IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 April 2026, accepted 20 May 2026, date of publication 29 May 2026, date of current version 8 July 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3698333

RESEARCH ARTICLE

Comparative Insights Into NeRF 3-D Scalability

EFSTRATIOS KAKALETSIS, GRIGORIOS ARIS CHEIMARIOTIS, PANOS K. PAPADOPOULOS, AND DIMITRIOS ZARPALAS

Center for Research and Technology, Information and Technologies Institute, Hellas, GR57001 Thesssaloniki, Greece

Corresponding author: Efstratios Kakaletsis (ekakalets@iti.gr)

The research leading to these results has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101092875 (DIDYMOS-XR). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein. The publication of the article in OA mode was financially supported by HEAL-Link.

ABSTRACT 3-D reconstruction for large scale indoor and outdoor environments constituted a challenging problem in inverse graphics and computer vision research community. The scalability issue that concerns only the geographic dimensions of the reconstructed scene in terms of scale is a inevitable bottleneck. This could be confronted by state-of-the-art methods which a comparative study should be reveal new insights of the digital environment representation. In this paper, a comparative study is facilitated including few literature methods with accordance by common evaluation metrics and computational complexity. A novel loss function referring to the scene details is also discussed by accessing the density variable of the NeRF multilayer perceptron output as alternative research direction. To this end, it presents efficiency advantange in terms of robustness to noise, scale and sentivity to distortions combined with sufficiency on accessed GP-GPUs memory and footprint in several datasets while it was validated and compared with state-of-the-art Inf-NeRF, Grid-NeRF and Nerfacto techniques.

INDEX TERMS Large scale 3-D reconstruction, Inf-NeRF, grid-NeRF, Nerfacto, density loss function, comparative study.

I. INTRODUCTION

3-D reconstruction for large scale indoor and outdoor environments based on neural implicit techniques is an interesting and rapidly growing field of computer vision. References [1], [2]. In this aspect, Neural Radiance Fields (NeRF) methods [3] are a powerful tool for generation of photorealistic 3-D models of scenes, exploiting 2D images [4] or videos [5]. Efforts towards tackling issues regarding the huge training time of such methods include the introduction of novel input encodings that allow for usage of smaller networks without quality drop [6]. The need of big amount of training data can been also addressed [7], although rendering time is slow and increases linearly with more input views that are necessary towards producing a better-quality output. The combination of Generative Adversarial Networks (GANs) and NeRF [8] can remove the need for accurate camera pose information for the training images and scene generalization can be addressed with the introduction of attention mechanisms [9], however the required training time for a single scene is still substantial. Recent NeRF methods have managed to speed up training and generation of novel views, by descriminating the 3-D volumes of the NeRF in an octree-based radiance field [10]. Time performance in training and rendering time has improved by efficiently sampling points of the scene [11].

Although, the performance gain is achievable with the use of high-end GP-GPUs through them recent developments in computer graphics and computer vision technologies refers to 3-D reconstruction literature foundational work and recent advances in cross-modal reasoning for 360 depth completion [12] and reconstruction for large-scale indoor environment with feature alignment network [13], graph neural networks [14], polyhedron-based graph neural network for 3-D building reconstruction from point clouds [15], including cases of few-shot generalization for single-image 3-D reconstruction via priors [16], or visual question answering for 3-D environments [17], allow creation of virtual environments for XR applications by low cost yet high-fidelity 3-D reconstruction methods and neural methodologies. In this

The associate editor coordinating the review of this manuscript and approving it for publication was Ayman El-Baz.

respect, a comparison study is presented in this paper on recent scientific advances offering a methodology for creating photorealistic 3-D reconstructions for Digital Twins. The aim is to establish a basis for the applicability of state-of-the-art inverse graphics, concretely for baseline NeRF models (e.g. Nerfacto) and latest high quality and performance large scale scenes models (i.e. Inf-NeRF, Grid-NeRF). These models can be ran for both indoor and outdoor enviroments as well as for precise metric estimation of complex Digital Twin for tourism and cities [18], [19]. The main NeRF techniques that are being evaluated in this paper are the NeRFacto [20], Inf-NeRF [21], Grid-NeRF [22] and a proposed one with a novel evaluation loss. In this topic, recent literature often refers to additional loss for learning radiance fields that takes advantage of superpixels texture constraints [23] or combination of explicit Gaussians and neural fields strengths in case of Hybrid Radiance Fields (HyRF) [24] novel scene representation decomposing the scene into set of explicit Gaussians storing only critical high-frequency parameters and grid-based neural fields that predict remaining properties. In this paper, an important and timely topic for scalability of Neural Radiance Fields in large-scale 3-D reconstruction attempts to combine a comparative study with a methodological contribution based on a modified loss function applied to Inf-NeRF. The main contribution relies on the introduction of a density-based loss term that incorporates the gradient and variance of the NeRF density field. It validates the detection of detailed space with more sampling points indicated by novel computation of variance and gradient of the density parameter instead of colour in NeRF's mathematical model [3]. To this end, rendering views gain better quality that are robust to noise, scale and sentivity to distortions as PSNR, SSIM and LPIPs performance shows respectively. Although, such pipelines typically require substantial computational resources, they are state-of-art 3-D reconstruction methodologies that use photogrammetry to create accurate virtual representations of spaces, enriched with the respective textures. However, these methods are selected due to the fact that are easily deployed within a normal computation facility equipped with sufficient required GP-GPUs resources.

The remainder of this paper is organized as follows. In Section II related literature for Neural Radiance Fields (NeRFs) 3-D reconstruction is presented, whereas in Section III we describe the details of proposed method which included in the comparative study. In Section IV experiments conducted to measure the algorithms performance are presented. Finally, Section V provides a discussion and Section VI presents conclusions and future work.

II. RELATED WORK

A. NEURAL RADIANCE FIELDS

NeRFs are known as differentiable novel-view synthesis models as they learn the implicit 3-D representations without 3-D supervision. This implies that every single pixel in the reconstructed render image (ray compound) is a differentiable function of the computerized 3-D world. NeRFs encode for each view and pixel the radiance (RGB values) accounting for its relative position and orientation of the camera viewpoint. Given this input data, the NeRF "Field" uses a fully connected neural network (i.e. Multilayer Perceptron) which is able to infer the volumetric properties (both surface volume and density estimation) of the space by sampling a set of rays for every pixel from the camera to the surface. By rendering the training views the model is able to represent the 3-D scene from novel views (subsequently validated with unseen views) using quality assessment metrics. Here the groundbreaking challenge solved by NeRFs is that these networks only require a set of monocular images, which are subsequently pre-processed with an SfM mechanism (relative transformation of location and rotation of the cameras). From the estimated location and direction of these rays, point clouds can be estimated and therefore a mesh or surface (3-D model) can be reconstructed and later texturized with the radiance values.

The main challenge with NeRFs relies on the computational cost related to sampling and encoding ray projections. As the resolution and number of images increase, both the training time and the required parameters to represent the scene scale accordingly. This process involves the interpretation of rays inside a frustum employing a specific uniform positional encoding as a coarse-fine structure. Mip-NeRF [25] reduces this cost with a versatile frustum which can sample the ray space in a unique integrated encoding network. Here Mip-NeRF interprets the ray space as a set of conical frustums, utilizing an integrated positional encoding. This technique approximates the frustum with a multivariate Gaussian and then computes the integral over the positional encodings of the coordinates within the Gaussian. It was demonstrated that with the same precision Mip-NeRF requires half the parameters (memory) on the same scene in comparison with "vanilla" NeRFs. Mip-NeRF360 [26] is a Mip-NeRF variant that addresses sampling and aliasing, using a non-linear scene parameterization, online distillation, and a novel distortion-based regularizer to overcome the challenges presented by unbounded scenes (both in speed and performance). Instant-NGP [6], employs a smaller input encoding network. This network is enhanced by a multiresolution hash table of trainable feature vectors, whose values are optimized through Stochastic Gradient Descent (SDG). Instant-NGP uses Neural Radiance Caching, which consists of running the MLP independently for each pixel, where feature buffers can be treated as per-pixel feature vectors that contain the 3-D coordinate as well as additional features, therefore applying the multiresolution hash encoding while treating all additional features as auxiliary encoded dimensions to be concatenated with the encoded position. The multiresolution structure allows the network to disambiguate hash collisions. This approach is easier to parallelize on modern Graphic Processing Units (GPUs), achieving the same training performance in

5 minutes while NeRFs require several hours to process the same scene. Zip-NeRF [27] integrates the general framework of Mip-NeRF360 with the featurization method of instant-NGP. Similar to Mip-NeRF, Zip-NeRF assumes each pixel corresponds to a cone. For a given interval along the ray, it constructs a set of multisamples to approximate the shape of the conical frustum. Additionally, it introduces an alternative loss function that, unlike Mip-NeRF 360's interlevel loss, is continuous and smooth with respect to the distance along the ray, thereby preventing z-aliasing. In addition to these challenges, NeRF-like architectures are designed to overfit a specific scene, therefore one entire network is necessary to reconstruct a single scene. This means that nor patterns nor distillation can be done with these networks, as these learn space-dependent (thus, scene-dependent) features.

Similarly to NeRFs, 3-D Gaussian Splatting (3DGS) [28] can reconstruct realistic views using oriented gaussians as a distinct mechanism (instead of rays) in order to represent the space volumetry with optimized details (non-uniform projections). Also, it is optimizing anisotropic covariance to achieve an accurate representation of the scene. While NeRFs are able to render point samples from the underlying data, point sample rendering suffers from holes, causes aliasing, and is strictly discontinuous. 3DGS addresses these issues by ''splatting'' point primitives with an extent larger than a pixel [29] as gaussians, e.g., circular or elliptic discs, ellipsoids, or surfels [30], [31], [32]. This algorithm is known to remain a good trade-off between training speed and quality.

Considering the vast amount of monocular image data being captured and shared online and the potency of deep learning with GPU parallelization, NeRFs and 3DGS algorithms are currently in a peak of research interest [33], [34] as have shown to perform with impressive results in distinct areas from automotive, robotics, manufacturing, architecture, marketing, and entertainment industries.

B. LARGE SCALE 3-D RECONSTRUCTION

Several 3-D reconstruction methods for large scale outdoor and indoor environments utilize NeRFs for the 3-D representation of the Digital Twins for smart cities and tourism. The following several works of recent state-of-the-art literature concerning 3-D reconstruction using Neural Radiance Fields for large scale scenes and environments are indicated. To begin with, method Inf-NeRF [21] is working with core a LoD structure using octrees where each node represents a specific cubic space of the scene, known as the node's Axis-Aligned Bounding Box (AABB). It is being executed efficiently in NVIDIA 3090 RTX by training a Inf-NeRF MLP focused on a large scale scene on platform of Nerfstudio. Method SMERF [35] utilizes distillation network teacher-student model by training MLP NeRF partitions of a respective large scale 3-D reconstruction NeRF representation. In same manner, Block-NeRF [36] utilizes large-scale scenes spanning multiple blocks that can be optimized independently. At such a scale, the data collected will necessarily have transient objects and variations in appearance for scaling the rendering of the NeRF 3-D scene. It should be noted that Mega-NeRF [37] introduces a sparse and spatially aware network structure along with a simple geometric clustering algorithm that partitions training pixels into different NeRF submodules which can be trained in parallel. A NeRF with learnable scene decomposition presented in publication Switch-NeRF [38] achieves both high-fidelity scene reconstruction and efficient computation. A compact multi-resolution ground feature plane representation to coarsely capture the scene, and complement it with positional encoding inputs through another NeRF branch for rendering in a joint learning fashion is described in Grid-NeRF [22] running in GP-GPUs for efficient performance. Finally, applying the alternating direction method of multipliers (ADMM) to achieve consensus among camera poses while maintaining parallel tile optimization, the ScaNeRF [39] partition scenes into tiles. Each with a multi-resolution hash feature grid and shallow chained diffuse and specular multi-layer perceptrons (MLPs) and decomposes the appearance with the specular MLP allowing a specular-aware warping loss to provide a second optimization path for camera poses.

III. PROPOSED METHOD

Inf-NeRF [21] is selected to be deployed in experimental evaluation for current task of large scale 3-D reconstruction methodology. This method employs a LoD tree to divide the scene in space and scale into cubes, which are reconstructed using many NeRFs. During rendering, Inf-NeRF includes selection retrieval of a minimal subset of nodes, significantly reducing memory requirements and I/O time for parameter retrieval from disk or cloud storage and also reducing aliasing artifacts. As Inf-NeRF is the indicated method for this task, baseline Nerfacto [20] and other large scale 3-D reconstruction methods as Grid-NeRF [22] and a proposed one accompanied with evaluated novel loss function are included in complementary comparison described as follows in Section III.

Method Inf-NeRF [21] can significantly improve both its efficiency and accuracy by concentrating computational effort on high-detail or high-error areas, implementing the adaptive sampling in octrees. To this end, a step-by-step plan tailored for adaptive sampling in Inf-NeRF's octree structure is proposed. The aim is to use adaptive sampling to selectively refine or subdivide the octree where the scene complexity, rendering error, or ray contribution is high — and sample less where the scene is simpler or well-approximated. Proposed method's novelty is approached on qualitative process of detailed space definition aiming to improve the grid size by subdivision. More specifically, defining the variance and gradient scores of density parameter and training the Inf-NeRF MLP using the proposed loss function that includes the additional parameter L_{dens} discussing also the case to connection to prior density regularizer, the qualitative and quantitative performance aim to to be sufficient for more

subdivided octrees. Covering the issue of insufficiency to translate grid features of complex scenes (i.e., Grid-NeRF) which results to inaccurate geometry and missing large amount of scene details or maintaining the octree size (i.e, Inf-NeRF). More specifically, an adaptive sampling pipeline for Inf-NeRF's ctree is developed. Firstly, defining refinement criteria, this stage includes chosen metrics to decide where to sample more densely: the gradient of the radiance field density indicates the large changes implying high detail (e.g. object edges). Secondly, a score for each octree node derived for each octree node (voxel) accumulating metrics such as absolute value of mean gradient (i.e., $|\nabla(\sigma)|$) and variance of density. Adding the ℓ_1 [40], [41] of these accumulated scores and filtered with an exponential decay in a window during training, a novel density loss for evaluation stage is defined. This sampling density loss is fused with the distortion loss, regularization loss and transparency loss with efficient weights with the rgb loss constituting the total loss function as described by the following equations:

$$L_{total} = L_{RGB} + w_1 * L_{distort} + w_2 * L_{reg} + w_3 * L_{trans} + w_4 * L_{dens} \quad (1)$$

where L_{dens} noted as follows:

$$L_{dens} = \ell_1(exp(\nabla(\sigma))) + \ell_1(exp(var(\sigma))) \quad (2)$$

which uniformly samples points within the AABB and uses ℓ_1 loss function to encourage their density to approach 0.

A prior density regularizer is also proposed aiming to exponential/logarithmic softening of large gradient and variance density (σ) scores. To this end, the following Gaussian regularizer is proposed by replacing the L_{dens} term of the total loss function with a $\mathbf{L}_2$ of density sampling points subtracted by (μ_σ):

$$L_{dens} = \ell_2(\sigma - \mu_\sigma) \quad (3)$$

where the w_4 fusion total loss parameter noted as follows:

$$w_4 = \frac{1}{2\sigma_0^2} \quad (4)$$

that indicates smaller w_4 parameter values for stronger regularization instead of larger w_4 values which achieves weaker regularization of the adaptive sampling-based refinement. The photorealistic representation of NeRF's mesh/pointcloud that colour and density parameters both are without uncertain sampling of scene details and filter outliers, artifacts and noise.

IV. EXPERIMENTAL EVALUATION

Firstly, COLMAP [42] features describing the extrinsic and intrinsic camera parameters, are extracted for each one of the dataset images. Experiments are executed into Nerfstudio [43] software environment or in explicitly created python environment using 2 GP-GPUs Nvidia RTX 3090 on Linux operating system ensuring its reproducibility and verification.

TABLE 1. Configuration Training Parameters.

	Nerfacto	Inf-NeRF	Proposed
l_r	0.001	0.0001	0.01
patch-size	1	1	1
num-rays-per-batch	4096	4096	4096
eval-num-rays-per-batch	4096	16384	16384
train-num-per-batch	4096	65536	65536
distortion-loss-mult	0.002	0.002	0.002
prop-interlevel-loss-mult	1.0	1.0	0.01
transparency-loss-mult	-	0.001	0.001
density-loss-mult	-	-	0.002

Defining the training parameters (e.g. Table 1), a comparative study is conducted between large scale 3-D reconstruction methods Inf-NeRF [21] and Grid-NeRF [22] with the small scale 3-D reconstruction method nerfacto [20] as well as the proposed one, for comparison purposes. These methods are trained after parameter fine-tunning on input arguments. The study is conducted in the following datasets including several large scale cases such as indoor environment (FactoryDT-dataset-v1) and outdoor scenes (i.e., VNG-real-dataset-v2 [44], Residence [45], VNG-real-dataset-v1).

- VNG-real-dataset-v1: The VNG dataset contains real-world outdoor scenes from Vilanova, Barcelona, captured by a car equipped with a sparse 16-line LiDAR and four fisheye cameras. This dataset is considerably more challenging than the KITTI360 [46] benchmark due to several factors: a large part of the camera's view is obstructed by the car, the recordings exhibit rolling shutter effects, and the camera annotations are partly inaccurate because they rely on poor GPS data.
- VNG-real-dataset-v2 [44]: The VNG-real-dataset-v2 includes around 70 short video clips, divided into three recording sessions with time duration 1 hour and 45 minutes obtained a new multi-sensor dataset and a surround-view system, LiDAR, and odometry, captured by 5MPx Basler camera. This camera uses a higher-resolution sensor with a lower field of view (FOV) and a global shutter. This enables higher quality with less fisheye distortion and movement distortion. While these images are not representative of the technologies implemented in current private cars, they allow us to validate the reconstruction algorithms in an ideal scenario, without being limited by the cameras.
- FactoryDT-dataset-v1 (FactoryDT): Using the autonomous mobile robot of IW, (i.e., the iw.hub), different datasets were gathered in an industrial environment. The datasets were collected over multiple runs of the iw.hub, also reflecting real-world scenarios that the robot is exposed to. The iw.hub is equipped with an onboard camera, the LIPSedge AE400 3-D Stereo Camera, and 2 LiDARs. Images are captured through the iw.hub's Stereo camera, and the robot localization relies on indoor LiDAR-based positioning. The ground truth odometry provided in the dataset is the output of the

TABLE 2. Comparative study of large scale 3-D reconstruction (best performed metric in bold compared to Inf-NeRF and Nerfacto).

	VNG-real-dataset-v2 [44]		FactoryDT-dataset-v1		VNG-real-dataset-v1	Residence [45]
	Inf-NeRF	Proposed	Inf-NeRF	Proposed	nerfacto	nerfacto
mean DISTS ↓	**0.105**	0.109	**0.143**	0.149	0.256	0.464
mean LPIPS ↓	**0.353**	0.358	**0.431**	0.440	0.623	0.722
mean MSGMSD index ↑	0.083	**0.085**	**0.132**	0.126	0.249	0.277
mean VSI index ↑	**0.989**	0.988	0.968	**0.970**	0.893	0.810
mean GMSD index ↑	0.092	**0.093**	**0.135**	0.130	0.243	0.270
mean MSSSIM index ↑	**0.925**	0.920	0.885	**0.898**	0.446	0.080
mean SSIM index ↑	**0.909**	0.904	0.865	**0.879**	0.375	0.064
mean PSNR index ↑	**29.33**	29.03	24.377	**24.91**	15.59	10.70

TABLE 3. Comparative study of large scale 3-D reconstruction (best performed metric in bold).

	VNG-real-dataset-v2 [44]			FactoryDT-dataset-v1		
	↑ PSNR	↑ SSIM	↓ LPIPS	↑ PSNR	↑ SSIM	↓ LPIPS
nerfacto [20]	29.364	0.838	0.396	24.566	0.838	0.284
Inf-NeRF [21]	**29.334**	**0.909**	0.353	24.377	**0.865**	0.431
Grid-NeRF [22]	26.159	0.652	**0.263**	23.695	0.824	**0.298**
Proposed	29.037	0.846	0.358	**24.382**	0.824	0.431

TABLE 4. Computation complexity for NeRF training.

	Time	iterations	GPU memory	memory footprint (num_rays_per_sec)		voxel/octree size	FLOPs
				VNG-real-dataset-v2 [44]	FactoryDT-dataset-v1		
nerfacto [20]	∼20min	30k	∼6GB	104884	608723	2048/-	2.3×10^{16}
Inf-NeRF [21]	∼9h:15min	10k	∼38GB	259420	252976	2048/4681	2.1×10^{16}
Grid-NeRF [22]	∼14h:30min	120k	∼36GB	67108864	67108864	2048/3072	2.4×10^{16}
Proposed	∼12h:19min	10k	∼37GB	151274	252976	2048/4681	2.1×10^{16}

navigation module running on the iw.hub in production. Camera frames and robot poses are synchronized before being recorded to generate the FactoryDT-dataset-v1 dataset.

Quantitative comparison Table 2, 3 includes the performance evaluation of Inf-NeRF, Grid-NeRF and Nerfacto 3-D model renderings from several view points measuring several quality metrics. It includes metrics such as mean PSNR, mean SSIM, mean LPIPS, mean DISTS, mean MSGMSD, mean VSI index, mean GMSD index, mean MSSIM index presented in Table 2. In Table 4, the computation complexity in terms of time and respective iterations for training, memory footprint, GPU utilization, voxel/octree size, and FLOPs per iteration are noted.

Considering an ablation study to compare proposed method's efficiency, Table 5 shows the case that L_{dens} loss function term is omitted. The proposed method's PSNR outperforms in adequate cases (Table 3) as well as the ablation study (Loss Function Term: L_{dens}) shows that the proposed method works efficiently while it presents computational complexity advantage (GPU memory, memory footprint, FLOPs) compared to Inf-NeRF (baseline) and Grid-NeRF, alternatively including faster Nerfacto training.

V. DISCUSSION

Section V approaches new insights of scalability in NeRF's 3-D reconstruction methodologies discussing comparison issues and limitations.

A. COMPARISON ISSUES

Regarding the comparative results in Table 3, the Inf-NeRF method outperforms the other two large scale methods indicating that is the baseline method. The baseline large scale Inf-Nerf outperforms the other large scale methods (Grid-NeRF, proposed) in PSNR and SSIM metrics in VNG-real-dataset-v2 [44] and in SSIM metric in FactoryDT-dataset-v1. The best performed method in terms of PSNR in FactoryDT-dataset-v1 is the proposed one. Regarding the secondary error-based metric, namely the Learned Perceptual Image Patch Similarity (LPIPS) [47], it computes the distance between feature representations of a pre-defined network on multiple levels and is known to match human perception well. A low LPIPS score means that image patches are perceptually similar between ground truth view and a distorted rendered one. To this end, creating NeRFs for large scale scenes, term scalability causes distortion artifacts as shows LPIPS. It should be noted that nerfacto (Table 3) performance

TABLE 5. Failure case analysis: ***Loss function terms ablation study.***

Loss Function Terms					Performance Evaluation		
L_{RGB}	$L_{distort}$	L_{reg}	L_{trans}	L_{dens}	PSNR ↑	SSIM ↑	LPIPS ↓
✓	✓	✓	✓	×	29.334	0.909	0.353
✓	✓	✓	✓	✓	29.037	0.846	0.358

impacts increased evaluation metrics (i.e., LPIPS, PSNR) because its NeRF's structure (point cloud / mesh) does not constitute from blocks with large dimensions instead of a lot discriminative separated blocks with enriched details offered by the resolution of grid/LoD octrees architecture.

The proposed adaptive-sampling-based approach achieves to improve robustness to noise, scale and sensitivity to distortions. Achieving the proposed method motivation, the following issues have been handled: a) Efficient handling of fine grained detail in large scenes is achieved in distorted regions (poor LPIPs performance Table 2), b) efficient sampling in informative regions refers to the loss function additional parameter which is computed in accumulated buffer of sequential octree leafs. c) sufficient density based refinement as proposed method PSNR outperforms (Table 1 - best performed metric in bold) and due to poor performance of LPIPS (Table 2 -best performed metric in bold).

Computational complexity (Table 4) presents limited cost of the proposed method in terms of time while maintaining the number of iterations fact that proves reasonable performance (i.e., training time $\in$ $[9h, 12h]$). In addition, Figures 4, 5 present the training loss convergence until 10k iterations. While maintaining the voxel size, proposed method presents improvement over current octree subdivision approaches as shown in Table 4. The octree size (i.e., current metric of octree subdivision) is constant compared to baseline (i.e., Inf-NeRF) and is increased compared to Grid-NeRF. This limited advantage follows intermediate but adequate GPU memory, memory footprint and most efficient number of FLOPs. Finally, the qualitative results in Figures 2, 3 includes proposed method rendering views as well as COLMAP representations in the Nerfstudio environment for the datasets.

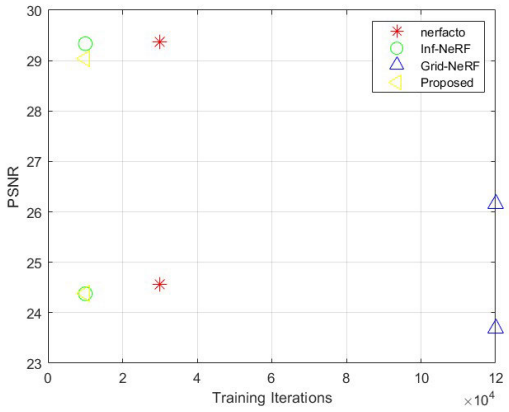


FIGURE 1. What are Proposed Method's Tradeoffs?.

FIGURE 2. Qualitative performance: Inf-NeRF rendering views/ COLMAP from "VNG-real-dataset-v2" dataset in the nerfstudio.

B. LIMITATIONS

Limited GP-GPUs hardware resources inevitably obstruct the NeRF's MLP training and mesh/pointcloud extraction. Only 2 GP-GPUs NVIDIA RTX 3090 are used for the presented comparative study. The number of training input images confirm the reduced memory footprint without pruning the weights or MLP layers or using other lightweight model versions.

The training input parameter fine-tuning is also differentiate the qualitative and quantitative NeRF's results. The scale and front/back boundaries determination also reduce the outliers and noise on rendered NeRF. Without loosing the applicable generalizability across different environment digital representations, NeRFs present adequate and convincing quality performance (Table 2).

C. NEW INSIGHTS

This comparison study offers new insights on large scale 3-D reconstruction concluded in the following issues: i) scalability in terms of scaled space dimensions could be achieved sufficiently even though using limited computational GP-GPUs resources. Robustness to scale presents adequate and sufficient outperformance, ii) the large number of training data (input images) could be confronted by using multitasking parallel programming architectures, iii) using octrees to discriminate the space in grid of volumes leverages scene complexity, rendering error and ray contribution with efficient manipulation of NeRF's density term, iv) noise and outliers could be reduced by density sampling in parts with

FIGURE 3. Qualitative performance: Inf-NeRF rendering views/ COLMAP from "FactoryDT-dataset-v1" dataset in nerfstudio.

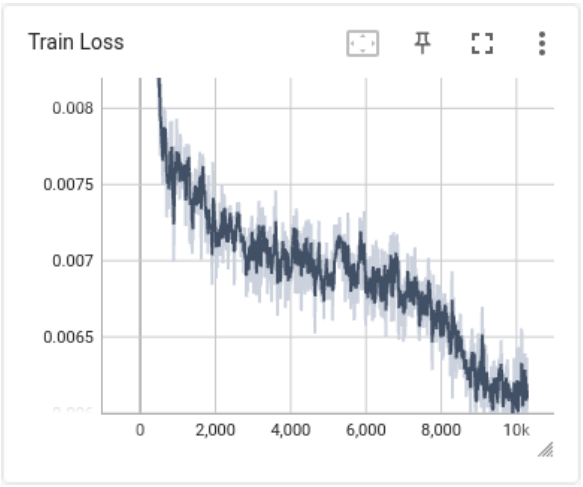


FIGURE 4. Training Loss in "FactoryDT-dataset-v1" dataset during Inf-NeRF MLP training.

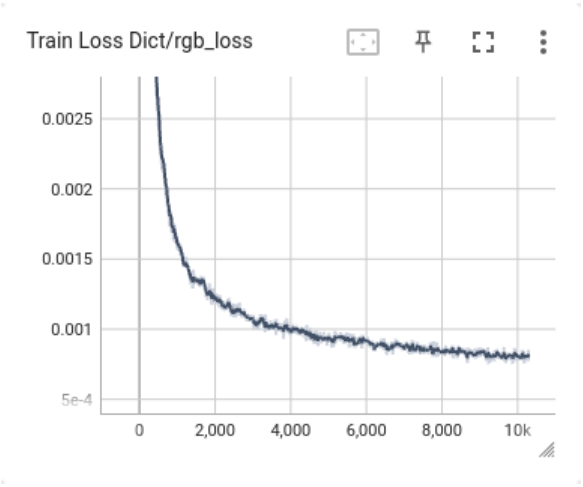


FIGURE 5. Training RGB Loss in "FactoryDT-dataset-v1" dataset during Inf-NeRF MLP training.

high details as PSNR outperformance indicates compared to baseline Inf-NeRF and Nerfacto methods in adequate cases, v) Tradeoffs that are being approached by the adaptive-sampling-based proposed method are robustness to noise, scale and sensitivity to distortions. More specifically, these could be discussed in sensitivity and specificity of best results that validate robustness to noise artifacts in rendering process as PSNR shows photorealism of 3-D models, good fidelity and perceptual realism (see Figure 1).

VI. CONCLUSION

In this paper, the large scale 3-D construction methods Inf-NeRF, Grid-NeRF and the small scale Nerfacto are deployed for comparison purposes with a proposed one which is based to Inf-NeRF method exploiting a novel loss function. The experimental performance evaluation and failure case analysis by ablation studies were presented, prove that the proposed approach outperforms the other large scale methods in limited cases but offering adequate training time. Efficient sampling only in informative regions is noted due to the loss function additional parameter that is computed in accumulated buffer of sequential octree leafs and sufficient density based refinement achieved in limited cases. Future work might include deployment in larger GP-GPUs equipment integrating the adaptive sampling strategy into the Inf-NeRF technique measuring the performance advantage.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101092875 (DIDYMOS-XR). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein. The publication of the article in OA mode was financially supported by HEAL-Link.

REFERENCES

[1] H. Luo, J. Zhang, X. Liu, L. Zhang, and J. Liu, "Large-scale 3-D reconstruction from multi-view imagery: A comprehensive review," *Remote Sens.*, vol. 16, no. 5, p. 773, Feb. 2024.

[2] M. Park, B. Yoo, J. Y. Moon, and J. H. Seo, "InstantXR: Instant XR environment on the Web using hybrid rendering of cloud-based NeRF with 3-D assets," in *Proc. 27th Int. Conf. 3-D Web Technol.*, Nov. 2022, pp. 1–9.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Jan. 2022.

[4] Y.-J. Yuan, Y.-T. Sun, Y.-K. Lai, Y. Ma, R. Jia, and L. Gao, "NeRF-editing: Geometry editing of neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18332–18343.

[5] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe, and Z. Lv, "Neural 3-D video synthesis from multi-view video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5521–5531.

[6] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–15, Jul. 2022.

[7] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "PixelNeRF: Neural radiance fields from one or few images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4576–4585.

[8] Q. Meng, A. Chen, H. Luo, M. Wu, H. Su, L. Xu, X. He, and J. Yu, "GNeRF: GAN-based neural radiance field without posed camera," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6331–6341.

[9] M. M. Johari, Y. Lepoittevin, and F. Fleuret, "GeoNeRF: Generalizing NeRF with geometry priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18344–18347.

[10] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "PlenOctrees for real-time rendering of neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5732–5741.

[11] T. Hu, S. Liu, Y. Chen, T. Shen, and J. Jia, "EfficientNeRF–efficient neural radiance fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12892–12901.

[12] W. Nie, C. Jiao, R. Chang, L. Qu, and A.-A. Liu, "CPG3D: Cross-modal priors guided 3-D object reconstruction," *IEEE Trans. Multimedia*, vol. 25, pp. 9383–9396, 2023.

[13] L. Lai, J. Chen, Z. Zhang, G. Lin, and Q. Wu, "CMFAN: Cross-modal feature alignment network for few-shot single-view 3-D reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 5522–5534, Mar. 2025.

[14] R. Liu, G. Zhang, J. Wang, and S. Zhao, "Cross-modal 360° depth completion and reconstruction for large-scale indoor environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 25180–25190, Dec. 2022.

[15] Z. Chen, Y. Shi, L. Nan, Z. Xiong, and X. X. Zhu, "PolyGNN: Polyhedron-based graph neural network for 3-D building reconstruction from point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 218, pp. 693–706, Dec. 2024.

[16] B. Wallace and B. Hariharan, "Few-shot generalization for single-image 3-D reconstruction via priors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3817–3826.

[17] Y. Etesam, L. Kochiev, and A. X. Chang, "3DVQA: Visual question answering for 3-D environments," in *Proc. 19th Conf. Robots Vis. (CRV)*, May 2022, pp. 233–240.

[18] M. Antoun et al., "Interactive digital twins enabling responsible extended reality applications," *Sci. Rep.*, vol. 15, no. 1, p. 34539, Oct. 2025.

[19] M. Vierimaa, M. Heiskanen, H. Kuppens, A. T. Islam, R. Khan, and S. Mohamed, "Digital twins benefits, challenges and future prospects from intelligent motion control point of view," *Microprocessors Microsystems*, vol. 122, Jun. 2026, Art. no. 105271.

[20] X. Zhang, P. P. Srinivasan, B. Deng, P. Debevec, W. T. Freeman, and J. T. Barron, "NeRFactor: Neural factorization of shape and reflectance under an unknown illumination," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–18, Dec. 2021.

[21] J. Liang, L. Zhang, Z. Zhao, and X. Xu, "InfNeRF: Towards infinite scale NeRF rendering with O(log n) space complexity," in *Proc. SIGGRAPH Asia Conf. Papers*, Dec. 2024, pp. 1–11.

[22] L. Xu, Y. Xiangli, S. Peng, X. Pan, N. Zhao, C. Theobalt, B. Dai, and D. Lin, "Grid-guided neural radiance fields for large urban scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8296–8306.

[23] G. Zhang, C. Xue, and R. Zhang, "SuperNeRF: high-precision 3-D reconstruction for large-scale scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.

[24] Z. Wang and D. Xu, "HyRF: Hybrid radiance fields for memory-efficient and high-quality novel view synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2025, pp. 18321–18344.

[25] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5835–5844.

[26] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-NeRF 360: Unbounded anti-aliased neural radiance fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5460–5469.

[27] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Zip-NeRF: Anti-aliased grid-based neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19640–19648.

[28] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3-D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–139, Aug. 2023.

[29] M. Botsch, A. Hornung, M. Zwicker, and L. Kobbelt, "High-quality surface splatting on today's GPUs," in *Proc. Eurographics/IEEE VGTC Symp. Point-Based Graph.*, Jun. 2005, pp. 17–141.

[30] H. Pfister, M. Zwicker, J. van Baar, and M. Gross, "Surfels: Surface elements as rendering primitives," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2000, pp. 335–342.

[31] L. Ren, H. Pfister, and M. Zwicker, "Object space EWA surface splatting: A hardware accelerated approach to high quality point rendering," *Comput. Graph. Forum*, vol. 21, no. 3, pp. 461–470, Sep. 2002.

[32] M. Zwicker, H. Pfister, J. van Baar, and M. Gross, "EWA volume splatting," in *Proc. Visualizat.*, Oct. 2001, pp. 29–538.

[33] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, "NeRF: Neural radiance field in 3-D vision: A comprehensive review (updated post-Gaussian splatting)," 2022, *arXiv:2210.00379*.

[34] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-time denoising and dereverberation wtih tiny recurrent U-Net," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5789–5793.

[35] D. Duckworth, P. Hedman, C. Reiser, P. Zhizhin, J.-F. Thibert, M. Lučić, R. Szeliski, and J. T. Barron, "SMERF: Streamable memory efficient radiance fields for real-time large-scene exploration," *ACM Trans. Graph.*, vol. 43, no. 4, pp. 1–13, Jul. 2024.

[36] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. P. Mildenhall, P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-NeRF: Scalable large scene neural view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8248–8258.

[37] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly- throughs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12912–12921.

[38] M. Zhenxing and D. Xu, "Switch-NeRF: Learning scene decomposition with mixture of experts for large-scale neural radiance fields," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022, pp. 1–15.

[39] X. Wu, J. Xu, X. Zhang, H. Bao, Q. Huang, Y. Shen, J. Tompkin, and W. Xu, "ScaNeRF: Scalable bundle-adjusting neural radiance fields for large-scale scene rendering," *ACM Trans. Graph.*, vol. 42, no. 6, pp. 1–18, Dec. 2023.

[40] J. Peng, C. Chen, Y. Liu, Y. Fan, and X. Song, "Radiance field reconstruction from noisy multiview images for view synthesis," *Proc. SPIE*, vol. 13653, pp. 192–197, Jul. 2025.

[41] R. Sheffer, C. Pinchover, H. Zisman, D. Ozeri, and R. Litman, "Under-canopy terrain reconstruction in dense forests using RGB imaging and neural 3-D reconstruction," 2026, *arXiv:2601.22861*.

[42] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

[43] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *Proc. ACM Conf. SIGGRAPH*, 2023, pp. 1–12.

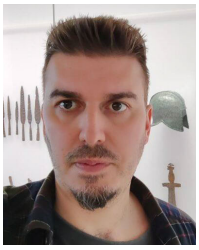
[44] *Dataset for 'Learning Scene Semantics From Vehicle-Centric Data for City-Scale Digital Twins'*, J. R. Ficosa International (Barcelona, Spain), Jul. 22, 2024.

[45] L. Lin, Y. Liu, Y. Hu, X. Yan, K. Xie, and H. Huang, "Capturing, reconstructing, and simulating: The UrbanScene3D dataset," in *Proc. ECCV*, 2021, pp. 93–109.

[46] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3-D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3292–3310, Mar. 2023.

[47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

EFSTRATIOS KAKALETSIS received the Diploma degree from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki (AUTh), Greece, in 2010, the master's degree in informatics and communications (digital media) from the Department of Informatics, AUTh, in 2014, and the Ph.D. degree in informatics from the Artificial Intelligence and Information Analysis Laboratory, AUTh, in 2024, conducting academic research. Since 2024, he has been employed as a Postdoctoral Researcher with the Center for Research and Technology, Information and Technologies Institute, Hellas. He has co-authored more than 15 papers in peer-reviewed international journals and conference proceedings. His current research interests focus on subject areas including computer graphics, computer vision, and machine learning.



GRIGORIOS ARIS CHEIMARIOTIS degree in electrical and computer engineering, the master's degree in medical informatics, and the Ph.D. degree from AUTH, in 2010, 2013, and 2020, respectively, with a dissertation focused on medical image analysis. Since 2014, he has been with AUTH, INAB/CERTH, and DUTH; and ITI, since 2024, contributing to European and national research projects. He has co-authored ten publications in international scientific journals and 19 presentations at international conferences.

PANOS K. PAPADOPOULOS received the B.Sc. and M.Sc. degrees in computer science and biomedical informatics and the Ph.D. degree from the University of Thessaly, in 2015, 2017, and 2020, respectively. He subsequently was a Postdoctoral Researcher on deep learning-based optimizations for video coding and streaming with the Department of Informatics and Telecommunications, University of Thessaly. He is currently a Research Scientist with the Information Technologies Institute (ITI)-CERTH, where he is also a Project Manager in several research projects. He has (co-)authored more than 20 papers in peer-reviewed international journals, conferences, and book chapters and serves as a reviewer for several scientific journals and conferences. His research interests include video coding and compression, computer vision, deep learning, and generative AI.

DIMITRIOS ZARPALAS received the Diploma degree in electrical and computer engineer from the Aristotle University of Thessaloniki (AUTh), the M.Sc. degree in electrical engineering (focusing on computer vision) from The Pennsylvania State University, and the Ph.D. degree in medical informatics from the Health Science School, Department of Medicine, AUTh. He has been a member of the Information Technologies Institute, since 2007. His main research interest are on 3-D/4-D computer vision and machine learning, such as tele-immersion applications: 4-D reconstruction of moving humans, their ''hologram'' compression and transmission in real-time; 3-D motion capturing, analysis, and evaluation; 3-D object recognition and 3-D shape descriptor extraction; 3-D medical image processing, shape analysis of anatomical structures; while in the past has also worked in indexing, search and retrieval and classification of 3-D objects, proteins and 3-D model watermarking. He has (co-)authored more than 75 papers in peer reviewed international journals, conference proceedings, and books (including one IEEE Distinguished Paper and one IEEE Conference Best Paper Award).

The Open Access publication of this article was financially supported by HEAL-Link