2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3503-7770-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IROS58592.2024.10802480

Thermal-NeRF: Neural Radiance Fields from an Infrared Camera

Tianxiang Ye[1], Qi Wu[1], Junyuan Deng[1], Guoqing Liu[1],
Liu Liu[2], Songpengcheng Xia[1], Liang Pang[3], Wenxian Yu[1] and Ling Pei[1*]

***Abstract*— In recent years, Neural Radiance Fields (NeRFs) have demonstrated significant potential in encoding highly-detailed 3D geometry and environmental appearance, positioning themselves as a promising alternative to traditional explicit representation for 3D scene reconstruction. However, the predominant reliance on RGB imaging presupposes ideal lighting conditions—a premise frequently unmet in robotic applications plagued by poor lighting or visual obstructions. This limitation overlooks the capabilities of infrared (IR) cameras, which excel in low-light detection and present a robust alternative under such adverse scenarios. To tackle these issues, we introduce Thermal-NeRF, the first method that estimates a volumetric scene representation in the form of a NeRF solely from IR imaging. By leveraging a thermal mapping and structural thermal constraint derived from the thermal characteristics of IR imaging, our method showcases unparalleled proficiency in recovering NeRFs in visually degraded scenes where RGB-based methods fall short. We conduct extensive experiments to demonstrate that Thermal-NeRF can achieve superior quality compared to existing methods. Furthermore, we contribute a dataset for IR-based NeRF applications, paving the way for future research in IR NeRF reconstruction, see https://github.com/Cerf-Volant425/Thermal-NeRF.**

I. INTRODUCTION

Over the past decades, image-based 3D reconstruction has not only achieved remarkable success but also emerged as a pivotal area of research within the field of computer vision [1]. Its high accuracy, efficiency, and scalability have led to widespread applications across numerous domains, including artificial intelligence, robotics, autonomous driving, and virtual reality. Despite these advances, reconstructing 3D scenes within indoor environments presents unique challenges, due to the complexity of accurately modeling the geometry, spatial positioning, topological relationships, and semantic properties of these settings [2]. This challenge is critical, as overcoming it could significantly enhance tasks reliant on precise indoor positioning, such as semantic segmentation, scene understanding, and environmental perception.

The pressing need to address these intricate challenges has opened avenues for the exploration of innovative approaches. Neural Radiance Fields (NeRFs) have recently risen to prominence for tasks like 3D scene representation and novel view synthesis [3]. NeRFs combine a multilayer perceptron (MLP) for learning spatial information with differential rendering to realistically simulate light interactions within a scene. It marks a departure from explicit representation methods like voxel, point cloud, and mesh, adopting an implicit approach that excels in capturing detailed scene nuances [4]–[6]. Up to this date, NeRF has demonstrated effective performance under ideal conditions, characterized by stable and sufficient lighting and free of visual occlusions, with its efficacy proven on high-quality, real-world images captured in such optimal environments. However, real-world environments frequently deviate from the premise of photometric consistency, exhibiting challenges such as variable illumination and low light conditions. These factors significantly impair the performance of RGB-based NeRF approaches. Despite efforts to optimize NeRF for these conditions [7]–[9], such adaptations have fallen short in the face of severe visual degradation.

In the particularly demanding contexts of fire incidents and nighttime operations, the RGB cameras that feed NeRF with data fail to function effectively, rendering the system unsuitable. This highlights the necessity for infrared (IR) imaging approach [10] that can operate reliably in such challenging environments. IR cameras, with their ability to capture thermal signatures rather than relying on visible light, offer a distinct advantage in scenarios where conventional cameras falter. They excel in penetrating through smoke, detecting heat sources in the dark, and providing clear images despite the presence of obstructions, making them indispensable for scene reconstruction where darkness prevails and thermal sensitivity is crucial. However, the inherent characteristics of IR images, such as low contrast, sparse features and limited textures, which result in subtle pixel-level variations [11]–[13]. Since NeRF primarily constrains pixel-level loss, the nuanced variations in IR images significantly hinder its ability to accurately reconstruct scenes, posing challenges to NeRF reconstruction.

Given these learnings, we propose Thermal-NeRF, an innovative approach that tackles NeRF estimation from IR cameras, enabling the scene reconstruction from visually degraded environments. We apply a thermal mapping to model IR images' thermal value, ensuring the consistency in heat representation. And we introduce a structural thermal constraint to harness the structural information within images, offering vital constraints for IR images marked by sparse features and textures. Our dense experiments demonstrate that our method outperforms existing approaches in reconstruction quality. Furthermore, considering the current scarcity of IR datasets, we have compiled a targeted dataset for IR-based NeRF reconstruction. To summarise, the pri-

This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant No.62273229, No.61873163 separately and in part by smart city beidou spatial-temporal digital base construction and application industrialization (HCXBCY-2023-020).

[1]Shanghai Key Laboratory of Navigation and Location Based Services, Shanghai Jiao Tong University, [2]Hefei University of Technology, [3]Shanghai Slamtec Research.

*Corresponding author: Pei Ling ling.pei@sjtu.edu.cn

mary technical contributions are as follows:

1) We propose Thermal-NeRF, to the best of our knowledge, our approach is the first attempt to tackle NeRF estimation from IR cameras in situations where RGB cameras are inadequate.
2) We propose thermal mapping for modeling IR thermal values and and a structural thermal constraint on thermal distribution, which leverages IR's distinctive features for high-fidelity 3D representations.
3) We build an IR dataset featuring visually degraded scenes to validate our proposal and address the gap in existing IR datasets for NeRF. We will release the dataset to establish a benchmark for future work.

II. RELATED WORK

In this section, we examine key related works and explore their connections to our proposed method.

A. 3D Scene Representation

Addressing 3D scene representation has been a longstanding challenge in the fields of computer vision and computer graphics. Existing methods such as depth maps [14], point clouds [4], voxel grids [5], and meshes [6] have made significant contributions to this domain. However, each of these approaches comes with inherent limitations, be it in terms of resolution, computational efficiency, or the ability to capture complex geometries.

In contrast, the recent introduction of implicit coordinate-based representations [3], [15]–[17], exemplified by NeRF [3], represents a significant advancement. NeRF models a scene through a MLP as a continuous function of scene radiance and volume density, learned from a set of 2D RGB images. At test time, NeRF can render novel views from arbitrary 3D camera positions and viewing angles. However, a common limitation in most NeRF-based research is the need for high-quality input scene data, assuming ideal conditions for optimal performance.

On one hand, efforts have been made to enhance NeRF's adaptability to different lighting scenarios, RawNeRF [7] functions effectively in low-light environments, W-NeRF [8] is designed to cope with environments that exhibit changing lighting conditions and mitigates the effects of dynamic objects within a scene, LB-NeRF [9] manages scenes with transparent medium, overcoming challenges posed by light refraction, making NeRF more robust to light and broadening NeRF's applicability. On the other hand, [18]–[22] have seen significant improvements, as evidenced by various studies. Neus [18], for example, achieves high-fidelity reconstruction of objects and scenes from 2D RGB images, while ResNeRF [21] excels in synthesizing novel views with high fidelity, maintaining 3D structures in large-scale indoor scenes. Other researches [20], [22] focus on enhancing object textures, resulting in improved surface mesh quality.

Despite these developments, all these methods predominantly rely on traditional RGB camera inputs, which are less effective in dark or smoke-filled environments. Such challenging conditions are not only common but also crucial in the context of indoor scene reconstruction, underscoring a critical gap in the current state of the art.

B. Infrared Cameras

Infrared cameras [10], [23], which capture radiation in the infrared spectrum, specifically in the long-wave infrared (LWIR) range of 8 to 14 μm, produce spatial temperature distribution maps, independent of external light sources, capturing the infrared radiation emitted by objects. This technology has seen wide civilian application [24], ranging from fever scanners to insulation, due to reduced costs and enhanced portability. In addition, their high sensitivity has opened doors to various optical applications like fire prediction, electrical hotspot detection, and nighttime monitoring [25], offering advantages over RGB cameras, especially in distinguishing between objects based on heat signatures.

However, in the realm of IR scene reconstruction, distinct challenges are encountered, primarily attributed to the inherent properties of IR images. These images are typically characterized by sparsity in detail and diminished contrast [10]. Moreover, the acquisition of camera parameters, with a specific emphasis on extrinsic variables, proves infeasible through standard Structure from Motion (SfM) methodologies that are predicated on feature matching. Notably, techniques like COLMAP [26], which have found widespread application in NeRF for pose estimation, encounter limitations in this context. To tackle these issues, contemporary research has involved the fusion of IR with other modalities [27]–[30], compensating for the spatial data limitations of IR images. For example, Ma et al. [28] integrates RGB images to enrich the feature set and enhance depth perception, and Lang et al. [29] merges with Inertial Measurement Unit (IMU) data to provide pose information. Also, several studies have capitalized on the unique aspects of IR imaging using deep learning methods [11], [12], [30], concentrating on augmenting IR-specific features. Notably, X-NeRF [30] is the sole existing NeRF-based approach that concerns IR images, it creates a cross-spectral scene representation and learns the relative poses between IR and RGB sensors. In contrast, our approach is distinguished by its exclusive reliance on the IR modality for NeRF without auxiliary modalities.

III. METHODOLOGY

In this section, we introduce Thermal-NeRF, a new framework for scene representation with IR images. We first briefly introduce the background of original NeRF (Sec. III-A). Then we implement thermal mapping (Sec. III-B) to model thermal values and finally we innovate a structural thermal constraint (Sec. III-C) to leverage structural information. The overall framework is illustrated in Fig. 1 and described in detail below.

A. Background: Neural Radiance Field

NeRF [3] optimizes a neural radiance field parameterized by an MLP network $f_\Theta : (\mathbf{x}\ \mathbf{d}) \quad (\mathbf{c}\ \sigma)$, which maps Cartesian input coordinates $\mathbf{x} \quad \mathbb{R}^3$ and viewing direction

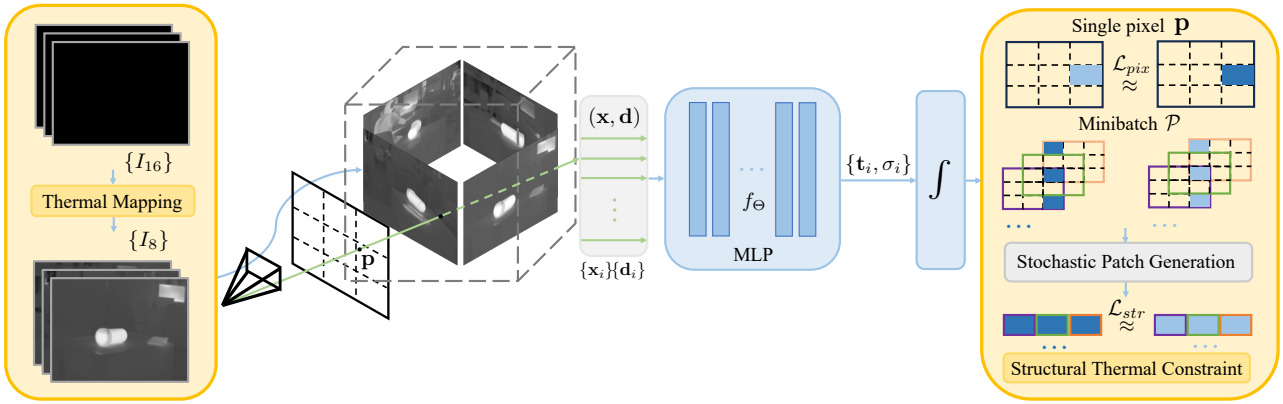

Fig. 1. Overview of the proposed Thermal-NeRF. Initially, a set of 16-bit IR images $\{I_{16}\}$, undergoes thermal mapping to be transformed into 8-bit images $\{I_8\}$. The method includes a scene contraction step, which compresses the indoor space into a predefined, fixed-size bounding box. Utilizing the camera parameters, ray bundles are generated through the contracted indoor scene. These bundles are then sampled to yield sampling points and directions $\{\mathbf{x}_i, \mathbf{d}_i\}$. The samples are encoded and fed into the MLP f_Θ. This step aggregates the output radiance $\mathbf{t}_i$ and densities σ_i to compute the thermal value. A unique structural thermal constraint is proposed to optimize the loss within mini-patches formed by stochastic pixels, see Equation 8.

$\mathbf{d}$ $\mathbb{S}^2$ to the predicted color $\mathbf{c}$ $\mathbb{R}$ and volume density σ $\mathbb{R}$.

Considering the errors introduced by applying the sensor calibration and readout delay of the data transmission, we apply pose refinement to optimize the camera pose and the neural field simultaneously to improve the image quality [31], [32]. We optimize camera views jointly with scene representation through an $SE(3)$ transformation, the camera views are then utilized to create ray bundles. To render an image from the NeRF model, the color at each pixel $\mathbf{p}$ $\mathbb{Z}^2$ on the image is obtained by volume rendering, aggregating the radiance along a ray $\mathbf{r}$ shooting from the camera position $\mathbf{o}_i$, passing through the pixel $\mathbf{p}$ into the volume [33]

$$\boldsymbol{C}(\mathbf{p}) = \int_{h_n}^{h_f} T(h)\sigma(\mathbf{r}(h))\mathbf{c}(\mathbf{r}(h)\ \mathbf{d})dh \tag{1}$$

where $T(h) = \exp(-\int_{h_n}^{h} \sigma(s)ds)$ denotes the accumulated transmittance along the ray, and $\mathbf{r}(h) = \mathbf{o} + h\mathbf{d}$ denotes the camera ray that starts from camera origin $\mathbf{o}$ and passes through $\mathbf{p}$, with near and far bounds h_n and h_f.

The original NeRF minimizes a least squares error between the rendered prediction colors $C(\mathbf{p})$ and ground truth colors $C(\mathbf{p})$ provided by the images. Akin to NeRF, Thermal-NeRF minimizes the error between rendered prediction thermal values $T(\mathbf{p})$ and ground truth thermal values $T(\mathbf{p})$, which is detailed in the next section.

B. Thermal Mapping

Diverging from RGB imaging, IR imaging encapsulates thermal values at each pixel, reflecting temperature variations vital for accurately interpreting thermal scenes. The data for our Thermal-NeRF are a set of IR images $I_k\ _{k=1}^{N}$ in 16-bit single-channel format, where each pixel corresponds to a thermal value. The transformation of the thermal value of T_{16} at a given 16-bit pixel $\mathbf{p}$ for camera ray $\mathbf{r}$ can be formulated as a linear mapping [10], the final conversion is given by the equation

$$T_{16}(\mathbf{p}) = \frac{\mathbf{p}}{\mathbf{k}} + \mathbf{b} \tag{2}$$

where $\mathbf{b}$ and $\mathbf{k}$ are two fixed imaging coefficients indicated by the IR camera.

Then we apply min-max scaling to the extreme thermal values within each data sequence $I_k\ _{k=1}^{N}$, ensuring heat consistency across all IR images, this approach also maximizes the contrast of images. To further aid visual task performance, we convert thermal values to 8-bit format, the conversion is written as

$$T(\mathbf{p}) = \frac{T_{16}(\mathbf{p}) - \min_{x\in\mathcal{P}} T_{16}(x)}{\max_{x\in\mathcal{P}} T_{16}(x) - \min_{x\in\mathcal{P}} T_{16}(x)} 255 \tag{3}$$

where $\mathcal{P}$ represents the set comprising every pixel within the data sequence, with 255 signifying the range from 0 to 255 in an 8-bit format.

The rendered thermal value $T(\mathbf{p})$ can then be compared against the corresponding ground truth thermal value $T(\mathbf{p})$, for all the pixels $\mathcal{P}$. We perform a least squares minimization of the pixel-wise thermal loss

$$\mathcal{L}_{pix}(\Theta) = \frac{1}{\mathcal{P}} \sum_{\mathbf{p}\in\mathcal{P}} T(\mathbf{p}) - T(\mathbf{p})\ ^2 \tag{4}$$

While pixel-wise thermal loss plays a crucial role in refining details at the pixel level, its focus on high-frequency changes limits its effectiveness in capturing the subtler, low-contrast nuances of IR imaging [34], exploring structural constraints becomes essential to accurately capture the spatial and textural nuances of thermal scenes.

C. Structural Thermal Constraint

Addressing the challenge of sparse textures and feature scarcity in IR imaging, where thermal radiation distribution is uneven, it's essential to focus on structural information. Before introducing the structural thermal constraint, we first need to define the evaluation metric based on thermal values to accurately assess structural information. To this end, we leverage Structural Similarity (SSIM) [35] index, which adeptly captures the structural similarities or differences in images, and allows for localized assessment focusing on

areas with concentrated thermal radiation. SSIM combines the three components of luminance, structure and contrast to comprehensively measure image quality. Let X be the reference image and Y be the test image, which is described as follows

$$\mathrm{SSIM}(X\ Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X{}^2 + \mu_Y{}^2 + C_1)(\sigma_X{}^2 + \sigma_Y{}^2 + C_2)} \tag{5}$$

where μ and σ denote the mean and standard deviation respectively, and σ_{XY} is the cross-correlation between X and Y. C_1 and C_2 two constants, are stable coefficients when the mean value and the variance are close to zero.

Additionally, IR imaging primarily detects variations in thermal values, reflecting temperature disparities rather than the brightness levels typically associated with RGB image quality assessment, so we should remove the luminance part. Then the heat-based metric HSSIM can be written as

$$\mathrm{HSSIM} = (X\ Y) = \frac{2\sigma_{XY} + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2} \tag{6}$$

we set $C_2 = 9 \times 10^{-4}$ in our work. Accounting for global information variability, we compute the mean HSSIM using local statistics derived from a sliding window W with kernel size $l \times l$, moving across the image with stride s. Local statistics are assessed within each window to compute HSSIM, with the final metric being the average of these values.

During stochastic training of NeRF, the randomly sampled pixels in a minibatch $\mathcal{P}$ do not constitute a coherent local patch, leading to a total loss of their spatial interrelation. As suggested by [36], for each sampled minibatch of pixels $\mathcal{P}$, they can form a stochastic patch through a patch generation function $G(\mathcal{P})$, which is initialized randomly each time when called. In this way, the structural loss can capture the non-local structural thermal information across all the training images. Let $\mathcal{T}$ and $\mathcal{T}$ be the sampled pixels transformed to the format of thermal values, where $\mathcal{T} = \ T(\mathbf{p})\ \mathbf{p}$ $\mathcal{P}$ and $\mathcal{T} = \ T(\mathbf{p})\ \mathbf{p}$ $\mathcal{P}$. The corresponding patches are generated as $G(\mathcal{T})$ and $G(\mathcal{T})$ respectively. Typically, areas with higher temperatures tend to exhibit more detailed features, implying that pixel intensity can reflect thermal levels. This suggests that the intensity of a pixel can serve as a gauge for the thermal target. Hence, we define $E(\mathcal{P})$, which represents the expected thermal intensity within a patch, to compute the average thermal intensity

$$E(\mathcal{P}) = \frac{1}{h \times w} \sum_{i=1}^{h \times w} T(\mathbf{p}_i) \tag{7}$$

where h, w denotes the height and width of the formed patch. The intensity within this range is confined between 0 and 1, which is conveniently used as a weighting factor for the structural loss function. Then the structural loss can be written as

$$\mathcal{L}_{str}(\Theta) = E(\mathcal{P})(1 - \mathrm{HSSIM}(G(\mathcal{T})\ G(\mathcal{T}))) \tag{8}$$

At this point, we can obtain the total loss, which is a combination of

$$\mathcal{L}_{tot} = \mathcal{L}_{pix} + \mathcal{L}_{str} \tag{9}$$

D. Implementation Details

Our code is based on nerfacto model proposed by nerfstudio [37] applying the thermal mapping and thermal constraint based on heat information. We train the model for $3 \cdot 10^4$ iterations on one NVIDIA RTX 3090 with the default optimizer and hyper-parameters as in nerfacto, the training usually converges in about 15 to 20 minutes. And we set the sliding window's kernel size to 4 and the stride to 4, enabling us to compute the mean HSSIM without overlapping pixels.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we validate the efficacy of Thermal-NeRF through experiments on a self-collected IR dataset of indoor environments, focusing on novel view synthesis and 3D object reconstruction. We conduct a detailed comparison of Thermal-NeRF with representative methods including the original NeRF [3], Mip-NeRF 360 [38], and DVGO [39].

A. Self-Collected IR Dataset

To assess the Thermal-NeRF with IR cameras, due to the current lack of IR datasets for NeRF, we built a new dataset utilizing the Optris PI 450i IR camera and will make it publicly available. The ground truth camera poses were recorded using VICON, a Motion capture (MoCap) system. The camera is equipped with several infra-reflective markers to form a rigid body, positioning it within the MoCap's coordinate system W. To increase the precision of camera pose, we conducted a hand-eye calibration, achieving the transformation from T_R^W to T_C^W, T_R^W represents the pose of the rigid body, and T_C^W represents the pose of the camera, both relative to the MoCap system's coordinate system W, where R and C correspond to the coordinate systems of the rigid body and the camera, respectively. For every image I_k and the corresponding pose T_k captured by the MoCap systems, we used spherical linear interpolation

(a) Low-Light Condition (b) Light-Changing Condition (c) Smoky Condition

Fig. 2. This illustration highlights that our actual sequences were documented in demanding conditions, including low and fluctuating lighting, as well as smoke. The RGB images, taken by camera RealSense D435, demonstrate the visual outcomes. The frames marked in red represents the same area as captured by both IR and RGB cameras, albeit with differing resolutions and fields of view.

TABLE I

QUANTITATIVE EVALUATION ON SELF-COLLECTED SEQUENCES UNDER LOW-LIGHT CONDITIONS OF DIFFERENT MODELS

Model	Sequnce 1 train/test views: 243/26			Sequence 2 train/test views: 186/20			Sequence 3 train/test views: 187/20		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [3]	27.57	0.83	0.47	21.32	0.82	0.45	26.22	0.75	0.26
Mip-NeRF 360 [38]	27.96	0.84	0.45	20.92	0.82	0.44	26.15	0.75	0.24
DVGO [39]	19.18	0.79	0.57	13.16	0.71	0.63	15.83	0.65	0.41
Thermal-NeRF w/o pose [32]	28.19	0.88	0.37	17.60	0.85	0.37	25.93	0.77	0.21
Thermal-NeRF	**32.41**	**0.90**	**0.36**	**29.79**	**0.88**	**0.37**	**30.41**	**0.79**	**0.20**

TABLE II

QUANTITATIVE EVALUATION ON SELF-COLLECTED SEQUENCES UNDER LIGHT-CHANGING CONDITIONS OF DIFFERENT MODELS

Model	Sequnce 1 train/test views: 195/21			Sequence 2 train/test views 213/23			Sequence 3 train/test views 158/17		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [3]	27.62	0.83	0.48	27.48	0.76	0.25	30.57	0.84	0.39
Mip-NeRF 360 [38]	27.63	0.82	0.46	27.54	0.76	0.24	31.08	0.84	0.38
DVGO [39]	18.69	0.78	0.63	18.50	0.68	0.36	20.86	0.74	0.50
Thermal-NeRF w/o pose [32]	28.30	0.89	0.40	27.09	0.87	0.22	29.69	0.91	0.38
Thermal-NeRF	**31.27**	**0.89**	**0.38**	**32.89**	**0.89**	**0.22**	**34.17**	**0.93**	**0.37**

TABLE III

QUANTITATIVE EVALUATION ON SELF-COLLECTED SEQUENCES UNDER SMOKY CONDITIONS OF DIFFERENT MODELS

Model	Sequnce 1 train/test views: 207/22			Sequence 2 train/test views: 223/24			Sequence 3 train/test views: 337/37		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [3]	26.80	0.83	0.38	24.55	0.84	0.46	25.76	0.84	0.45
Mip-NeRF 360 [38]	27.17	0.84	0.38	24.58	0.84	0.45	26.10	0.86	0.44
DVGO [39]	14.28	0.70	0.64	15.10	0.74	0.64	18.76	0.80	0.55
Thermal-NeRF w/o pose [32]	26.78	0.88	0.30	23.36	0.85	0.39	25.59	0.87	0.36
Thermal-NeRF	**32.31**	**0.89**	**0.29**	**25.88**	**0.85**	**0.35**	**26.78**	**0.87**	**0.35**

for the rotational component of T_k, and linear interpolation for the translational component of T_k.

Our IR dataset features three distinct scenarios, each designed to replicate conditions similar to those in visually degraded environments: environments with low lighting, fluctuating lighting and smoke, illustrated by Fig. 2. We simulated thermal sources using heated objects like cups of hot water and heat-emitting electronic devices. Under these varied conditions, our IR cameras demonstrated consistent performance, in contrast to RGB cameras, which experienced diminished functionality in the visible spectrum. For comprehensive scene coverage, we performed 360-degree photography around each indoor scenario, comprising four sequences per scenario, with each sequence containing between 200 to 300 images. Our dataset will be open-sourced for further research.

B. Experimental Setup

Our experimental framework encompasses novel view synthesis and 3D object reconstruction to comprehensively evaluate the effectiveness and robustness of our proposed method. Additionally, we have verified the critical importance of pose refinement based on ground truth camera poses in enhancing the performance of our approach. In assessing novel view synthesis, we employ widely recognized evaluation metrics, including the Peak Signal-to-Noise Ratio (PSNR) [33], SSIM [34], and the Learned Perceptual Image Patch Similarity (LPIPS) [40] using VGGNet [41], to ensure a thorough and objective assessment. Notably, we intentionally transform IR images into pseudo-color representations at the evaluating stage to improve our analysis by highlighting features that are less visible in grayscale, the jet colormap array is adopted for color conversion, this conversion process does not impact the assessment outcomes and aligns with conventional NeRF evaluation standards. For the 3D object reconstruction aspect, our focus is on reconstructing thermal objects within the scene, we utilize the marching cube method for mesh extraction [42], enabling a qualitative evaluation of the reconstructed objects.

C. Novel View Synthesis

We employ both quantitative and qualitative assessments to evaluate the effectiveness of Thermal-NeRF of novel view synthesis experiments. Quantitative and qualitative analysis, leveraging three challenging scenes from our custom dataset, is succinctly presented in Tables I, II, III and

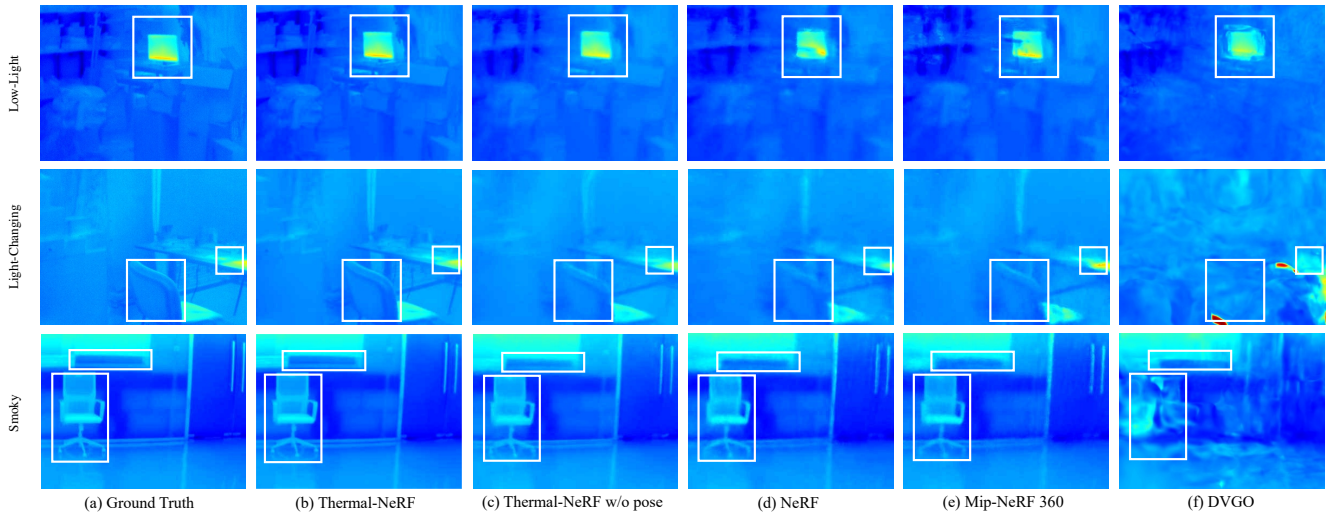

Fig. 3. Qualitative evaluation on self-collected sequences under challenging conditions. To enhance the visual assessment of the effects, IR images are intentionally transformed into pseudo-color representations by employing the jet colormap array for the color conversion process. Model: original NeRF [3], Mip-NeRF 360 [38], DVGO [39], and Thermal-NeRF. Specifically, we present results of Thermal NeRF without pose refinement to underscore the significance of pose optimization in achieving optimal outcomes.

Fig. 3 respectively. In our findings, Thermal-NeRF excels at capturing thermal intensities with remarkable precision, crucial for comprehensive thermal analysis. Specifically, our experiments illuminate that integrating pose refinement with a structural thermal constraint improves image quality, outperforming the results achieved by applying either technique in isolation, which enables Thermal-NeRF to attain its best performance. In contrast to competing models, Thermal-NeRF produces images that are sharp and clear scenes including the details, avoiding the common issue of blurriness. While models like original NeRF and Mip-NeRF 360 capture essential scene features, they struggle with achieving clear images. DVGO, aiming for faster rendering speeds, compromises on rendering accuracy. The superiority is also consistently reflected in our PSNR metrics, where our model surpasses others. Our approach also achieves superior structural and textural accuracy, essential for maintaining scene integrity, as indicated by the SSIM scores. In terms of perceptual alignment with human vision, especially in understanding the subtleties of thermal images, our model shows notable advancements. The LPIPS scores highlight this improvement, demonstrating our model's enhanced ability to interpret semantic details more effectively.

Moreover, although the PSNR values for Thermal NeRF might occasionally fall slightly below those of the more time-consuming methods like NeRF and Mip-NeRF 360 when pose refinement is not used, this phenomenon can be rationalized by considering the distinct focus of our method. PSNR is a metric that emphasizes pixel-level differences, which does not necessarily lead to better visual outcomes. In contrast, our approach prioritizes the visual quality of the images, see Fig. 3, as proven by the superior LPIPS, SSIM scores. These metrics assess structural integrity and perceptual similarity, respectively, and are more aligned with human visual perception by prioritizing aspects that contribute to a visually appealing image over mere pixel accuracy. This underscores the importance of the integration of structural constraints, which are key to enhancing rendering quality.

D. 3D Object Reconstruction

To assess Thermal-NeRF's efficacy in 3D reconstruction, we employ marching cube algorithm for mesh extraction of localized heat sources within the scenes, the specific parameters of the algorithm are adjusted based on each model. In our evaluation, we tested Thermal-NeRF against models including original NeRF, Mip-NeRF 360 and DVGO. The resulting meshes, simulating heat sources with objects like a cup and a kettle containing hot water, are flawlessly shaped and hole-free, underscoring the model's precision, as depicited in (a) of Fig. 4. Notably, while NeRF and Mip-NeRF 360 demonstrated satisfactory performance at the image level as discussed in Sec. IV-C, their corresponding meshes were similarly plagued by significant noise. This underscores their limitations in precisely capturing the scene's depth information.

We conjecture that the inherent sparsity and low contrast typical of IR images pose significant obstacles for general models, particularly affecting their ability to accurately estimate density. This often results in outputs marred by considerable noise. In contrast, by incorporating structural constraints, Thermal-NeRF adeptly navigates these challenges. These constraints enable the model to concentrate on areas with pronounced thermal activity, effectively prioritizing the reconstruction of critical heat-emitting objects within a scene.

E. Ablation Studies

In our study, we conduct a series of ablation experiments by individually altering components of our method. We evaluate the performance impact of these changes using average image metrics across six self-collected sequences in three

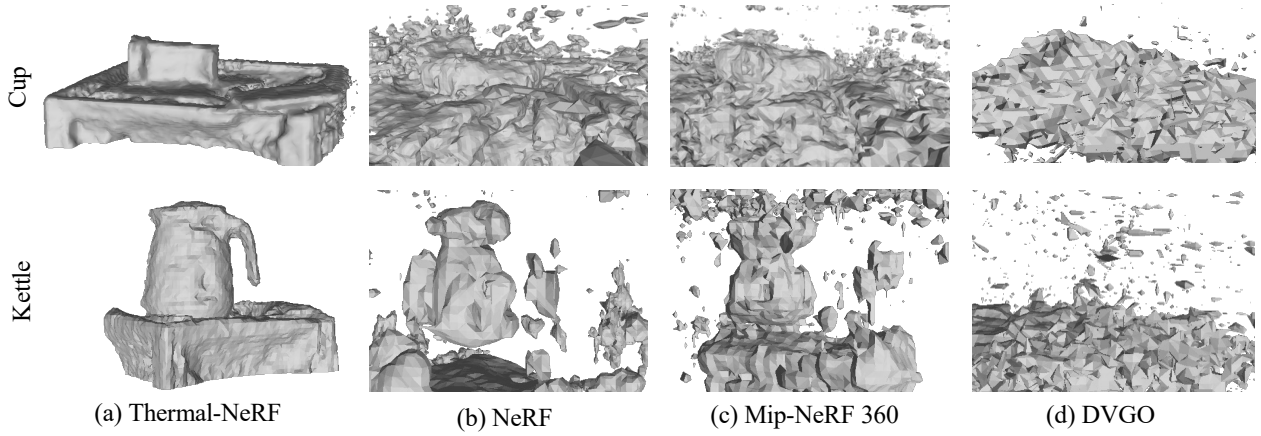

Fig. 4. Examples of mesh reconstruction of heat source objects in a scene. Explicit meshes of cup and kettle exported by Thermal-NeRF and other models are shown respectively.

TABLE IV
QUANTITATIVE ABLATION STUDY ON THE SELF-COLLECTED DATASET.

Thermal-NeRF method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o thermal mapping	22.64	0.71	0.69
w/o structural thermal constraint	24.70	0.79	0.64
Thermal-NeRF (default)	**30.52**	**0.88**	**0.41**

challenging environments. The results of these experiments are illustrated in the accompanying Tab. IV and Fig. 5.

Our first ablation involves the mapping method. We replace our thermal mapping approach with a trivial pixel mapping method, which involves scaling the pixel values of single images from a 16-bit format to an 8-bit format using min-max normalization. This modification results in a significant drop in performance, characterized by blurring and ghosting effects. This outcome is anticipated, as such a change violates the thermal consistency inherent in IR imaging. Without mapping the images to a uniform temperature range, pixel values for the same spatial point vary with the viewpoint, leading to inconsistencies. This effect can be shown in Fig. 5, where the trivial pixel mapping results in a substantially different pixel distribution compared to results applying thermal mapping.

Subsequently, we ablate the thermal structural constraint component of our method, Equation 8. This alteration results in a degradation of image quality, with noticeable ghosting in detail features and overall blurring. It becomes clear that pixel-level loss is insufficient for IR imaging, which typically have low contrast. The introduction of structure thermal constraint significantly mitigates artifacts in uniformly colored areas and effectively recovers detailed features in areas with concentrated heat, see Fig. 5.

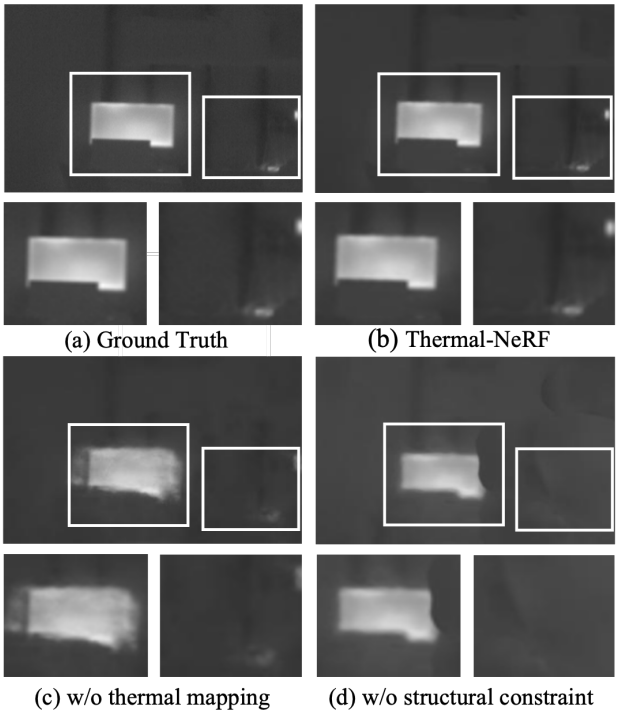

Fig. 5. Ablation qualitative example. Here we show renderings from different Thermal-NeRF ablation variants. The full model produces the best results. We zoom in on crops to highlight differences in the rendered images.

V. CONCLUSIONS

In this study, we introduce Thermal-NeRF, the first approach for reconstructing neural radiance fields exclusively from IR imaging, particularly beneficial in visually degraded robotics scenarios, including low-light, light-changing, and smoky environments. Thanks to the proposed combination of thermal mapping and a structural thermal constraint, Thermal-NeRF outperforms existing methods on our custom IR dataset, delivering improved quality in both image rendering and mesh reconstruction of heat sources. Thus, this paper extends the spectrum of practical IR-based techniques with a 3D representation learning approach. Future work could involve integrating depth supervision to enable comprehensive scene-level reconstruction.

REFERENCES

[1] Z. Ma and S. Liu, "A review of 3d reconstruction techniques in civil engineering and their applications," *Advanced Engineering Informatics*, vol. 37, pp. 163–174, 2018.

[2] Z. Kang, J. Yang, Z. Yang, and S. Cheng, "A review of techniques for 3d reconstruction of indoor environments," *ISPRS International Journal of Geo-Information*, vol. 9, no. 5, p. 330, 2020.

[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," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[4] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[5] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 303–312.

[6] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Mesh optimization," in *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, 1993, pp. 19–26.

[7] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron, "Nerf in the dark: High dynamic range view synthesis from noisy raw images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 190–16 199.

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7210–7219.

[9] T. Fujitomi, K. Sakurada, R. Hamaguchi, H. Shishido, M. Onishi, and Y. Kameda, "Lb-nerf: light bending neural radiance fields for transparent medium," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2142–2146.

[10] M. Vollmer, "Infrared thermal imaging," in *Computer Vision: A Reference Guide*. Springer, 2021, pp. 666–670.

[11] K. Ko, K. Shim, K. Lee, and C. Kim, "Large-scale benchmark for uncooled infrared image deblurring," *IEEE Sensors Journal*, 2023.

[12] X. Kuang, X. Sui, Y. Liu, Q. Chen, and G. Gu, "Single infrared image enhancement using a deep convolutional neural network," *Neurocomputing*, vol. 332, pp. 119–128, 2019.

[13] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Information fusion*, vol. 24, pp. 147–164, 2015.

[14] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2320–2327.

[15] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural volumes: Learning dynamic renderable volumes from images," *arXiv preprint arXiv:1906.07751*, 2019.

[16] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[17] J. Deng, Q. Wu, X. Chen, S. Xia, Z. Sun, G. Liu, W. Yu, and L. Pei, "Nerf-loam: Neural implicit representation for large-scale incremental lidar odometry and mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8218–8227.

[18] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[19] N. Kulkarni, J. Johnson, and D. F. Fouhey, "Directed ray distance functions for 3d scene reconstruction," in *European Conference on Computer Vision*. Springer, 2022, pp. 201–219.

[20] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or, "Latent-nerf for shape-guided generation of 3d shapes and textures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 663–12 673.

[21] Y. Xiao, Y. Zhao, Y. Xu, and S. Gao, "Resnerf: Geometry-guided residual neural radiance field for indoor scene novel view synthesis," *arXiv preprint arXiv:2211.16211*, 2022.

[22] J. Tang, H. Zhou, X. Chen, T. Hu, E. Ding, J. Wang, and G. Zeng, "Delicate textured mesh recovery from nerf via adaptive surface refinement," *arXiv preprint arXiv:2303.02091*, 2023.

[23] F. Bao, X. Wang, S. H. Sureshbabu, G. Sreekumar, L. Yang, V. Aggarwal, V. N. Boddeti, and Z. Jacob, "Heat-assisted detection and ranging," *Nature*, vol. 619, no. 7971, pp. 743–748, 2023.

[24] Y. He, B. Deng, H. Wang, L. Cheng, K. Zhou, S. Cai, and F. Ciampa, "Infrared machine vision and infrared thermography with deep learning: A review," *Infrared physics & technology*, vol. 116, p. 103754, 2021.

[25] J.-H. He, D.-P. Liu, C.-H. Chung, and H.-H. Huang, "Infrared thermography measurement for vibration-based structural health monitoring in low-visibility harsh environments," *Sensors*, vol. 20, no. 24, p. 7067, 2020.

[26] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[27] R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, "Vif-net: An unsupervised framework for infrared and visible image fusion," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 640–651, 2020.

[28] Y. Ma, Y. Wang, X. Mei, C. Liu, X. Dai, F. Fan, and J. Huang, "Visible/infrared combined 3d reconstruction scheme based on nonrigid registration of multi-modality images with mixed features," *IEEE Access*, vol. 7, pp. 19 199–19 211, 2019.

[29] S. Lang and K. Jäger, "3d scene reconstruction from ir image sequences for image-based navigation update and target detection of an autonomous airborne system," in *Infrared Technology and Applications XXXIV*, vol. 6940. SPIE, 2008, pp. 535–543.

[30] M. Poggi, P. Z. Ramirez, F. Tosi, S. Salti, S. Mattoccia, and L. Di Stefano, "Cross-spectral neural radiance fields," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 606–616.

[31] S. Katragadda, W. Lee, Y. Peng, P. Geneva, C. Chen, C. Guo, M. Li, and G. Huang, "Nerf-vins: A real-time neural radiance field map-based visual-inertial navigation system," *arXiv preprint arXiv:2309.09295*, 2023.

[32] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "Nerf–: Neural radiance fields without known camera parameters," *arXiv preprint arXiv:2102.07064*, 2021.

[33] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on communications*, vol. 43, no. 12, pp. 2959–2965, 1995.

[34] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[36] Z. Xie, X. Yang, Y. Yang, Q. Sun, Y. Jiang, H. Wang, Y. Cai, and M. Sun, "S3im: Stochastic structural similarity and its unreasonable effectiveness for neural fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 024–18 034.

[37] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23, 2023.

[38] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5470–5479.

[39] C. Sun, M. Sun, and H.-T. Chen, "Improved direct voxel grid optimization for radiance fields reconstruction," *arXiv preprint arXiv:2206.05085*, 2022.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[42] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353.