]. PixelNeRF [35] uses training images during test-time rendering, a feature not considered by traditional NeRF. It projects the convolutional features of training images onto rays from novel viewpoints, serving as a conditional embedding for MLP inference. Similarly, IBRNet [34] employs a comparable approach but integrates an additional ray transformer to enhance density estimation. MetaNeRF [36] suggests initializing the MLP weights through pre-training on a comprehensive dataset, followed by scene-specific fine-tuning. DietNeRF [37] introduces a pairwise loss that enhances multi-view consistency by minimizing the cosine distance between high-level semantic features across different viewpoints. RegNeRF [39] synthesizes image patches from unobserved camera positions and enforces consistency in the RGB values using a trained normalizing flow model, while also applying a smoothness loss to the density values. DSNeRF [40] leverages sparse depth data produced by COLMAP [13] as direct supervision for the rendered depth maps. Our approach aligns with DSNeRF in using depth information as a supervisory signal, yet we innovate by using this data to create novel training samples through view synthesis technologies. In this study, we focus on enhancing depth maps derived from external depth estimation methods to augment data, noting that our methodology is compatible with all NeRF-based models. In this paper, we contrast our findings with DSNeRF, highlighting the efficiency of our method.

4. Proposed approach

In this section, we elaborate on a methodology designed to augment datasets for NeRF models, particularly when the number of available images is insufficient for comprehensive training. Our approach involves the addition of virtually synthesized images named augmented images to the training dataset through a Depth Image-based Rendering (DIBR) [41] process.

As mentioned in Section 1, the efficiency of NeRF models is significantly influenced by the quantity and diversity of input training images. A rich dataset, characterized by a wide range of angles, positions, lighting conditions, and distances, contributes to a more detailed and accurate 3D representation. Such diversity is essential for training models like NeRF to generalize effectively across unseen views, which enhances the accuracy and robustness of 3D reconstructions and renderings. On the other hand, a training dataset lacking in diversity or quantity may affect the model's ability to be overfitted on the available views and fail to reconstruct unseen scenes accurately. In our previous studies [42,43], we demonstrated that adding more images contributes to an enhancement in the quality of the NeRF model.

Algorithm 1 Depth-aware Data Augmentation for NeRF (C2.R2.S2)

Input: Original dataset
Output: Enhanced NeRF model trained on original and augmented data
function CalibrateDataset(dataset)
    return cameraParameters(dataset)
end function
function CreateSubset(data, size)
    return extractSubset(data, size)
end function
function TrainNeRF(data)
    model ← initializeNeRFModel()
    model.train(data)
    return model
end function
function SynthesizeViews(model, test_set)
    return model.generateViews(test_set)
end function
function ExtractDepthMaps(data)
    return depthEstimation(data)
end function
function SynthesizeMissingImages(images, depth_maps)
    return synthesizeImages(images, depth_maps)
end function
function AugmentData(original, augmented)
    return original + augmented
end function
dataset ← LoadDataset
subsets ← [4, 3, 2]
for size in subsets **do**
    subset ← CreateSubset(calibrated_data, size)
    model ← TrainNeRF(subset)
    test_set ← getTestSet(size)
    synthesized_views ← SynthesizeViews(model, test_set)
    depth_maps ← ExtractDepthMaps(subset)
    missing_images ← SynthesizeMissing(subset, depth_maps)
    augmented_data ← AugmentData(subset, missing_images)
    retrained_model ← TrainNeRF(augmented_data)
    final_output ← SynthesizeViews(retrained_model, test_set)
end for
return results

To augment the training dataset for NeRF, we propose a sequence of steps outlined in our methodology (shown in Fig. 1, and Algorithm 1): (1) Calibration of original images, (2) Depth map generation, (3) Synthesis of virtual images, (4) Incorporation into the training pool, (5) Training of the NeRF model, and (6) View synthesis with the trained model.

4.1. Calibration of original images

Initially, the available images are calibrated to extract camera parameters using Structure-from-Motion (SfM) techniques, such as COLMAP, or using the OpenCV method. The result of this process is the extraction of both intrinsic and extrinsic camera parameters. These parameters are subsequently used in tasks such as depth estimation, view synthesis, and model training.

4.2. Depth map generation

Subsequently, depth maps are generated from the existing images utilizing the MPEG Depth Estimation Reference Software (DERS) [19], or MPEG Immersive Video Depth Estimation (IVDE) [45]. Both tools are known for their high-quality depth estimation capabilities. This step is bypassed if depth maps are already available within the dataset.

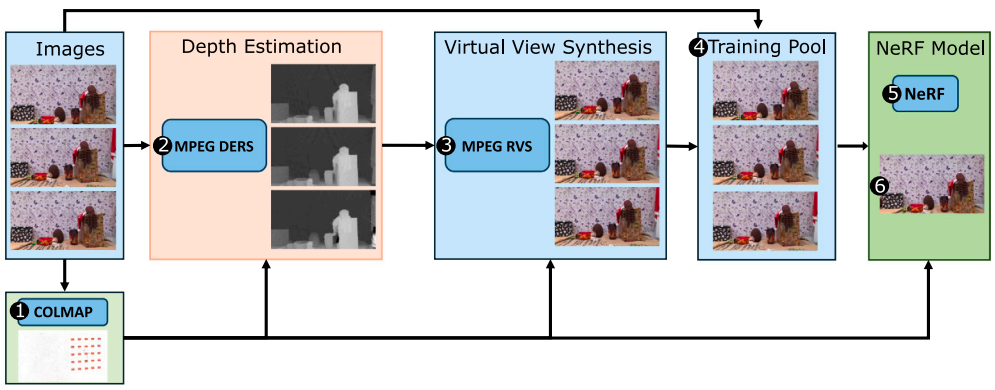

Fig. 1. Pipeline of the proposed method: (1) Calibration using structure-from-motion (COLMAP) [13]), (2) Depth estimation (DERS [19]), (3) View synthesis (RVS [23]), (4) Dataset pool preparation, (5) Training NeRF model (NeRF-pytorch [44], and (6) Target view synthesis. (C2.R2.S1).

4.3. Synthesis of virtual images

Employing the MPEG Reference View Synthesizer software (RVS), virtual images are synthesized to fill the gaps of missing views. This process leverages the original images, along with their corresponding depth maps and camera parameters, to generate these synthesized images. In particular, RVS utilizes depth maps to accurately position and orient virtual views in the 3D space of the scene. Depth maps inform the synthesizer about the relative distances of objects from the viewpoint, enabling it to reconstruct scene geometry with high precision. By integrating depth maps into the synthesis process, RVS can more effectively handle occlusions and varying scene complexities, resulting in more realistic and coherent virtual images. The depth-aware synthesis not only improves the fidelity of the interpolated views but also enhances the overall photorealism of the augmented dataset, significantly boosting the performance of the subsequent NeRF model training. (C9.R2.S1)

4.4. Incorporation into training pool

The synthesized virtual images, along with their corresponding camera parameters, are added to the pool of available images for training.

4.5. Training of the NeRF model

The enriched dataset pool is then used to train the NeRF model, aiming to enhance its performance and generalizability.

4.6. View synthesis with trained model

we proceed to synthesize target views to evaluate both the quality of the augmented images and the effectiveness of the trained model. This synthesis involves generating images from viewpoints that were not explicitly represented in the training set. The primary metric for assessing the quality of these rendered images is the Peak Signal-to-Noise Ratio (PSNR), which provides a quantitative measure of image fidelity. By analyzing the PSNR values, we can measure the accuracy and visual quality of the synthesized views, thus verifying the model's ability to reconstruct high-quality images from the learned data. This step is crucial in determining the practical utility of the augmented dataset and the robustness of the NeRF model in producing visually compelling and accurate representations from novel viewpoints. (C8.R2.S1)

Several critical points have to be considered in our methodology. Firstly, the positioning of virtual images must be within the original images' field of view to avoid the introduction of new occlusions that could degrade the quality of the trained model. Additionally, the accuracy and clarity of the depth maps are critical; low-quality depth maps can lead to artifacts in the augmented images, negatively impacting model quality. Our approach, through careful augmentation and consideration of these factors, aims to significantly enhance the training dataset for NeRF models, facilitating superior 3D scene reconstruction and rendering.

5. Dataset description

This section details the datasets used in our research to validate the proposed methodology. We conducted experiments using two distinct approaches. Initially, we established a structured dataset for proof of concept through specific, controlled datasets; subsequently, we employed benchmark datasets to evaluate the performance enhancements afforded by our technique in scenarios characterized by limited data availability.

The choice of datasets can introduce biases that may affect the generalizability of the results. Our experiments primarily utilized three datasets, which were selected for their diversity in scene complexity and lighting conditions and both real-world scene and synthetic dataset. Then as a second type of experiments, we used two NeRF datasets to benchmark our work. To mitigate potential biases, we ensured a balanced representation of different scene types and conducted additional validation using synthetic datasets, thus providing a controlled environment to test our augmentation technique.(C3.R2.S2)

5.1. Structured dataset for proof of concept

For the preliminary phase of our investigation, we selected three 5×5 input images, each serving to demonstrate the feasibility and effectiveness of our approach under controlled conditions. The first dataset, a subset derived from the ULB toys table dataset [46–48], comprises images arranged in a 5×5 grid with a fixed baseline distance of 32 mm between adjacent captures. This dataset is referenced in multiple sources, indicating its validity and reliability for this type of research. The second dataset, similarly organized in a 5×5 configuration, was obtained using an Azure Kinect camera system, which was positioned on a movable frame at the Universidad Politécnica de Madrid (**UPM** dataset hereafter). This setup maintained a baseline distance of 20 mm between each captured view. The third dataset is extracted from the ETRI Garage dataset, a synthetically generated set using Blender software [49], by the Electronics and Telecommunications Research Institute (ETRI). This dataset, arranged in a 5×5 grid, differs from the previous two with a larger baseline of 60 mm between captures. The selection of these varied datasets, encompassing both actual and virtual environments, allows for a comprehensive evaluation of our methodology across different baseline distances and conditions.

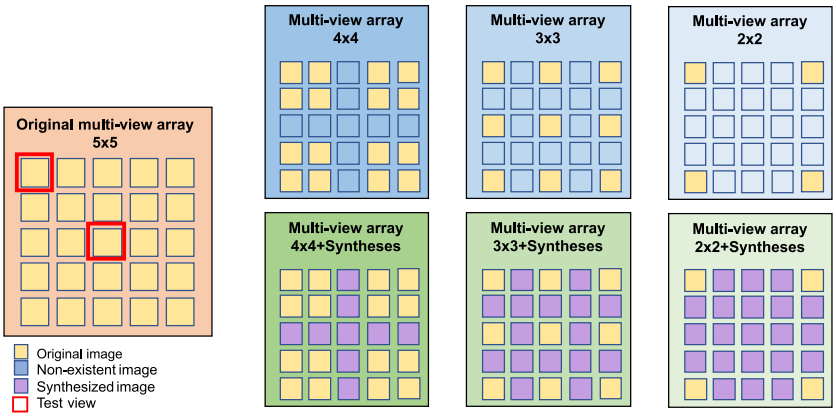

Fig. 2. Experiment configurations. From left: The first configuration is for full original 5 × 5images available (yellow squares), Second column shows the 4 × 4 configuration with just original images (up), and original + synthesized (purple squares) images (bottom), Third column shows the 3 × 3 configuration with just original images (up), and original + synthesized (purple squares) images (bottom), and the last column shows the 2 × 2 configuration with just original images (up), and original + synthesized (purple squares) images (bottom).

5.2. Benchmarking dataset

For a more extensive evaluation, we applied our methodology to two additional datasets from NeRF real datasets, named Orchid [50] and Fortress [12], to assess the improvement in data quality enabled by our approach, particularly in situations with lack of enough training images. The structure of these datasets including the position of the cameras, depth of objects, disparity, and number of images are close to the conditions of the datasets we proposed for proof of concept in the Section. 5.1. Also these datasets are frequently employed in the evaluation of 3D technologies, including neural radiance fields (NeRF) and Local Light Field Fusion (LLFF) [12]. The Orchid dataset comprises 25 images, whereas the Fortress dataset contains 42 images. For each dataset, we executed two sets of experiments: one using the full original available images, and the other employing a minimal dataset configuration, typically including four images. This minimal dataset was then expanded through image augmentation to illustrate the beneficial impact of augmented data within this context. In these experiments, we have the same test set for the Orchid dataset as the previous datasets, but for the Fortress dataset where the dataset inherently contains a larger number of images, we designated five images as a consistent test set across all trials, to ensure uniformity in evaluation metrics.

6. Experiment conditions

In order to make clear the details of the proposed pipeline, configurations of using depth estimation tool to generate the depth maps, configurations of view synthesizer, and configurations of training NeRF are discussed.

All these experiments were conducted using the nerf-pytorch [44] implementation of NeRF. Each experiment utilized a 5 × 5 dataset configuration (two additional unstructured datasets are used as well, which are introduced in Section 5). Camera parameters were determined via COLMAP, and depth maps for the images were generated employing MPEG's Depth Estimation Reference Software (DERS), with consideration for all 5 × 5 images. The test image was consistently placed in the upper left corner labeled V_0, and the center of the dataset structure labeled V_{12} (Fig. 2). For augmentation purposes at each experimental stage, synthesized images were generated using RVS, based on the available images in each step of experiments, excluding the test images to preserve the experimental conditions integrity.

On the depth estimation side, we used two approaches to estimate the depth of our datasets to be used in the data-augmentation method. In the first approach, we used all original images to estimate the best depth maps (hereafter named D_H). This approach helps to have more accurate and sharp depth maps with fewer holes in it. The method is close to the datasets that already have a good depth map. For the second approach, we used just available original images (shown in Fig. 2 - yellow squares) in each step to do the depth estimation (hereafter named D_L). This approach is closer to the real cases in which we do not have good depth maps. But in this process, we suffer from holes in depth maps (especially in 2 × 2 datasets) that cause artifacts and bad images in augmented data.

On the view synthesizer side, we used two approaches. First, synthesizing the non-existent images using the D_H (hereafter DA_1), which gives the best quality augmented images. In this method, available original images in each configuration and their corresponding best quality depth maps are used to synthesize the views. Second, using D_L. In this approach, the same as the previous approach, the available images are used but with their corresponding D_L (hereafter DA_2).

On the NeRF side, we executed two parallel sets of experiments to examine the performance of the NeRF model across different conditions of data availability, with an emphasis on the integration of extra augmented images. The aim is to evaluate the impact of this data augmentation on the NeRF model's capacity to generate precise and high-quality renderings. Through these experimental series, our objective is to illuminate the potential enhancements in model performance facilitated by the strategic incorporation of synthesized data, offering insights into optimizing NeRF models for improved neural rendering capabilities.

The first set of experiments was committed to evaluate the NeRF model utilizing solely the original images. Initiating with a comprehensive dataset including 5 × 5 original images, we decreased step by step the dataset size to 4 × 4, 3 × 3, and 2 × 2 subsets for training. This decremental strategy (illustrated in Fig. 2 - top row) enabled a structural examination of the consequences of decreasing training data on model efficacy. This experiment established a baseline for the NeRF model's performance under constrained data conditions, reflecting common challenges in real-world deployments.

Subsequently, the second experiment series was designed to investigate the benefits of supplementing the original dataset with synthesized images. We enhanced the datasets by integrating synthesized images, produced in the prior phase, into the vacant positions of the original dataset's missing views. This integration yielded an augmented 5 × 5 dataset, represented by a combination of original (marked as yellow squares) and synthesized augmented (denoted as purple squares) images (depicted in Fig. 2 - bottom row).

7. Results and discussions

In this section, we show the outcomes observed at various stages of our investigation. First, we delve into the results related to the

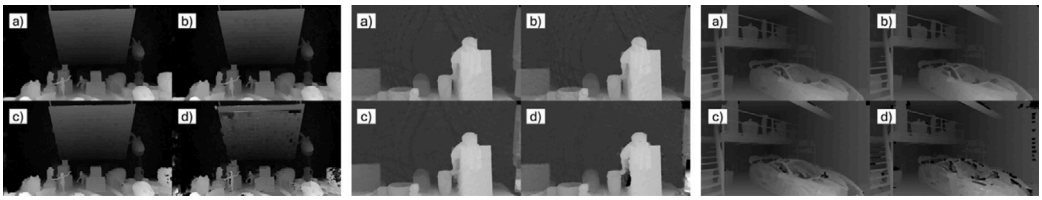

Fig. 3. Generated depth maps for ULB Toys Table dataset (left), UPM dataset (middle), and ETRI Garage dataset (right), in each column: (a) high-quality depth map using all original 5 × 5 images of the dataset, (b) using available images for 4 × 4 configuration, (c) for 3 × 3 configuration, and (d) for 2 × 2 configuration.

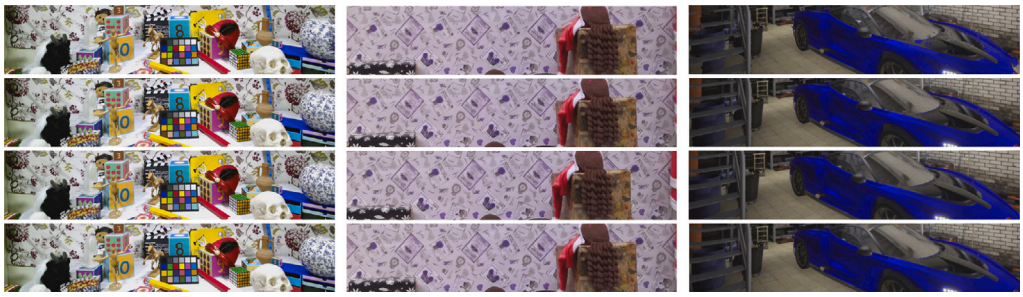

Fig. 4. ULB Toys Table dataset augmented images (left column), UPM dataset augmented images (middle column), and Garage dataset augmented images (right column) from top to bottom: Using depth maps with available images for 2 × 2 configuration, 3 × 3 configuration, 4 × 4 configuration, and best quality.

generation of depth maps across different configurations of reference images and their subsequent influence on the augmented images. Subsequently, we explore the impact of these augmented dataset pools on the NeRF model's performance.

7.1. Quality of depth maps and augmented images

In this section, we show the quality of the generated depth maps based on the quality of the augmented images.

Fig. 3 presents the outcomes of depth map generation for the ULB Toys table dataset, UPM dataset, and ETRI Garage dataset employing two distinct methodologies. The first methodology involves the creation of high-quality depth maps (D_H) using the entire original images from the datasets (labeled a, in figures). While this approach requires a longer execution time for depth map estimation, it results in better quality and higher accuracy (shape edges and hole-free surfaces in the D_H). On the other hand, the second methodology generates depth maps based on the images available within each dataset configuration (D_L). For instance, within a 2 × 2 configuration of original images, depth maps are produced for the available image positions using the corresponding 2 × 2 images. This method is characterized by a reduced computational time compared to the first approach, at the expense of lower-quality depth maps. Holes can be seen in this type of depth map (labeled b, c, and d in Fig. 3).

Fig. 4 shows the augmented images derived using the aforementioned depth map methodologies. It is evident that augmented images constructed from D_H exhibit fewer artifacts and attain a higher Peak Signal-to-Noise Ratio (PSNR), underscoring the benefits of this approach in terms of image quality. Fig. 5 further illustrates the heatmaps of the augmented images, describing variations across different configurations and methodologies. This visual representation aids in comprehensively understanding the impact of each depth map generation approach on the synthesized images and, by extension, on the overall performance of the NeRF model. Based on this information, best augmented images, happen when good quality depth maps are available. Through this analysis, we provide insights into the strategic optimization of dataset augmentation for enhancing the fidelity and accuracy of neural rendering.

7.2. Impact of data augmentation on NeRF model

This section delves into both objective and subjective analyses of our findings. Objective assessments, quantified using the Peak Signal-to-Noise Ratio (PSNR), are depicted in Fig. 6 and Table 1, for structured 5 × 5 datasets and for NeRF benchmarking datasets.

Within Fig. 6, the blue line represents scenarios without data augmentation and using just original available images in each configuration of experiments, demonstrating that objective quality incrementally rises with an increase in the original views for all datasets examined. The maximum quality of the rendered images is in this configuration with all original available images. And the worst quality happens when there are a minimum number of original images available.

More crucially, the incorporation of synthesized views using D_L acquired with available images in each configuration, as represented by the orange line (indicative of data augmentation), substantially boosts the quality across various subsets of the ULB Toys Table dataset, aligning the results more closely with the best possible outcomes which belong to all original available images. This quality increase is approximately 11 dB for 2 ×2, and 3 × 3 configurations.

For the UPM dataset, this approach helps with the 2 × 2 configuration but not for the rest. This quality increase is approximately 5 dB for 2 × 2 configuration for both DA_1 and DA_2. For other configurations of this dataset, there is a quality decrease (6 dB). This is caused by the artifact of the augmented images due to the low quality generated depth maps.

On the other hand, the Garage dataset exhibits a unique behavior; while data augmentation enhances quality for the 2 × 2 and augmented subsets using D_H, around 1 dB, it fails to yield similar improvements in other experimental setups. This discrepancy could likely be due to the lower quality synthesized images' and greater noise levels, possibly resulting from depth map quality, especially considering the dataset's baseline size and depth of objects.

For both NeRF benchmarking datasets, two experiments were conducted. One with the minimum amount of original images, and another using all original images available. Results in both datasets demonstrate a comprehensive increase in quality, 7 dB for the Orchid dataset, and 8 dB for the Fortress dataset using DA_1, and 6 dB for the orchid dataset and 2 dB for the Fortress dataset using DA_2 approach.

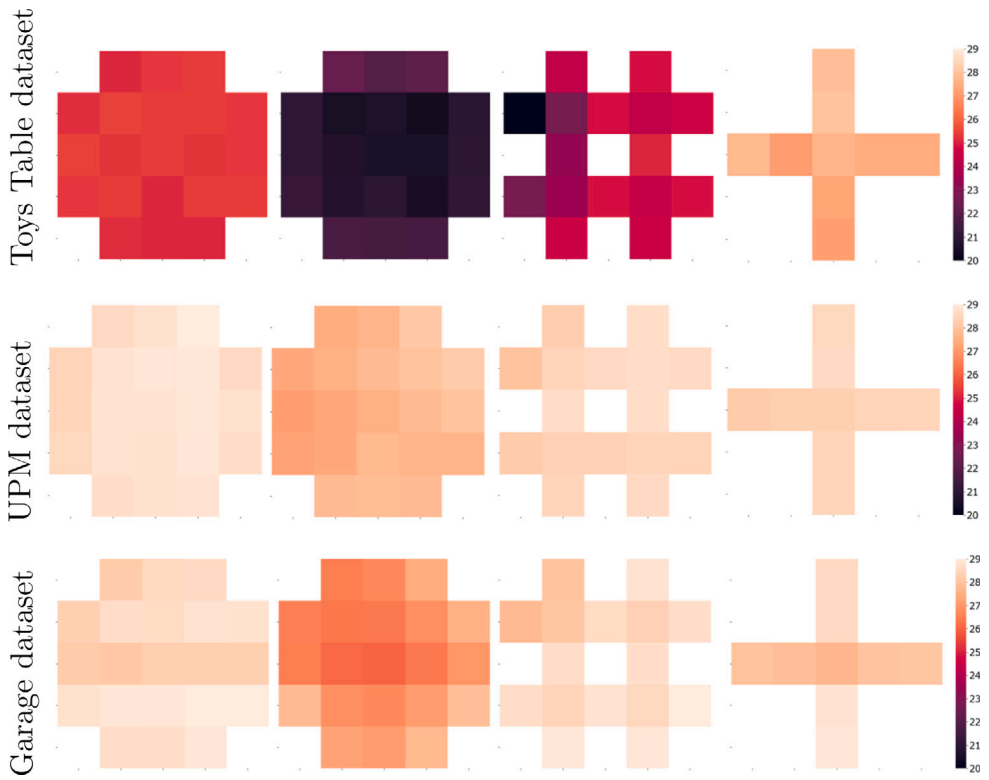

Fig. 5. Heatmaps of augmented images using D_H, and D_L, and available depth maps for ULB dataset (top row), UPM dataset (middle row), and ETRI dataset (bottom row), in each row from left heatmap for augmented images based on: D_H, 4 × 4, 3 × 3, and 2 × 2 configurations.

Table 1
PSNR (dB) of the rendered views for different datasets with different configurations, Δ (original and DA_2) represents the difference of PSNR between synthesized images using just original images with using augmented images using D_H.

Dataset	ULB		UPM		ETRI		Orchid		Fortress	
	V_0	V_{12}	V_0	V_{12}	V_0	V_{12}	V_0	V_{12}	V_0	V_{12}
2 × 2 original	9.3997	7.9484	15.7232	16.5038	27.7962	26.9812	11.2032	11.4837	17.9764	17.9276
2 × 2 + DA_1	**20.9295**	**23.7172**	**23.2256**	**23.3347**	**28.0849**	**27.9023**	**19.3729**	**19.2861**	**25.9875**	**25.3657**
2 × 2 + DA_2	20.7085	20.5553	22.9703	23.0747	27.0182	26.0962	17.9765	16.1535	19.965	19.5241
Δ	11.3088	11.3088	7.5024	11.3088	0.2887	0.9211	8.1697	7.8024	8.0111	7.4381
3 × 3 original	10.1339	9.4127	**30.6063**	**31.2085**	**33.0976**	**34.5235**	–	–	–	–
3 × 3 + DA_1	**21.2753**	**24.2113**	23.5323	23.6884	29.6938	29.1556	–	–	–	–
3 × 3 + DA_2	21.1718	23.7248	23.382	23.6342	29.5615	29.3973	–	–	–	–
Δ	11.1414	14.7986	−7.074	−7.5201	−3.4038	−5.3679	–	–	–	–
4 × 4 original	**22.358**	25.1364	**31.0326**	**31.5676**	**34.7867**	**34.3756**	–	–	–	–
4 × 4 + DA_1	21.9503	**25.3178**	23.6616	23.6682	33.323	30.0184	–	–	–	–
4 × 4 + DA_2	21.0965	25.2238	23.7469	23.5852	33.4814	29.2804	–	–	–	–
Δ	−0.4077	0.1814	−7.371	−7.8994	−1.4637	−4.3572	–	–	–	–
5 × 5 original	**22.9636**	**26.5057**	**31.0853**	**31.8266**	**34.8668**	**35.8774**	**22.9363**	**23.2146**	**29.3476**	**27.5579**

From a subjective viewpoint, the evaluations offer profound insights. A comparative analysis between the initial experimental series, utilizing purely original images (shown in Figs. 7, 11-left, 12-left) and the subsequent series, incorporating both original and synthesized images, show improvements.

The results for the first series of experiments based on rendered images using the original available images show that using minimum available images has the worst results, as was mentioned in objective results. This is the case that the training images are limited and is the main case for data augmentation. The inclusion of augmented images enhances the quality of synthesized target views, with notable reductions in blur and sharper edges. This improvement is visually shown in Fig. 8 for the ULB Toys table dataset, in Fig. 9 for the UPM dataset, in Fig. 11 for the Orchid dataset, and in Fig. 12 for the Fortress dataset, but **NOT** in Figs. 10 for the ETRI Garage dataset, expressing zoomed-in sections of the test image rendering. Based on the objective findings, enhancements in image quality are observed in the ULB and UPM datasets. Yet, for the Garage dataset, subjective assessments do not indicate significant differences, verifying the objective and subjective analysis. This happened because of the object's depth which is related to the disparity between images. It shows for NeRF, where the disparity between the images is low, data augmentation does not help. This consistency suggests that the synthesized images' quality and attributes critically impact the NeRF model's performance across varied dataset scenarios.

The experimental results show that data augmentation can provide better learning of the 3D structure of scenes in NeRF training. Therefore, it can be known that by employing the proposed data augmentation to provide sufficient information about the 3D structure, good results can be achieved when synthesizing novel viewpoints using NeRF.

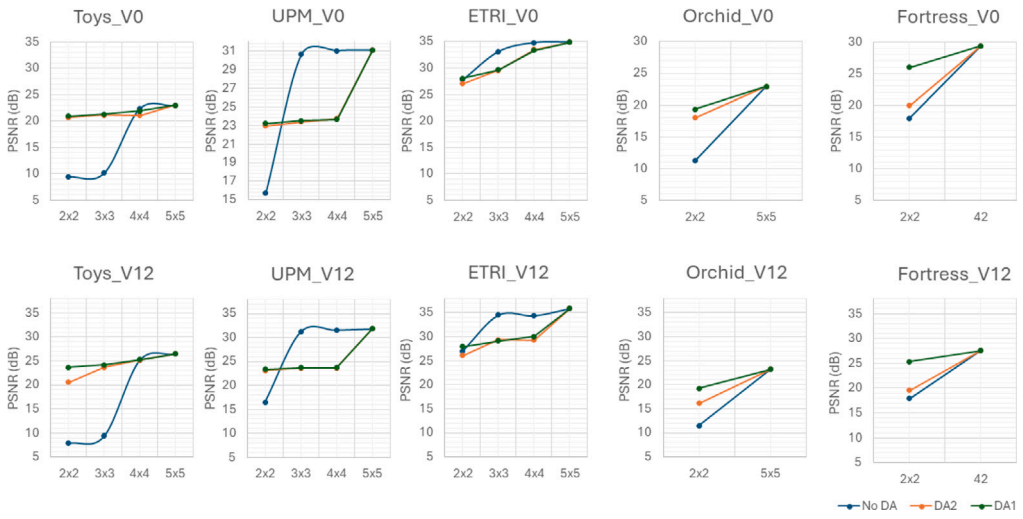

Fig. 6. Objective Results, the blue line demonstrates the training model just with available original images, the orange line shows adding augmented images using depth maps generated with available images, and the green line shows the results for adding augmented images using D_H.

Fig. 7. ULB Toys table dataset subjective results (left column), UPM dataset (middle column subjective results), and ETRI Garage subjective results (right column), from top to down: Original image, rendered v0 for 2 × 2 configuration, 3 × 3 configuration, 4 × 4 configuration, and 5 × 5 configuration.

Fig. 8. NeRF rendered v0 (ULB Toys Table dataset) for 2 × 2 configuration (left column), for 3 × 3 configuration (middle column), and for 4 × 4 configuration (right column), from top to bottom: original images, original + augmented using available depth maps, original + augmented using D_H.

Fig. 9. NeRF rendered v0 (UPM dataset) for 2 × 2 configuration (left column), for 3 × 3 configuration (middle column), and for 4 × 4 configuration (right column), from top to bottom: original images, original + augmented using available depth maps, original + augmented using D_H.

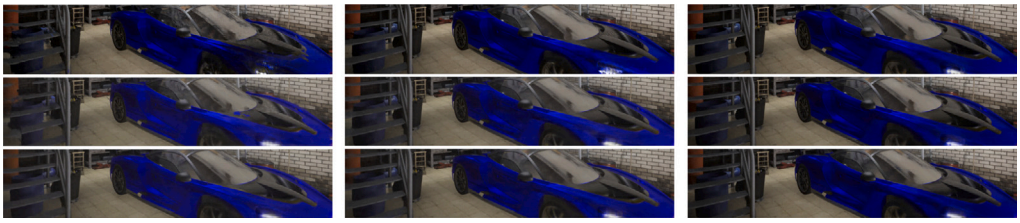

Fig. 10. NeRF rendered v0 (Garage dataset) for 2 × 2 configuration (left column), for 3 × 3 configuration (middle column), and for 4 × 4 configuration (right column), from top to bottom: original images, original + augmented using available depth maps, original + augmented using D_H.

Fig. 11. Left column: NeRF rendered v0 (Orchid dataset) for 2 × 2 configuration, top: original images, middle: rendered view using original 2 × 2 configuration, bottom: rendered view using all original images. Right column: NeRF rendered v0 (Orchid dataset) for 2 × 2 configuration, top: original images, middle: original + augmented using available depth maps, bottom: original + augmented using D_H.

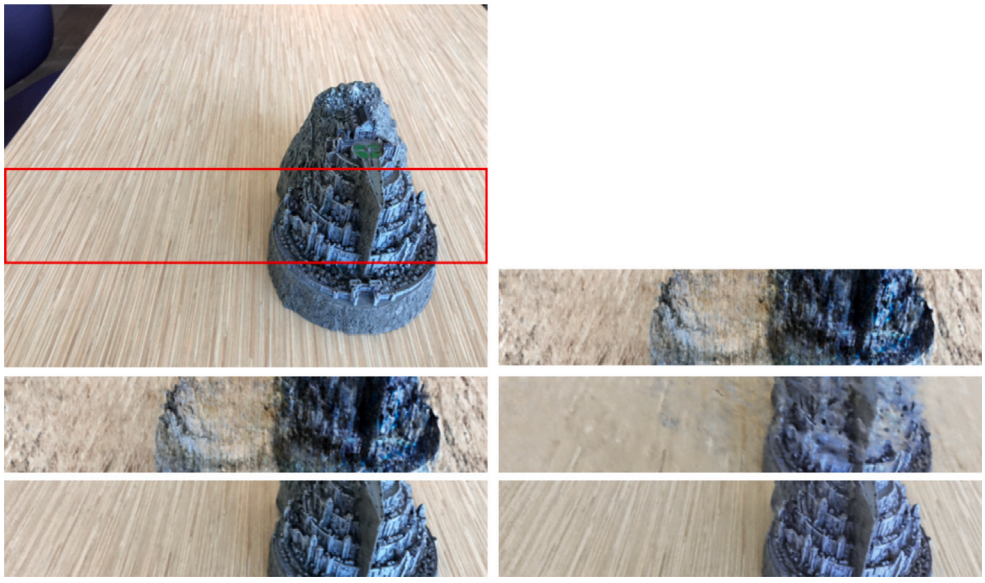

Fig. 12. Left column: NeRF rendered v0 (Fortress dataset) for 2 × 2 configuration, top: original images, middle: rendered view using original 2 × 2 configuration, bottom: rendered view using all original images. Right column: NeRF rendered v0 (Fortress dataset) for 2 × 2 configuration, top: original images, middle: original + augmented using available depth maps, bottom: original + augmented using D_H.

Fig. 13. Subjective results for Orchid dataset (top row), and Fortress dataset (bottom row), from left to right: trained with standard NeRF, trained with DSNeRF, trained with DA1 method, trained with DA2 method.

Table 2
PSNR (dB) of the rendered test set for Orchid and Fortress datasets for 2 × 2 configuration using standard NeRF, DSNeRF, DA1 method, and DA2 method. The last row shows the quality of the rendered images based on training standard NeRF using all original ground truth images.

Dataset	Orchid		Fortress	
	V_0	V_{12}	V_0	V_{12}
NeRF	11.2032	11.4837	17.9764	17.9276
NeRF + DA_1	17.9765	16.1535	19.965	19.5241
NeRF + DA_2	**19.3729**	**19.2861**	**25.9875**	**25.3657**
DSNeRF	11.0741	10.8955	23.8515	20.9667
5 × 5 original	**22.9363**	**23.2146**	**29.3476**	**27.5579**

7.3. Benchmark comparison

This section assesses the perceptual quality of novel view synthesis on the Fortress and Orchid datasets using three approaches. In this section, we evaluate the perceptual quality of novel view synthesis on the Fortress and Orchid datasets using three approaches. We compare three distinct methods: (1) the standard NeRF, (2) NeRF enhanced with sparse depth supervision, referred to as DSNeRF, and (3) NeRF integrated with our proposed data augmentation technique.

The results of this comparative analysis are presented in Table 2 and Fig. 13, under uniform experimental conditions. According to Table 2, our method shows an enhancement in the mean PSNR of the standard NeRF by 8 dB with the DA1 technique and 4 dB with the DA2 technique. Furthermore, it outperforms DSNeRF by 7 dB with the DA1 method and 3 dB with the DA2 method.

Based on the results, we quantified the computational time required for each process. The standard NeRF takes approximately 4 h to train the model and render the target views. In comparison, DSNeRF requires approximately 12 h to complete. Our method, however, demonstrates more time efficiency: the DA1 approach takes about 10 h, while the DA2 approach completes in approximately 5 h, encompassing depth estimation, model training, and target rendering. These findings suggest that our approach not only consumes less time on average but also enhances quality.

While DA4NeRF significantly enhances the capability of Neural Radiance Fields to generate photorealistic views from sparsely sampled data, it does present certain limitations. Key among these is the reliance on high-quality depth data, where inaccuracies can propagate through the augmentation process, potentially degrading the model output. Moreover, the computational demands of our approach may restrict its application in real-time scenarios or on devices with limited processing capabilities. Future work will need to address these challenges, possibly through the adoption of more efficient computational strategies and broader testing across diverse and dynamic environments to ensure the technique's robustness and scalability. (C5.R2.S2)

8. Conclusion

In summary, our research introduces a data augmentation pipeline designed for Neural Radiance Fields, addressing the challenge of insufficient training images for better scene reconstruction (briefly shown in Table 3). By synthesizing non-existent images from depth maps, either sourced from existing datasets or created through depth estimation software, we expand the training dataset available for the NeRF framework. This method effectively enriches the training data by using depth map information. Our experiments demonstrate that this approach significantly enhances the quality of images rendered by NeRF, particularly when working with very sparse datasets.

The dataset characteristics, particularly the baseline distance and the depth of objects, emerge as pivotal factors influencing the augmentation's success. Our findings explain that datasets characterized by larger disparities between images benefit more from our augmentation method, whereas NeRF inherently performs well with datasets exhibiting minimal disparities. Notably, the ETRI Garage dataset did not necessitate our augmentation approach, but our approach outperformed the rest of the datasets.

In conclusion, the DA4NeRF approach, integrating depth-aware augmentation techniques for Neural Radiance Fields, promises significant advancements in the rendering of photorealistic scenes from sparse datasets. Our methodology not only enhances the generalizability and accuracy of NeRF models but also sets a framework for future explorations into efficient data augmentation strategies in 3D rendering fields. The potential applications of this research span a broad array of technologies including virtual reality, augmented reality, and autonomous vehicle navigation, where accurate and detailed environmental representations are crucial. Our work, not only contributes to the theoretical and methodological growth in computer graphics but also paves the way for innovative practical applications that leverage deep learning for enhanced visual computing tasks. (C5.R2.S1)

Moreover, the quality of depth maps stands out as a critical variable affecting performance across different datasets. To forge a path toward a more nuanced understanding of synthetic datasets, future studies should incorporate an analysis of ground truth depth maps. Such an exploration is anticipated to unlock further optimizations of the NeRF model, tailoring it more effectively to diverse dataset characteristics. All these numbers are approximate and averaged over several datasets and several experiments were done on a single NVIDIA RTX-3090 GPU.

Our DA4NeRF technique was evaluated across a diverse array of datasets, including structured synthetic environments and complex real-world scenes. This diversity not only tests the robustness of our approach under varied conditions but also demonstrates its potential applicability across different domains. (C6.R2.S1)

Table 3
Summary of experiment results. PSNR (dB) of the rendered V_0 for sparse views, and the impact of adding augmented images to training set. C4.R2.S1.

Dataset	ULB	UPM	ETRI	Orchid	Fortress
Sparse Images	9.3997	15.7232	27.7962	11.2032	17.9764
Orig + DA	20.9295	23.2256	28.0849	19.3729	25.9875
Original 5 × 5	22.9636	26.5057	34.8668	22.9363	29.3476

CRediT authorship contribution statement

Hamed Razavi Khosroshahi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Jaime Sancho:** Writing – review & editing, Methodology, Investigation, Data curation. **Gun Bang:** Validation, Data curation. **Gauthier Lafruit:** Writing – review & editing, Supervision. **Eduardo Juarez:** Writing – review & editing, Supervision. **Mehrdad Teratani:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Wordtune in order to improve readability of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hamed Razavi Khosroshahi reports financial support and equipment, drugs, or supplies were provided by ARIAC by DIGITALWALLONIA4.AI. Gun Bang reports financial support was provided by Korea government (MSIT). Mehrdad Teratani reports financial support was provided by Emile DEFAY 2021. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Service Public de Wallonie Recherche under grant n° 2010235 – ARIAC by DIGITALWALLONIA4.AI; in part by the FER 2021 project (n∘ 1060H000066-FAISAN); in part by the Emile DEFAY 2021 project (n∘ 4R00H000236); in part by the FER 2023 project (n∘ 1060H000075), Belgium; and in part by Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera media; in part by the project AIMS5.0, which is supported by the Chips Joint Undertaking and its members, including the top-up funding by National Funding Authorities from involved countries under grant agreement no. 101112089; and in part by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00072).

Data availability

Data will be made available on request.

References

[1] M. Xu, S. Yoon, A. Fuentes, D.S. Park, A comprehensive survey of image augmentation techniques for deep learning, Pattern Recognit. 137 (C) (2023) http://dx.doi.org/10.1016/j.patcog.2023.109347.

[2] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., 2012, URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[3] X. Cui, V. Goel, B. Kingsbury, Data augmentation for deep neural network acoustic modeling, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (2014) 5582–5586, http://dx.doi.org/10.1109/ICASSP.2014.6854671.

[4] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 779–788, http://dx.doi.org/10.1109/CVPR.2016.91.

[5] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2022) 3523–3542, http://dx.doi.org/10.1109/TPAMI.2021.3059968.

[6] X. Chen, Z. Song, J. Zhou, D. Xie, J. Lu, Camera and LiDAR fusion for urban scene reconstruction and novel view synthesis via voxel-based neural radiance fields, Remote Sens. 15 (18) (2023) http://dx.doi.org/10.3390/rs15184628.

[7] Z. Bao, G. Liao, Z. Zhao, K. Liu, Q. Li, G. Qiu, 3D reconstruction and new view synthesis of indoor environments based on a dual neural radiance field, 2024, arXiv:2401.14726.

[8] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, Nerf: representing scenes as neural radiance fields for view synthesis, Commun. ACM 65 (1) (2021) 99–106, http://dx.doi.org/10.1145/3503250.

[9] J.F. Hughes, A. van Dam, M. McGuire, D.F. Sklar, J.D. Foley, S. Feiner, K. Akeley, Computer Graphics: Principles and Practice, third ed., Addison-Wesley, Upper Saddle River, NJ, 2013.

[10] N. Max, Optical models for direct volume rendering, IEEE Trans. Vis. Comput. Graphics 1 (2) (1995) 99–108, http://dx.doi.org/10.1109/2945.468400.

[11] M. Levoy, P. Hanrahan, Light field rendering, in: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '96, Association for Computing Machinery, New York, NY, USA, 1996, pp. 31–42, http://dx.doi.org/10.1145/237170.237199.

[12] B. Mildenhall, P.P. Srinivasan, R. Ortiz-Cayon, N.K. Kalantari, R. Ramamoorthi, R. Ng, A. Kar, Local light field fusion: practical view synthesis with prescriptive sampling guidelines, ACM Trans. Graph. 38 (4) (2019) http://dx.doi.org/10.1145/3306346.3322980.

[13] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, http://dx.doi.org/10.1109/CVPR.2016.445.

[14] O. Özyeşil, V. Voroninski, R. Basri, A. Singer, A survey of structure from motion, Acta Numer. 26 (2017) 305–364, http://dx.doi.org/10.1017/S096249291700006X.

[15] M. Poggi, A. Conti, S. Mattoccia, Multi-view guided multi-view stereo, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2022, pp. 8391–8398, http://dx.doi.org/10.1109/IROS47612.2022.9982010.

[16] H. Zhang, C. Wang, S. Tian, B. Lu, L. Zhang, X. Ning, X. Bai, Deep learning-based 3D point cloud classification: A systematic survey and outlook, Displays 79 (2023) 102456, http://dx.doi.org/10.1016/j.displa.2023.102456.

[17] J. Choe, B. Joung, F. Rameau, J. Park, I.S. Kweon, Deep point cloud reconstruction, 2022, arXiv:2111.11704.

[18] C. Buzón, A. Perez-Romero, J.L. Castro, I. Ben Jerbania, F. Manzano-Agugliaro, Photogrammetry as a new scientific tool in archaeology: Worldwide research trends, Sustainability 13 (2021) 5319, http://dx.doi.org/10.3390/su13095319.

[19] S. Rogge, D. Bonatto, J. Sancho, R. Salvador, E. Juarez, A. Munteanu, G. Lafruit, MPEG-i depth estimation reference software, in: International Conference on 3D Immersion (IC3D), IEEE, 2019, http://dx.doi.org/10.1109/IC3D48390.2019.8975995.

[20] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, I. Kweon, Accurate depth map estimation from a lenslet light field camera, 2015, http://dx.doi.org/10.1109/CVPR.2015.7298762.

[21] S. Rogge, I. Schiopu, A. Munteanu, Depth estimation for light-field images using stereo matching and convolutional neural networks, Sensors 20 (21) (2020) http://dx.doi.org/10.3390/s20216188.

[22] T.-C. Wang, A. Efros, R. Ramamoorthi, Occlusion-aware depth estimation using light-field cameras, 2015, pp. 3487–3495, http://dx.doi.org/10.1109/ICCV.2015.398.

[23] S. Fachada, D. Bonatto, M. Teratani, G. Lafruit, View Synthesis Tool for VR Immersive Video, in: D.B. Sobota (Ed.), 3D Computer Graphics, IntechOpen, Rijeka, 2022, http://dx.doi.org/10.5772/intechopen.102382.

[24] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536, http://dx.doi.org/10.1038/323533a0.

[25] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551, http://dx.doi.org/10.1162/neco.1989.1.4.541.

[26] E.R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, G. Wetzstein, Pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis, 2021, arXiv:2012.00926.

[27] M. Niemeyer, A. Geiger, GIRAFFE: Representing scenes as compositional generative neural feature fields, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 11448–11459, http://dx.doi.org/10.1109/CVPR46437.2021.01129.

[28] Z. Li, S. Niklaus, N. Snavely, O. Wang, Neural scene flow fields for space-time view synthesis of dynamic scenes, 2021, http://dx.doi.org/10.1109/CVPR52729.2023.00801, arXiv:2011.13084.

[29] W. Xian, J.-B. Huang, J. Kopf, C. Kim, Space-time neural irradiance fields for free-viewpoint video, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 9416–9426, http://dx.doi.org/10.1109/CVPR46437.2021.00930.

[30] M. Boss, R. Braun, V. Jampani, J.T. Barron, C. Liu, H.A. Lensch, Nerd: Neural reflectance decomposition from image collections, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 12664–12674, http://dx.doi.org/10.1109/ICCV48922.2021.01245.

[31] P.P. Srinivasan, B. Deng, X. Zhang, M. Tancik, B. Mildenhall, J.T. Barron, Nerv: Neural reflectance and visibility fields for relighting and view synthesis, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 7491–7500, http://dx.doi.org/10.1109/CVPR46437.2021.00741.

[32] J. Zhang, X. Liu, X. Ye, F. Zhao, Y. Zhang, M. Wu, Y. Zhang, L. Xu, J. Yu, Editable free-viewpoint video using a layered neural representation, ACM Trans. Graph. 40 (4) (2021) 1–18, http://dx.doi.org/10.1145/3450626.3459756.

[33] S. Liu, X. Zhang, Z. Zhang, R. Zhang, J. Zhu, B. Russell, Editing conditional radiance fields, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 5753–5763, http://dx.doi.org/10.1109/ICCV48922.2021.00572.

[34] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J.T. Barron, R. Martin-Brualla, N. Snavely, T. Funkhouser, Ibrnet: Learning multi-view image-based rendering, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 4688–4697, http://dx.doi.org/10.1109/CVPR46437.2021.00466.

[35] A. Yu, V. Ye, M. Tancik, A. Kanazawa, Pixelnerf: Neural radiance fields from one or few images, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 4576–4585, http://dx.doi.org/10.1109/CVPR46437.2021.00455, URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00455.

[36] M. Tancik, B. Mildenhall, T. Wang, D. Schmidt, P.P. Srinivasan, J.T. Barron, R. Ng, Learned initializations for optimizing coordinate-based neural representations, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, Los Alamitos, CA, USA, 2021, http://dx.doi.org/10.1109/CVPR46437.2021.00287.

[37] A. Jain, M. Tancik, P. Abbeel, Putting NeRF on a diet: Semantically consistent few-shot view synthesis, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 5865–5874, http://dx.doi.org/10.1109/ICCV48922.2021.00583.

[38] K. Deng, A. Liu, J.-Y. Zhu, D. Ramanan, Depth-supervised nerf: Fewer views and faster training for free, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 12872–12881, http://dx.doi.org/10.1109/CVPR52688.2022.01254.

[39] M. Niemeyer, J.T. Barron, B. Mildenhall, M.S.M. Sajjadi, A. Geiger, N. Radwan, RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 5470–5480, http://dx.doi.org/10.1109/CVPR52688.2022.00540.

[40] Z. Yan, C. Li, G.H. Lee, Nerf-DS: Neural radiance fields for dynamic specular objects, 2023, arXiv:2303.14435.

[41] M. Schmeing, X. Jiang, Depth image based rendering, in: P.S.P. Wang (Ed.), Pattern Recognition, Machine Intelligence and Biometrics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 279–310, http://dx.doi.org/10.1007/978-3-642-22407-2_12.

[42] H.R. Khosroshahi, G. Bang, J.L.G. Lafruit, M. Teratani, NeRF for view synthesis using subaperture views of multiview plenoptic 2.0 images, ISO/IEC JTC 1/SC 29/WG 4 m63096, Türkiye, Antalya, 2023.

[43] H.R. Khosroshahi, G. Bang, J. Lee, G. Lafruit, M. Teratani, Impact of number of sub-aperture images in training NeRF, ISO/IEC JTC 1/SC 29/WG 4 m64173, Switzerland, Geneva, 2023.

[44] L. Yen-Chen, Nerf-pytorch, 2020, https://github.com/yenchenlin/nerf-pytorch/.

[45] Manual of IVDE 3.0, ISO/IEC JTC1/SC29/WG4 MPEG2020/ N0058, Online, Jan. 2021, 2021.

[46] D. Bonatto, S. Fachada, G. Lafruit, ULB ToysTable, 2021, http://dx.doi.org/10.5281/zenodo.5055542.

[47] A. Schenkel, D. Bonatto, S. Fachada, H.-L. Guillaume, G. Lafruit, Natural scenes datasets for exploration in 6DOF navigation, in: 2018 International Conference on 3D Immersion (IC3D), IEEE, Brussels, Belgium, 2018, pp. 1–8, http://dx.doi.org/10.1109/IC3D.2018.8657865.

[48] D. Bonatto, A. Schenkel, T. Lenertz, Y. Li, G. Lafruit, [MPEG-I Visual] ULB High Density 2D/3D Camera Array data set, version 2 [m41083], Torino, Italy, ISO/IEC JTC1/SC29/WG11 MPEG2017/M41083, 2017.

[49] B.O. Community, Blender - A 3D Modelling and Rendering Package, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018, URL http://www.blender.org.

[50] D. Apriyanti, L. Spreeuwers, P. Lucas, R. Veldhuis, Orchid flowers dataset-v1.1.zip, in: Orchid Flowers Dataset, Harvard Dataverse, 2020, http://dx.doi.org/10.7910/DVN/0HNECY/GSZICH.