IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 July 2023, accepted 7 August 2023, date of publication 8 September 2023, date of current version 13 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3313184

RESEARCH ARTICLE

Complex-Motion NeRF: Joint Reconstruction and Pose Optimization With Motion and Depth Priors

HYUNJIN KIM[1], (Student Member, IEEE), DAEKYEONG LEE[1], SUYOUNG KANG[2], (Student Member, IEEE), AND PYOJIN KIM[3], (Member, IEEE)

[1]Department of Mechanical Systems Engineering, Sookmyung Women's University, Seoul 04310, South Korea
[2]Department of Electronics Engineering, Sookmyung Women's University, Seoul 04310, South Korea
[3]School of Mechanical Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, South Korea

Corresponding author: Pyojin Kim (pjinkim@gist.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant NRF-2021R1F1A1061397.

ABSTRACT We present Complex-Motion Neural Radiance Fields (CM-NeRF), which is a method that leverages motion and depth priors to optimize neural 3D scene representations and complex 6-DoF camera motions jointly. Although NeRF has achieved remarkable success in neural rendering and reconstruction, they require accurate camera poses and sufficient input for realizing high-quality novel view synthesis. We aim to recover accurate camera motion and NeRF simultaneously by effectively using motion and depth priors when a few input images are available. Moreover, our approach enables stable pose estimation and efficient view recovery in challenging and complex camera movements in addition to forward-facing camera motions. Considering the confidence of the depth, we use the depth map to guide the ray sampling and leverage depth information to accelerate the NeRF network training. Our experiments demonstrate the effectiveness of the CM-NeRF method in real-world scenarios involving challenging and complex camera motions. These results are non-trivial and may present significant variations compared to state-of-the-art techniques. The CM-NeRF demonstrates stable camera-pose estimation and efficient view recovery with only five training views of real-world data.

INDEX TERMS Neural radiance fields, neural rendering, novel view synthesis.

I. INTRODUCTION

Realistic novel view synthesis is essential in various domains, including robotics, computer vision, computer graphics, and augmented reality/virtual reality (AR/VR) applications wherein virtual environments and people interact. Recently, neural rendering techniques based on implicit representations (as in [1], [2]), such as Neural Radiance Fields (NeRF) [3], have progressed significantly in the realization of photo-realistic novel view synthesis. NeRF has the ability to realistically represent complex scenes without constraining the view synthesis resolution and while efficiently memorizing the 3D space. This technique uses weights of multi-layer perceptrons (MLPs) trained on numerous RGB images and their corresponding poses as inputs to represent scenes, which showcases its superior view synthesis capabilities. NeRF relies on Structure-from-Motion (SfM) methods such as COLMAP [4] to obtain accurate camera poses as their input. However, the pre-processing step of SfM can be time-consuming and resource-intensive. Several recent studies [5], [6], [7], [8], [9], such as Bundle-Adjusting NeRF (BARF) [8] and NeRF−− [9], have been focused on jointly optimizing neural 3D representations and six-degree-of-freedom (6-DoF) camera poses. Although these methods have provided promising results with unknown camera poses, they have been mainly assessed using forward-facing scenes with fewer variations in camera pose rotation and translation. In scenarios wherein the camera undergoes a more diverse motion, many RGB images and considerable training time are required to learn the scene representations and camera poses from scratch, while pose convergence is occasionally unsuccessful, as illustrated in Fig. 1.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

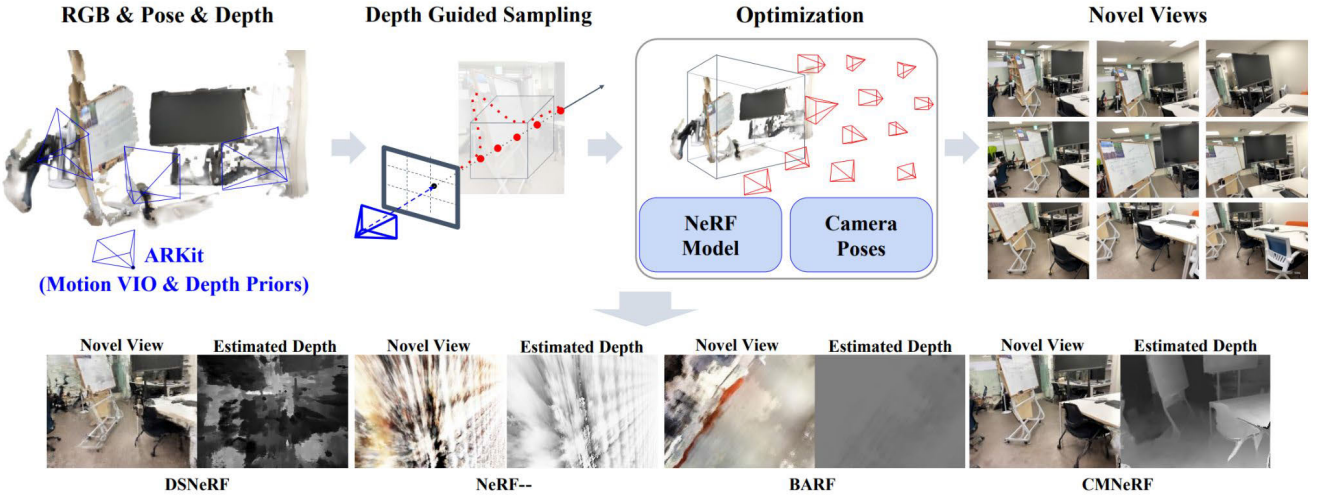

FIGURE 1. We leverage motion and depth priors for jointly registering camera poses and learning neural 3D representations of complete rooms when only sparse input views are available. Our motion and depth supervision facilitates the synthesis of high-quality novel views from a small number of input images captured using various camera motions in addition to forward-facing scenes.

To generate 3D scenes using limited RGB images, recent NeRF-variant approaches [10], [11] have incorporated additional depth priors from SfM or monocular depth completion techniques. However, these methods still require ground-truth camera poses computed from SfM for realizing effective learning, which can be a time-intensive process. Moreover, they rely heavily on SfM tools such as COLMAP, and their performance can be impacted by errors in the estimated poses or depths. Permutations in the SfM result can also significantly affect the final output of these methods.

To address the challenges of NeRF reconstruction, we propose a Complex-Motion NeRF (CM-NeRF), which is a novel approach that leverages motion and depth priors to estimate complex camera movements accurately and reconstruct scenes using only five input views. CM-NeRF overcomes the limitations of traditional methods by generating novel views from different angles. This provides a wider range of rotation and more natural, dynamic camera movements that extend beyond the constraints of the 2D plane. As a result, the reconstructed scene exhibits more realistic motion. Moreover, the CM-NeRF incorporates an optimization process for refining camera poses during reconstruction. This ensures a stable view synthesis and accurate pose estimation, even in the presence of an initial input noise or uncertainty caused by unexpected camera movements or perturbations.

We use motion and depth priors from commercial libraries, such as Apple ARKit's visual-inertial odometry (VIO) [12], to improve the camera pose estimation. VIO sensors are commonly found in smartphones and robotics, which makes them easy to use and practical. Obtaining motion priors from VIO is a straightforward process. Using VIO-generated outputs is a time- and resource-efficient alternative to the conventional COLMAP. We deliberately incorporated VIO into our approach owing to its superior accuracy and convenience. Our approach integrates seamlessly with VIO systems and prioritizes ease of use over traditional SfM algorithms such as COLMAP. By leveraging motion priors generated by VIO, we achieve faster and more accurate convergence to precise 6-DoF camera poses during the optimization [13]. Our approach differs from the original BARF [8] in camera initialization. Instead of initializing the identity transformation, we leverage motion and depth priors to provide better initial values that are close to the optimal solution. This prevents the occurrence of local minima during training, thus ensuring stable and accurate camera-pose estimation and high-quality view synthesis, even with a sparse input and complex camera motion in real-world scenes. The CM-NeRF is particularly well-suited for robotics, especially drones with complex motions. It can create detailed 3D models using sparse input data, which enhances the perception, localization, and mapping. This makes it highly effective in real-world scenarios with diverse robot movements. Our main contributions are as follows:

- We present a novel approach, CM-NeRF, that jointly optimizes 6-DoF camera poses and scene reconstruction using motion and depth priors.
- The CM-NeRF is a practical and feasible approach that integrates motion and geometric priors from VIO to improve accuracy and demonstrate robust performance in complex motion scenes with sparse inputs.
- The CM-NeRF is an effective solution for robotics that perform complex motions, as it can create detailed 3D models using sparse input data, while demonstrating improved perception, localization, and mapping capabilities in real-world scenarios.

II. RELATED WORK

The NeRF [3] is a powerful neural rendering technique that represents 3D scenes as implicit neural functions. The NeRF has demonstrated excellent performance in synthesizing new

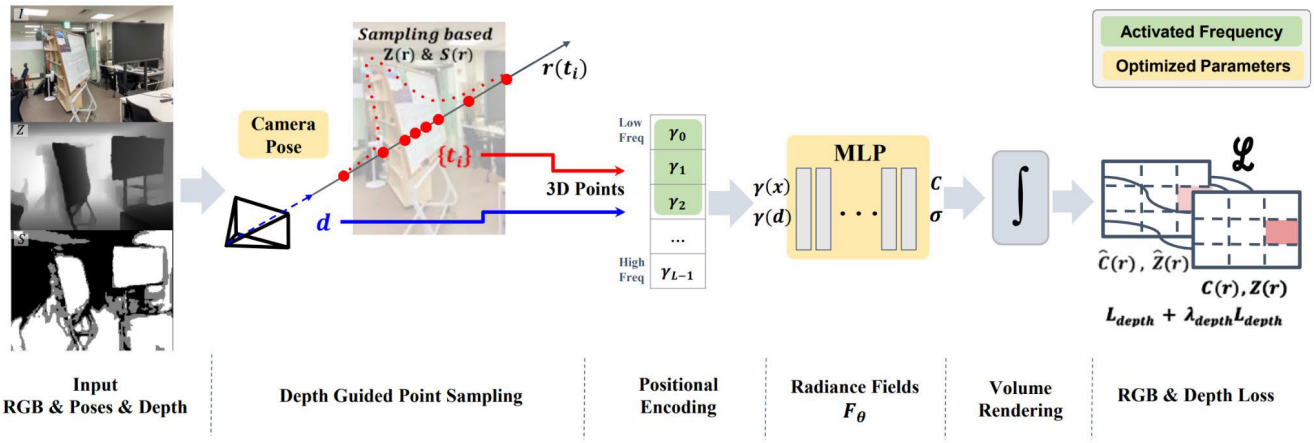

FIGURE 2. Overview of our CM-NeRF pipeline. Given a small set of RGB images *I*, depth maps *Z*, confidence maps *S*, and camera poses, we can determine the sampling bound by focusing on the surface via depth priors. The discrete point t_i is then sampled along the direction d of the camera ray to render the pixel color on the image based on our depth-guided sampling range. The position *x* of the sampled 3D points and camera viewing direction *d* are input into the NeRF through the smooth positional encoding γ, which results in $\hat{C}$ through volume rendering. At this stage, a smooth mask is applied to the PE that gradually activates a high frequency that is proportional to the optimization process. The output color and density from the volume rendering are integrated to obtain the color and depth of the final pixel. As the entire pipeline is differentiable, we can jointly optimize the 6-DoF camera poses and MLP to minimize the color and depth loss.

views, but its high computational time limits its practical applicability. To address this limitation and enhance the capabilities of the NeRF, several studies have been conducted, which include the use of priors for realizing a performance improvement [6], [10], [11], [14], speed-up techniques [15], [16], re-lighting methods [17], [18], and large-scale extensions [19], [20].

A. DEPTH SUPERVISION NeRF

In several studies, depth information was used in the NeRF, which enhanced the synthesis robustness with the use of a small number of input images. In depth oracle NeRF [21], a "depth oracle" network is introduced that effectively reduces the number of local samples of each ray needed near the surface, which results into an improved rendering speed. However, the application of this approach to real-world data is challenging as it requires a dense depth map. The dense depth priors NeRF (DDPNeRF) [11] estimates dense depth maps using a depth completion network on sparse point clouds obtained from COLMAP. It then incorporates the predicted dense depth for ray sampling while taking into consideration the depth uncertainty. The depth-supervised NeRF (DSNeRF) [10] trains the NeRF using the camera pose and sparse point cloud estimated in COLMAP. It constructs a loss function based on the predicted termination depth along each key point in the ray. This facilitates the faster training of DSNeRF with fewer views. DSNeRF and DDPNeRF are heavily dependent on accurate camera poses from COLMAP. However, COLMAP can be affected by input sparsity and percussion in rendering results. This is especially challenging for scenes comprising texture-less walls or repetitive patterns, in which case COLMAP may have difficulty estimating the camera poses and point clouds. The CM-NeRF addresses this challenge by leveraging the ARKit VIO system. This approach eliminates the need for preprocessing using COLMAP, thus simplifying the process and improving the accuracy. The CM-NeRF also focuses on seamlessly integrating with commercial VIO systems by prioritizing ease of use. In contrast to previous approaches, which require preprocessing through COLMAP to obtain motion and depth information, our method directly use ARKit to obtain motion and depth priors. Therefore, the CM-NeRF jointly optimizes the camera poses and NeRF network using these priors, thus enabling training with imperfect or unknown camera poses.

B. OPTIMIZING CAMERA POSES

In the field of simultaneous localization and mapping (SLAM), several methods address the challenges of handling large datasets and optimizing camera poses and scene representations. iMAP [22] uses the NeRF MLP to store 3D information in space. It can generate high-quality 3D models, but it has high memory requirements. iNeRF [23] is a new pose estimation method that leverages pre-trained NeRF networks. The iNeRF is less memory-intensive than the iMAP, but its results are dependent on the quality of the training data to a greater extent. NICE-SLAM [24] integrates a hierarchical scene representation to integrate multi-level local information. NICE-SLAM can generate detailed reconstructions of large indoor scenes, but it is computationally expensive and may not be able to handle all types of scenes.

To address the problem of NeRF reconstruction in the case of unknown camera poses, GNeRF [25] integrates generative adversarial networks (GAN) with the NeRF. However, this combination increases the computational complexity. As another method, the scene representation

transformer (SRT) [7] generates novel views from RGB images with or without poses. The SRT adopts geometry-free learning-based approaches, which makes it more suitable for large datasets. While self-calibrating NeRF [5], NeRF−− [9], and BARF [8] jointly optimize the camera poses and scene representations, but they struggle with sparse input views. It should be noted that the BARF and NeRF−− set the initial poses as identity matrices for real-world data, and thus require sufficient input images for stable pose optimization and high-fidelity view synthesis. Moreover, these methods have been primarily tested using forward-facing poses with minimal pose variations. Previous studies have emphasized the importance of obtaining a good initial guess in robotics, with contributions in 2D SLAM [26], [27] and visual-inertial navigation [28], [29]. In our approach, we leverage motion and depth priors to initialize near-optimal solutions, thus enhancing the pose convergence and facilitating stable pose estimation even in the case of challenging movements. We use the reliable initial guess obtained from ARKit as motion priors, which improves the stability and accuracy of the pose estimation. Furthermore, we employ depth priors to guide the scene sampling, thus allowing us to perform novel view and depth estimation using a small number of images and even with inaccurate poses.

III. BACKGROUND

We propose a novel optimization process that optimizes poses and NeRF networks. Our approach leverages motion and depth priors with the use of a small number of inputs to enhance the optimization. Our model receives a set of posed images $\{I_i \in \mathbb{R}^{H\times W\times 3},\ p_i \in SE(3)\}_{i=1}^{M}$, depth maps $\{Z_i\}_{i=1}^{M} \in [0, 4.5]^{H\times W}$, and depth confidence maps $\{S_i\}_{i=1}^{M}$ (see Fig. 2). The RGB images, depth, and confidence maps have a resolution of 192 × 256. The depth confidence map is a $H \times W$ value matrix consisting of 0, 1, or 2. A higher value indicates higher confidence in the depth measurement. In the confidence map, brighter regions indicate higher confidence levels. White, gray, and black denote confidence 2, 1, and 0, respectively, as shown in Fig. 2. The distribution of the confidence levels varies across different datasets. In the dataset used in this study, confidence 2, which represents a high reliability, accounts for an average of approximately 80.52% in each image.

A. NEURAL RADIANCE FIELDS

NeRF [3] uses neural networks to represent the entire 3D space by receiving input images of known poses to render unseen views. To render a pixel $\mathbf{u}$ on an image, the NeRF first casts a ray $\mathbf{r}(\mathbf{t}) = \mathbf{o} + \mathbf{td}$ passing through the camera center $\mathbf{o}$ along the direction $\mathbf{d}$ of the pixel. Then, the NeRF samples a set of points between the near and far bounds in each ray. The sampled 3D point enters the input of multi-layer perceptrons (MLP) through positional encoding, which maps $\mathbf{x}$ and $\mathbf{d}$ to higher dimensions for a high-fidelity view synthesis. The MLP is parameterized by Θ and encodes the input 3D point location $\mathbf{x} \in \mathbb{R}^3$ and viewing directions $\mathbf{d} \in \mathbb{R}^2$ into volume density σ and color $\mathbf{c}$: $F_\Theta(\mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$. The volume rendering determines the final color and density of the pixel by approximating the radiance from direction $\mathbf{d}$ along the pixel's ray. The MLP parameters are optimized by minimizing a differentiable color rendering loss.

B. BUNDLE ADJUSTING NEURAL RADIANCE FIELDS

The BARF [8] is a modified version of the NeRF that optimizes a single MLP without additional hierarchical sampling. The BARF jointly optimizes the camera poses and MLP parameters, thus facilitating rendering even when no camera pose is available. The BARF applies a smooth mask to the positional encoding to optimize the camera poses in proportion to the training process. The smooth mask gradually increases the frequency of the incrementally high positional encoding during the optimization process, which improves the model's ability to localize and synthesize high-fidelity views.

IV. PROPOSED METHOD

A. POSE ESTIMATION USING MOTION PRIORS

Accurate camera pose estimation is crucial for high-quality rendering in a NeRF. A previous research [13] highlights the importance of a good initial guess for realizing accurate camera pose estimation. To enhance the stability and fidelity, we leverage Apple ARKit's precise initial guesses from its VIO algorithm, which is known for its accuracy and stability [30]. In contrast, the BARF's identity matrix initialization can result in local minima and longer training times, which makes it suitable only for forward-facing scenes [31], [32] with minimal camera movement. Our approach addresses these limitations by seamlessly integrating with VIO systems such as ARKit, thus enabling stable pose estimation and high-fidelity rendering. The integration of the motion and geometric priors from off-the-shelf VIO systems enhances the accuracy and robustness.

Our proposed CM-NeRF approach offers improved freedom of movement and delivers more flexible and accurate results using just five sparse inputs. This is a significant improvement over previous forward-facing methods and results in stable poses and high-quality rendering. Moreover, our algorithm is designed to be practical, easily implementable, and widely adoptable. It provides flexibility, accuracy, and stability in CM-NeRF with only five sparse inputs.

B. OPTIMIZATION WITH DEPTH

We present an approach using motion and depth prior to the BARF for the MLP and pose optimization. For each sampling location, we sample N discrete points, denoted as $t_i \in \left[t_n, t_f\right]$, along the ray, where the lengths $\delta_i = t_{i+1} - t_i$. The estimated color $\hat{C}(r)$ of the camera ray $r(t)$, with near and far bounds t_n and t_f, is calculated using a weighting distribution of the

points sampled along the ray.

$$\hat{C}(r) = \sum_{i=1}^{N} w_i c_i, \tag{1}$$

$$w_i = T_i(1 - exp(-\sigma_i \delta_i)), \tag{2}$$

where

$$T_i = exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \tag{3}$$

Based on [11], the estimated depth $\hat{z}(r)$ and estimated depth confidence $\hat{s}(r)$ are calculated using w_i. The uncertainty of the estimated depth at a given point is denoted by $\hat{s}(r)$. The standard deviation $\hat{s}(r)$ is calculated by obtaining the standard deviation of the depth values in the predicted depth map $\hat{z}(r)$. A higher standard deviation indicates more significant uncertainty in the depth estimate $\hat{z}(r)$. It is important to note that a higher standard deviation generally results in a lower accuracy.

$$\hat{z}(r) = \sum_{i=1}^{N} w_i t_i \tag{4}$$

$$\hat{s}(r)^2 = \sum_{i=1}^{N} w_i (t_i - \hat{z}(r))^2 \tag{5}$$

Given images $\{I_i\}_{i=1}^{M}$ and depth $\{Z_i\}_{i=1}^{M}$, the goal is to optimize the NeRF network parameter Θ and the 6-DoF camera poses $\{p_i\}_{i=1}^{M}$ corresponding to the images.

$$\min_{p_1,\ldots p_M,\Theta} \sum_{i=1}^{M} \sum_{r \in R} \left(L_{color} + \lambda_{depth} L_{depth}\right) \tag{6}$$

L_{color} is the mean squared error term for the color and L_{depth}, based on by [11], is the mean Gaussian negative log-likelihood term for the depth, where λ_{depth} is a hyper-parameter balancing depth supervision.

$$L_{color} = \left\| \hat{C}(r; p_i, \Theta) - C_i(r) \right\|_2^2 \tag{7}$$

$$L_{depth} = \begin{cases} log(\,\hat{s}(r; p_i, \Theta)^2) + \dfrac{(\hat{z}(r; p_i, \Theta) - z(r))^2}{\hat{s}(r; p_i, \Theta)^2} & if\ P \\ 0 \quad else \end{cases} \tag{8}$$

$$, where\ P = z(r) < 4.5\ and\ s(r) = 2 \tag{9}$$

A depth loss is applied at a ray that satisfies the following conditions. First, the target depth $z(r)$ is less than 4.5. As explained for the iPhone's LiDAR sensor in [33], the LiDAR sensor provides a reliable accuracy up to 4.5 meters. Second, only the ray $r(t)$ with the target depth confidence $s(r)$ of 2 should be applied. Our approach prioritizes highly reliable depth values to ensure accurate results, thus preserving the accuracy and avoiding distortion caused by less reliable values.

C. DERIVING DEPTH-SUPERVISION SAMPLING

The majority of the scenes consist of empty or opaque spaces, the contribution of which to the weighted contribution function in Eq.1 is negligible [3]. To address this issue, we adopt an approach similar to that of [34], while increasing the sampling distribution of the points near the surface using depth priors. This ensures that these points contribute to the final color rendered. In the training process, we maintain the same number of samples N as the original BARF. We enhance their effectiveness by sampling half the points near the surface based on depth priors. The remaining half of the points are extracted between the existing near and far boundaries. For ray directions with a low depth confidence (where $s(r)$ is zero), we extract N points between the existing near and far boundaries. This approach effectively reduces the noise in the rendered images and improves the quality of the final rendered color.

V. EXPERIMENTS

We evaluate the effectiveness of our method through a baseline comparison and an ablation study using real-world datasets collected by the authors through ARKit. ARKit is a software framework developed by Apple that includes a VIO algorithm for producing augmented reality applications. We collect custom data using an iPad Pro 11 running iOS 16.3.1 and equipped with LiDAR sensors, including RGB images, ARKit 6-DoF poses, depth maps, and depth confidence maps. We uniformly selected five training views from the entire VIO sequence at regular intervals. We randomly selected 20 views for testing from the entire sequence, while excluding the training views. The proposed method was trained using a set of five views, and for validation, a hold-out set of four views was used.

Evaluation Metrics: To evaluate the effectiveness of our approach, we quantitatively assess both the accuracy of the optimized poses and the quality of the synthesized views. We compute a variety of perceptual metrics that are widely used for evaluating view synthesis results in the NeRF series and that include the mean peak signal-to-noise ratio (PSNR), structural similarity index map (SSIM) [35], and learned perceptual image patch similarity (LPIPS) [36] to evaluate the image rendering quality. In addition, we report the average rotation and translation errors of the estimated poses. To validate the accuracy of our approach, we calculate the deviation between the optimized poses and the ground truth poses obtained from OptiTrack.

A. BASELINE COMPARISON

We compare the following NeRF-variant methods, which are trained on five input views:

- NeRF [3]: A neural rendering technique that learns continuous volumetric representations of scenes from a set of 2D images and their corresponding camera poses.
- DSNeRF [10]: The depth-supervised variant of the NeRF that guides the learning of scene geometry and

FIGURE 3. We compare the estimated RGB and depth of our proposed method with the ground truth RGB. Furthermore, we render novel views and depth for various NeRF models trained on five views. The experimental results show that our proposed method outperforms the other evaluated methods in terms of image quality and depth accuracy. This demonstrates the effectiveness of our method in accurately capturing the scene geometry and appearance and robustly estimating the camera poses.

TABLE 1. Quantitative results for five input views. Our model outperforms the previous state of the art across all datasets.

Scene	PSNR ↑					SSIM ↑					LPIPS ↓				
	NeRF	DSNeRF	NeRF−−	BARF	Ours	NeRF	DSNeRF	NeRF−−	BARF	Ours	NeRF	DSNeRF	NeRF−−	BARF	Ours
Lab	14.27	16.68	7.16	8.47	**20.13**	0.4	0.56	0.21	0.24	**0.75**	0.48	0.35	0.78	0.77	**0.17**
Flowerpot	15.43	19.49	9.4	11.53	**21.74**	0.28	0.54	0.17	0.18	**0.59**	0.42	0.29	0.73	0.51	**0.17**
Lounge	19.64	17.98	9.16	8.95	**21.62**	0.66	0.45	0.19	0.27	**0.82**	0.24	0.69	0.72	0.7	**0.08**
Hallway	16.22	16.25	10.47	10.94	**22**	0.58	0.61	0.44	0.33	**0.81**	0.37	0.34	0.79	0.81	**0.18**
Stair	16.81	20.13	13.44	14.11	**24.12**	0.41	0.56	0.25	0.28	**0.69**	0.37	0.34	0.71	0.72	**0.21**
Bin	15.5	16.02	11.19	12.68	**19.7**	0.45	0.52	0.38	0.36	**0.71**	0.45	0.45	0.8	0.55	**0.18**
Sink	16.41	17.98	9.13	11.24	**20.84**	0.6	0.71	0.25	0.44	**0.75**	0.46	0.31	0.78	0.75	**0.24**
Mean	16.33	17.79	9.99	11.13	**21.45**	0.48	0.56	0.27	0.3	**0.73**	0.4	0.4	0.76	0.69	**0.18**

TABLE 2. Rotation Matrix Difference (RMD) in training datasets.

Method	Rot (°) ↑	
	RMSE	MAX
BARF	14.13	25.21
CMNeRF	**37.33**	**51.07**

appearance using a separate depth map as the input during training.

- NeRF−− [9]: The method learns camera parameters implicitly from input images without requiring known camera parameters.
- BARF [8]: The method that jointly optimizes camera poses and scene geometry using bundle adjustment, while taking RGB images and camera intrinsics as the input.

$$\arccos\left(Tr(R_1{}^\top R_i - 1)/2\right) \tag{10}$$

$$\max \arccos\left(Tr(R_1{}^\top R_i - 1)/2\right), \ \textit{where } i \in \{2, 3, 4, 5\} \tag{11}$$

Our proposed method defines complex motion as camera movements involving greater rotation changes and unrestricted translations beyond the 2D plane, which surpasses the limitations of forward-facing motions. Table 2 provides insights into the rotation matrix difference (RMD) [37] observed in training datasets. Here, $R \in SO(3)$ represents the rotation matrix of a 6-DoF camera motion. In the training data, we calculate the RMD between the remaining frames based on the first frame as shown in Eq.10. We calculate the maximum angle of the RMD as Eq.11. Table 2 shows that our approach allows up to 51 degrees of the RMD compared to the rest of the pose based on the first frame. Furthermore, our method demonstrates camera pose variations approximately twice as large as the forward-facing motions observed in the BARF.

Table 1 shows that our method outperforms the baselines across all metrics. The BARF and NeRF−− encounter issues in optimizing the camera poses when applied to complex camera pose trajectories with sparse input views or challenging camera motion, thus resulting in significant geometry and color defects. This problem is particularly noticeable in environments comprising a large range of camera motion, which results in inaccurate pose estimates. For instance, the BARF failed to estimate poses, thus rendering artifacts in

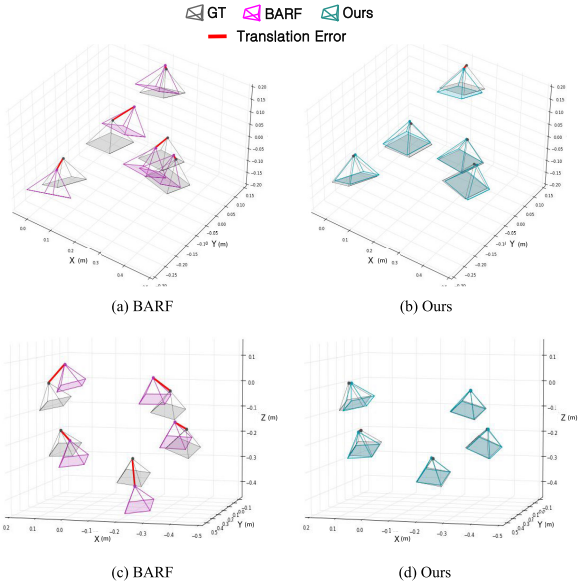

FIGURE 4. We visualize and compare the optimized camera poses to the OptiTrack ground truth poses. The first and second rows are the result of pose optimization for different datasets. Our proposed method (right) achieves high consistency with the ground truth, while the BARF method (left) produces sub-optimal results.

several regions, such as the flowerpot and the table in Fig. 3. Our method demonstrates a significant reduction in pose instability compared to the BARF and NeRF−−, owing to the use of both initial guidance and depth information. Our approach exhibits superior robustness in pose estimation with sparse input views. As shown in Fig. 4, our method outperforms the BARF in terms of robustness in the pose estimation. Furthermore, our approach demonstrates a superior performance in the camera pose estimation even with more complex and less restrictive camera motions, as demonstrated by the quantitative error metric in Table 3. Our results highlight the importance of our method for handling challenging camera trajectories beyond what is achievable using the baseline comparison methods.

The DSNeRF can be sensitive to errors in the depth estimates from COLMAP, especially in the case of scenes comprising texture-less walls or repetitive patterns. AS DSNeRF directly incorporates the uncertain depth estimated from COLMAP, errors are more likely to occur, which results in defects in the geometry and color. For example, an erroneous

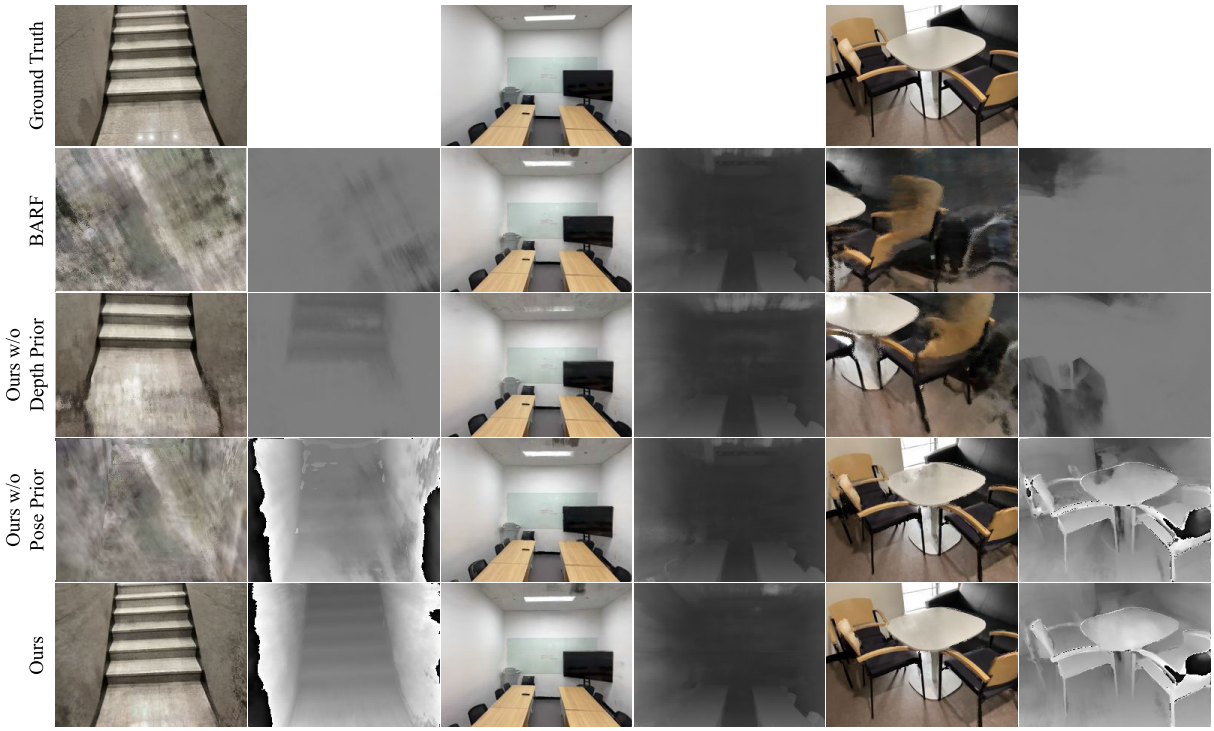

FIGURE 5. We visualize the estimated RGB and depth images. We observe that when the camera motion is small, the method performs well even without motion and depth priors. However, for more complex camera motions, the integrating of motion and depth priors is necessary to achieve accurate view synthesis results. These visualizations demonstrate the effectiveness of motion and depth priors for accurately capturing the scene geometry and appearance.

input in the area of the table and chair (third dataset in Fig. 3) causes much more significant deficiencies in geometry and color in the case of the DSNeRF compared to the other approaches. Our method uses depth priors from off-the-shelf VIO systems to guide ray sampling during neural representation training. This improves the efficiency and effectiveness of the training, thus resulting in better estimates of geometry and color and more robust pose estimation. The results in Fig. 3 and Table 1 show that our method outperforms the baseline methods in robustness, even with sparse depth inputs or initial values subject to perturbations.

B. ABLATION STUDY

We conduct experiments to analyze the impact of the depth prior and motion prior. In the motion prior ablation experiment, we set the initial camera poses as identity matrices. In the depth prior ablation experiments, we do not use the depth prior and, therefore, do not apply depth loss and depth-guided sampling. Table 3 and Fig. 4 demonstrate that our method, comprising the use of depth and motion priors, achieves better performance in image quality and pose optimization than other methods. This consistency is also observed in the qualitative results presented in Fig. 5.

1) WITHOUT POSE AND DEPTH PRIORS

This condition is the same as that of the BARF. We observed that scenes with less movement, particularly forward-facing scenes, produced higher-quality view synthesis results (second dataset in Fig. 5). However, we observed that the pose estimation was more defective for scenes with greater movement, which resulted in color and depth estimation failures, as shown in Fig. 5 and Table 3.

TABLE 3. Quantitative analysis of camera pose optimization using motion and depth priors for high-quality view synthesis.

Method	Pose Estimation		View Synthesis		
	Rotation (°) ↓	Translation (m) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
BARF	25	0.1	11.82	0.33	0.66
Ours w/o Depth	5.58	0.03	15.51	0.48	0.4
Ours w/o Pose	25	0.1	12.57	0.4	0.6
Ours	**1.79**	**0.02**	**20.34**	**0.69**	**0.19**
ARKit	2.13	0.07	-	-	-

2) WITHOUT DEPTH PRIORS (WITH POSE PRIORS)

We observed a significant reduction in the pose estimation error compared to the previous condition. However, the pose estimation error is still high in complex camera motion scenes with a sparse input. This is evidenced by the stair and table examples in Fig. 5 and Table 3. The rotation error increased, and the translation error decreased compared to the initial ARKit pose. Moreover, realizing the estimation of the geometric structure based solely on the pose prior is challenging. Our results suggest that the incorporation of the depth prior in the sparse input improves the geometry-structure estimation.

3) WITHOUT POSE PRIORS (WITH DEPTH PRIORS)

When using only the depth prior, the rendering results obtained can sometimes be better than those obtained using only the pose prior, as demonstrated in the last example of Fig. 5. However, the pose estimation fails in the majority of cases, as can be observed from the pose estimation results in Table 3. This suggests that the motion prior is necessary for successful pose optimization. Furthermore, the color estimation has some defects, but the geometric structure demonstrates the effectiveness of incorporating depth prior in the estimation results.

Our experiments show that our model is robust to complicated camera trajectories with a sparse input. This is evident from the quantitative results listed in Table 3, which show that our model's optimized poses exhibit lower error rates than the initial ARKit 6-DoF poses. This indicates that the optimized poses are more accurate and contribute to generating precise view synthesis results. This improvement in pose estimation is due to the use of motion and depth priors. These priors aid in mitigating the deficiencies in the view synthesis and pose estimation that can occur with the use of a sparse input. As shown in Fig. 5, the use of the depth and motion priors results in a more accurate view synthesis and pose estimation.

VI. CONCLUSION

We propose CM-NeRF, which is a novel approach that integrates motion and depth priors into the optimizing NeRF and camera poses. Our experimental results on real-world datasets demonstrate the superior performance of our method, even in the case of challenging scenarios with inaccurate camera poses. By leveraging motion and geometric priors from off-the-shelf VIO systems, we have effectively combined existing components and realized a strong performance at the system level. We envision opportunities to enhance the CM-NeRF's performance on large-scale and dynamic scenes and explore other input data types, such as LiDAR or stereo images. Moreover, the incorporation of semantic information or object-level priors is promising as future research directions.

REFERENCES

[1] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5932–5941.

[2] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, and A. Eriksson, "Implicit surface representations as layers in neural networks," in *Proc. ICCV*, 2019, pp. 4743–4752.

[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] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. CVPR*, 2016, pp. 4104–4113.

[5] Y. Jeong, S. Ahn, C. Choy, A. Anandkumar, M. Cho, and J. Park, "Self-calibrating neural radiance fields," in *Proc. ICCV*, 2021, pp. 5846–5854.

[6] A. Rosinol, J. J. Leonard, and L. Carlone, "NeRF-SLAM: Real-time dense monocular SLAM with neural radiance fields," 2022, *arXiv:2210.13641*.

[7] M. S. M. Sajjadi, H. Meyer, E. Pot, U. Bergmann, K. Greff, N. Radwan, S. Vora, M. Lucic, D. Duckworth, A. Dosovitskiy, J. Uszkoreit, T. Funkhouser, and A. Tagliasacchi, "Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1–6.

[8] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "BARF: Bundle-adjusting neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5721–5731.

[9] Z. Wang, S. Wu, W. Xie, M. Chen, and V. Adrian Prisacariu, "*NeRF*−−: Neural radiance fields without known camera parameters," 2021, *arXiv:2102.07064*.

[10] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, "Depth-supervised NERF: Fewer views and faster training for free," in *Proc. IEEE CVPR*, Jun. 2022, pp. 12882–12891.

[11] B. Roessle, J. T. Barron, B. Mildenhall, P. P. Srinivasan, and M. Nießner, "Dense depth priors for neural radiance fields from sparse input views," in *Proc. IEEE CVPR*, Jul. 2022, pp. 12892–12901.

[12] *Apple ARKit*. Accessed: 2023. [Online]. Available: https://developer.apple.com/documentation/arkit/

[13] L. Carlone, R. Tron, K. Daniilidis, and F. Dellaert, "Initialization techniques for 3D SLAM: A survey on rotation estimation and its use in pose graph optimization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 4597–4604.

[14] M. Kim, S. Seo, and B. Han, "InfoNeRF: Ray entropy minimization for few-shot neural volume rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12902–12911.

[15] C. Reiser, S. Peng, Y. Liao, and A. Geiger, "KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14315–14325.

[16] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," 2022, *arXiv:2201.05989*.

[17] P. P. Srinivasan, B. Deng, X. Zhang, M. Tancik, B. Mildenhall, and J. T. Barron, "NeRV: Neural reflectance and visibility fields for relighting and view synthesis," in *Proc. IEEE CVPR*, Jul. 2021, pp. 7495–7504.

[18] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NERF in the wild: Neural radiance fields for unconstrained photo collections," in *Proc. IEEE CVPR*, Jun. 2021, pp. 1–12.

[19] 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.

[20] Y. Xiangli, L. Xu, X. Pan, N. Zhao, A. Rao, C. Theobalt, B. Dai, and D. Lin, "BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering," 2021, *arXiv:2112.05504*.

[21] T. Neff, P. Stadlbauer, M. Parger, A. Kurz, J. H. Mueller, C. R. A. Chaitanya, A. Kaplanyan, and M. Steinberger, "DONeRF: Towards real-time rendering of compact neural radiance fields using depth Oracle networks," *Comput. Graph. Forum*, vol. 40, no. 4, pp. 45–59, Jul. 2021.

[22] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "iMAP: Implicit mapping and positioning in real-time," in *Proc. IEEE ICCV*, Oct. 2021, pp. 6229–6238.

[23] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, "iNeRF: Inverting neural radiance fields for pose estimation," in *Proc. IEEE IROS*, Sep. 2021, pp. 1323–1330.

[24] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "NICE-SLAM: Neural implicit scalable encoding for SLAM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1–4.

[25] 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. CVPR*, Oct. 2021, pp. 6351–6361.

[26] L. Carlone, R. Aragues, J. A. Castellanos, and B. Bona, "A fast and accurate approximation for planar pose graph optimization," *Int. J. Robot. Res.*, vol. 33, no. 7, pp. 965–987, Jun. 2014.

[27] L. Carlone and A. Censi, "From angular manifolds to the integer lattice: Guaranteed orientation estimation with application to pose graph optimization," *IEEE Trans. Robot.*, vol. 30, no. 2, pp. 475–492, Apr. 2014.

[28] A. Martinelli, "Vision and IMU data fusion: Closed-form solutions for attitude, speed, absolute scale, and bias determination," *IEEE Trans. Robot.*, vol. 28, no. 1, pp. 44–60, Feb. 2012.

[29] A. Martinelli, "Closed-form solution of visual-inertial structure from motion," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 138–152, Jan. 2014.

[30] P. Kim, J. Kim, M. Song, Y. Lee, M. Jung, and H.-G. Kim, "A benchmark comparison of four off-the-shelf proprietary visual–inertial odometry systems," *Sensors*, vol. 22, no. 24, p. 9873, Dec. 2022.

[31] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, Y. Wang, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Niessner, J. T. Barron, G. Wetzstein, M. Zollhoefer, and V. Golyanik, "Advances in neural rendering," 2021, *arXiv:2111.05849*.

[32] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, "NeRF: Neural radiance field in 3D vision, a comprehensive review," 2022, *arXiv:2210.00379*.
[33] *Apple ARKit LiDAR*. Accessed: 2023. [Online]. Available: https://www.it-jim.com/blog/iphones-12-pro-lidar-how-to-get-and-interpret-data/
[34] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo," in *Proc. IEEE ICCV*, Oct. 2021, pp. 5610–5619.
[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[36] 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.
[37] P. Kim, B. Coltin, and H. J. Kim, "Low-drift visual odometry in structured environments by decoupling rotational and translational motion," in *Proc. ICRA*, 2018, pp. 7247–7253.

HYUNJIN KIM (Student Member, IEEE) received the B.S. degree in computer science from Sookmyung Women's University, in 2021, where she is currently pursuing the master's degree with the Department of Mechanical Systems Engineering. She is a Researcher with the Machine Perception and Intelligence Laboratory, Department of Mechanical Systems Engineering, Sookmyung Women's University. Her research interests include 3D computer vision, mapping, visual odometry, and localization.

DAEKYEONG LEE is currently pursuing the bachelor's degree with the Department of Mechanical Systems Engineering, Sookmyung Women's University. She is a Research Intern with the Electrified Mobility Control Laboratory, Department of Mechanical and System Engineering, Sookmyung Women's University. Her research interests include robotic perception and autonomous driving.

SUYOUNG KANG (Student Member, IEEE) is currently pursuing the bachelor's degree with the Department of Electronics Engineering, Sookmyung Women's University. She is a Research Intern with the Machine Perception and Intelligence Laboratory, Department of Mechanical and System Engineering, Sookmyung Women's University. Her research interests include 3D computer vision, visual odometry, and visual SLAM for robotics.

PYOJIN KIM (Member, IEEE) received the B.S. degree in mechanical engineering from Yonsei University, in 2013, and the M.S. and Ph.D. degrees from the Department of Mechanical and Aerospace Engineering, Seoul National University, Seoul, South Korea, in 2015 and 2019, respectively. He was a Research Intern with Google (ARCore Tracking), Mountain View, in 2018. He is currently an Assistant Professor with the School of Mechanical Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea. Before joining GIST, he was a Postdoctoral Researcher with Simon Fraser University, Canada. His research interests include indoor localization, 3D computer vision, visual odometry, and visual SLAM for robotics.