Implicit Neural Fusion of RGB and Far-Infrared 3D Imagery for Invisible Scenes

Xiangjie Li[1], Shuxiang Xie[1], Ken Sakurada[2], Ryusuke Sagawa[2], Takeshi Oishi[1]

***Abstract*— Optical sensors, such as the Far Infrared (FIR) sensor, have demonstrated advantages over traditional imaging. For example, 3D reconstruction in the FIR field captures the heat distribution of a scene that is invisible to RGB, aiding various applications like gas leak detection. However, less texture information and challenges in acquiring FIR frames hinder the reconstruction process. Given that implicit neural representations (INRs) can integrate geometric information across different sensors, we propose Implicit Neural Fusion (INF) of RGB and FIR for 3D reconstruction of invisible scenes in the FIR field. Our method first obtains a neural density field of objects from RGB frames. Then, with the trained object density field, a separate neural density field of gases is optimized using limited view inputs of FIR frames. Our method not only demonstrates outstanding reconstruction quality in the FIR field through extensive experiments but also can isolate the geometric information of the invisible, offering a new dimension of scene understanding.**

I. INTRODUCTION

In light of rapid advancements in sensor technology, recent optical sensors have emerged as a superior alternative to traditional imaging methodologies, especially in scenarios with insufficient visible light. One example is the far infrared (FIR) sensor, which can observe thermal data in the 3 to 100 μm range of the electromagnetic spectrum [1]. The utility of FIR sensors has extended to many areas, such as the identification of humans or animals, facilitation of behavioral observations, and anomaly detection. Consequently, these sensors are frequently equipped and utilized by both aerial drones and ground-based robots.

3D scene reconstruction in the FIR domain utilizes thermal information to map a scene's heat distribution, facilitating a range of applications that gain from thermal insights. For example, it can identify gas leaks in industrial environments by pinpointing areas with abnormal temperature fluctuations. Additionally, three-dimensional FIR data can help guide humans and robots away from hazardous zones, thereby improving safety and operational efficiency.

However, the reconstruction of a scene solely from FIR frames faces challenges due to their typically lower texture details compared to RGB frames. The distinct radiometric characteristics of the FIR signal further add to the complexity of the reconstruction task. Moreover, in time-sensitive scenarios such as gas leaks, adequate scene acquisition becomes almost impossible, thereby complicating the process of 3D scene reconstruction. As a result, conventional multi-view strategies [2], [3], [4] developed for RGB imagery often fail to present satisfying results. Regarding existing thermal-RGB(D) fusion methods, which frequently presuppose the presence of the surface of objects, their efficacy diminishes when handling targets such as gases that lack a solid surface.

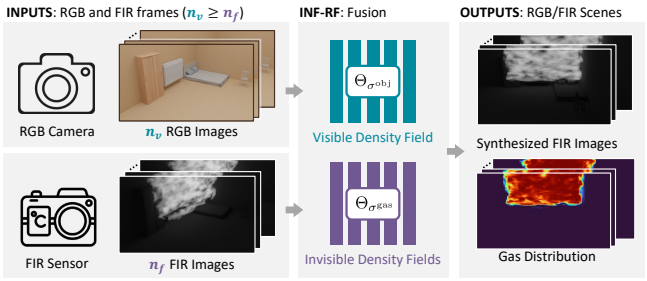

Fig. 1. Overview of our method. It receives full and limited views of RGB and FIR frames to obtain the neural density fields of visible objects and invisible gases respectively. It can reconstruct the scene through novel view synthesis and estimate the gas distribution.

Following the trend set by neural radiance fields (NeRF) [5], our approach utilizes implicit neural representations (INRs) for addressing these challenges through a sensor fusion technique combining RGB and FIR data. Through the application of INR, we can effectively capture and model scenes based on density fields, thereby offering a more appropriate framework for representing gases that are hard to observe. Moreover, the implicit and differentiable characteristics of INR facilitate the convenient fusion of data originating from diverse sensor modalities, as evidenced by [6], [7]. Also, by comparing INRs generated from both RGB and FIR data, we can distinguish and examine geometric features in FIR that remain invisible in the visible spectrum, introducing a novel layer of scene comprehension.

In this study, we present a novel methodology, termed INF-RF, aimed at applying both RGB and FIR datasets to effectively capture invisible scenes, such as airborne gases. Our method leverages a comprehensive collection of RGB frames in conjunction with a limited set of FIR frames as input data. Figure 1 provides an overview of the system. Initially, we utilize the RGB frames to construct a neural density field encapsulating the visible parts within the scene. Subsequently, this field is fused with the FIR frames to produce an integrated density field encompassing both visible and FIR characteristics of the scene. This novel framework facilitates the detection and separation of the

[1]X. Li, S. Xie and T. Oishi are with Institute of Industrial Science, The University of Tokyo, 153-8505, Tokyo, Japan {xiangjie, shxxie, oishi}@cvl.iis.u-tokyo.ac.jp

[2]K. Sakurada and R. Sagawa are with National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, 305-8568, Japan {k.sakurada, ryusuke.sagawa}@aist.go.jp

neural density field from the invisible components. Furthermore, our methodology enables novel view synthesis in the reconstructed FIR scenes.

Our contributions are summarized as follows:

- We introduce a sensor fusion method using INR that integrates geometric and thermal information from RGB and FIR sensors.
- We propose and validate a volume rendering equation for FIR that can distinguish and separate RGB and FIR components inside the scenes.
- We present an encoding technique in our framework targeting at few-shot scenarios that enhances the quality of reconstructed fields.

In this work, Sec. II comprehensively reviews pertinent literature and contextualizes the present study within the existing scholarly discourse. Sec. III delineates the proposed framework, with Section IV explaining the rendering equation. The methodology applicable to the few-shot scenario is discussed in Sec. V. Sec. VI presents and interprets the experimental findings, while Sec. VII encapsulates the conclusions of the paper.

II. RELATED WORK

This section gives a brief review of studies on FIR-RGB sensor fusion techniques and 3D reconstruction with sensor fusions based on INRs.

A. FIR-RGB Sensor Fusion

The fusion of FIR and RGB sensors finds extensive applications in areas such as person identification, autonomous driving, and rescue robots. Wu et al. introduced a deep zero-padding approach to identify common features between RGB and FIR images for person identification [8]. The utilization of a dual-stream network was proposed to bridge the sensor disparities [9], [10]. Chen et al. advanced this by enabling automatic selection of critical body parts for person identification through RGB-FIR fusion using a data-driven strategy [11]. He et al. adopted the Vision Transformer with added side information to extract distinguishing features, while Lu et al. designed their model to learn transferable features across sensors for person identification [12], [13].

In the domain of autonomous driving, Zhang et al. introduced an innovative model that combines an RGB/FIR camera system capable of producing pixel-wise aligned and temporally synchronized RGB and FIR frames [14]. A composite image dataset comprising RGB and FIR frames was developed for road detection tasks [15]. Their work also highlighted the effectiveness of FIR-RGB fusion through a fully convolutional neural network.

Thermal and RGB fusion proves beneficial in rescue operations as well. [16] achieved environmental recognition through semantic segmentation of RGB frames and identified individuals using object detection in FIR frames. While this system was primarily deployed on drones, its utility extends to ground-based vehicles within rescue contexts [17].

Despite considerable progress in various applications, the capability of FIR and RGB sensor fusion in 3D reconstruction has yet to be fully explored. RGB data offers abundant texture and depth details, unlike FIR data, which provides thermal insights invisible to the human eyes. Integrating these two types of sensors could lead to a more precise and holistic comprehension of scenes.

B. INR-based Sensor Fusion

Beyond traditional techniques such as point-based [18], [19], voxel-based [20], and mesh-based methods [21], there has been a notable shift towards using neural networks for three-dimensional scene representation in recent times. Park et al. [22] and Mescheder et al. [23] have each introduced methods for implicitly defining surfaces through signed distance functions and occupancy networks, respectively. Furthermore, Mildenhall et al. introduced NeRF, utilizing volume rendering to embed appearance and features within neural networks [5]. Building on NeRF's foundation, numerous studies like Mip-NeRF [24], Zipnerf [25], Instant-NGP [26], among others, have aimed to enhance NeRF's performance and speed.

NeRF has been expanded to include sensor fusion techniques as well. INF [6] presented a method based on NeRF for fusing LiDAR and RGB data while adjusting the extrinsic parameters between sensors. Similar to INF, MOISST [7] and SOAC [27] adopted approaches using INR for the spatio-temporal calibration between LiDAR and cameras. Cloner integrated LiDAR and RGB cameras to reconstruct expansive, open-air scenes [28]. In a similar manner, [29] combined RGB and depth information to achieve accurate scene reconstruction through INRs. In the context of combining RGB and thermal sensors, multimodal NeRF merged RGB, infrared, and point cloud data to precisely represent the 3D shapes of objects [30]. Similarly, X-NeRF brought together RGB, infrared, and multi-spectral data for novel view synthesis and fine-tuned the camera poses using normalized coordinates across devices [31].

Nevertheless, the mentioned studies do not target scenes invisible in the FIR spectrum, assuming that the geometric information is consistent across different sensor types. In contrast, recognizing that FIR cameras can capture elements invisible to RGB cameras, our method appreciates the variance in geometric data between RGB and FIR sensors and investigates their capacity for reciprocal information exchange.

III. IMPLICIT NEURAL FUSION OF RGB AND FIR

We propose an Implicit Neural Fusion system, which integrates RGB and FIR frames to achieve 3D reconstruction in FIR field that can recognize the invisible scene. Fig. 2 demonstrates the structure of our framework. The inputs are two distinct types of image data: full views of RGB frames $\left\{\mathcal{F}_1^{\mathrm{RGB}}, \mathcal{F}_2^{\mathrm{RGB}} \ldots, \mathcal{F}_{n_\mathrm{v}}^{\mathrm{RGB}}\right\}$ and limited views of FIR frames $\left\{\mathcal{F}_1^{\mathrm{FIR}}, \mathcal{F}_2^{\mathrm{FIR}} \ldots, \mathcal{F}_{n_\mathrm{f}}^{\mathrm{FIR}}\right\}$. n_v is the number of RGB frames, and n_f is the number of FIR frames. Usually $n_\mathrm{v} \geq n_\mathrm{f}$, but not necessarily. The outputs are neural density fields of the

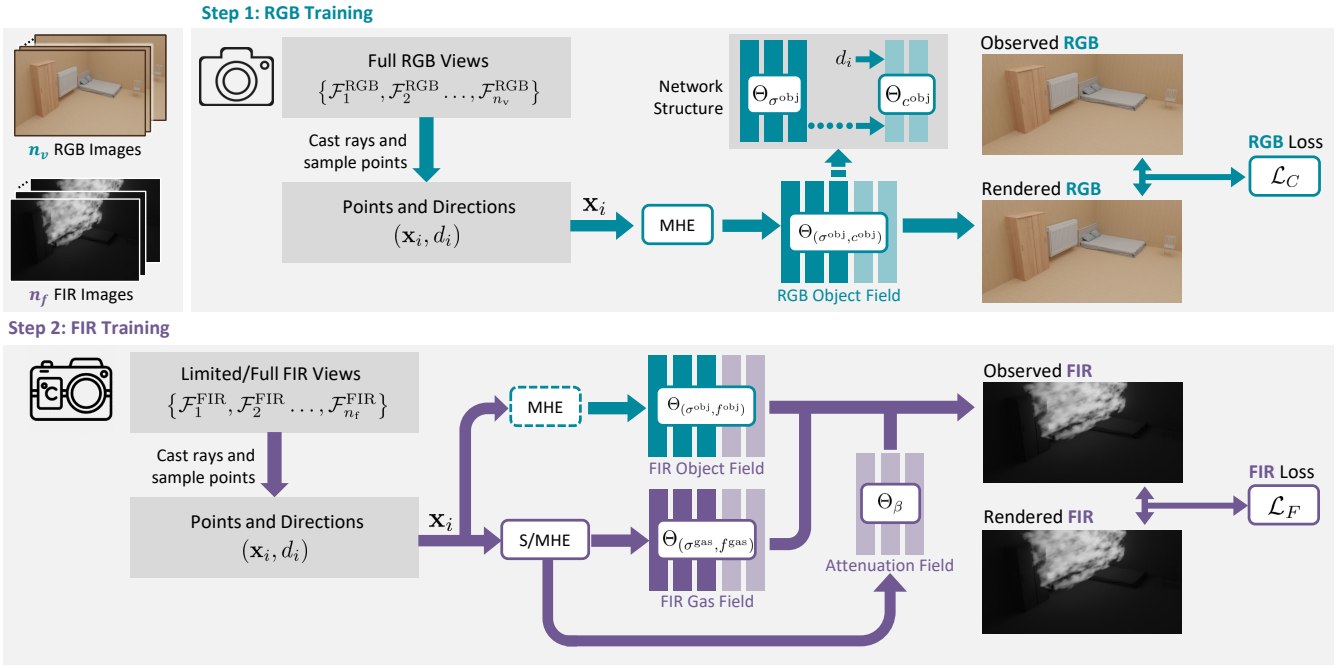

Fig. 2. General workflow of our method. We obtain a neural density field of objects from the training of RGB frames. Then, this geometric information of objects is referred to in the FIR training. By adding a new density field of gases, we can isolate the geometric distribution of gases and objects in the FIR scene.

visible and FIR scenes represented by MLPs: $\Theta_{\sigma^{\text{obj}}}$ and $\Theta_{\sigma^{\text{gas}}}$. We assume that the camera poses of FIR and RGB frames are known and in the same world coordinate.

A. Neural Modeling of RGB Scene

We describe the first step of our framework, which models the RGB scene using INRs. Assume that we sample the points $\mathbf{x}_i \in \mathbb{R}^3 (i = 1, 2, ..., n_{\text{s}})$ on the rays that pass through the pixels on the input RGB images. n_{s} is the number of sampling points along a ray. Consider $\mathcal{H}(\mathbf{x}_i; \theta)$ is the hash encoding function [26] that encodes a sampled point $\mathbf{x}_i$ with trainable encoding parameters θ . $\mathcal{H}(\mathbf{x}_i; \theta)$ is fed to an MLP $\Theta_{\sigma^{\text{obj}}}$ to obtain the neural density field of visible objects $\sigma_i^{\text{obj}} \in \mathbb{R}$: $\Theta_{\sigma^{\text{obj}}}(\mathcal{H}(\mathbf{x}_i; \theta)) \rightarrow \sigma_i^{\text{obj}}$. $\Theta_{\sigma^{\text{obj}}}$ is appended with a heading MLP $\Theta_{c^{\text{obj}}}$ which receives viewing direction $\mathbf{d}_i$ as additional input. $\Theta_{c^{\text{obj}}}$ learns the color $\mathbf{c}_i^{\text{obj}} \in \mathbb{R}^3$ of the point: $\Theta_{c^{\text{obj}}}(\mathbf{d}_i) \rightarrow \mathbf{c}_i^{\text{obj}}$. After obtaining the density and color of all n_{s} sample points along a ray, we render the pixel color of the ray $\widehat{C}^{\text{RGB}} \in \mathbb{R}^3$ by volume rendering [5]:

$$\widehat{C}^{\text{RGB}} = \sum_{i=1}^{n_{\text{s}}} T_i^{\text{RGB}} \left(1 - \exp\left(-\sigma_i^{\text{obj}} \delta_i\right)\right) \mathbf{c}_i^{\text{obj}}, \tag{1}$$

$$T_i^{\text{RGB}} = \exp\left(-\sum_{j=1}^{i-1} \sigma_j^{\text{obj}} \delta_j\right), \tag{2}$$

where δ_i denotes the distance between adjacent samples $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$. During the training, we optimize θ and the parameters of $\Theta_{\sigma^{\text{obj}}}$, $\Theta_{c^{\text{obj}}}$ through the difference between the rendered color and the input color $C^{\text{RGB}} \in \mathbb{R}^3$ with the following loss function:

$$\mathcal{L}_C = \left(\widehat{C}^{\text{RGB}} - C^{\text{RGB}}\right)^2. \tag{3}$$

B. Neural Modeling of FIR Scene

After the density field of the visible scene is learned in the above process, a heading MLP $\Theta_{f^{\text{obj}}}$ following $\Theta_{\sigma^{\text{obj}}}$ learns the signal intensity of objects $f_i^{\text{obj}} \in \mathbb{R}$ in the FIR field : $\Theta_{f^{\text{obj}}}(\mathbf{d}_i) \rightarrow f_i^{\text{obj}}$. Similarly, $\Theta_{\sigma^{\text{gas}}}$ appended with $\Theta_{f^{\text{gas}}}$ learns the density $\sigma_i^{\text{gas}} \in \mathbb{R}$ and signal intensity $f_i^{\text{gas}} \in \mathbb{R}$ of gases separately: $\Theta_{\sigma^{\text{gas}}}\left(\tilde{\mathcal{H}}(\mathbf{x}_i; \theta)\right) \rightarrow \sigma_i^{\text{gas}}$ and $\Theta_{f^{\text{gas}}}(\mathbf{d}_i) \rightarrow f_i^{\text{gas}}$. Here, $\tilde{\mathcal{H}}$ is the sliding multi-resolution hash encoding function explained later in Sec. V-B. We also use an MLP Θ_{β_i} for the attenuation field to model the attenuation of the intensity $\beta_i \in \mathbb{R}$ by scattering: $\Theta_{\beta_i}(\tilde{\mathcal{H}}(\mathbf{x}_i; \theta), \mathbf{d}_i) \rightarrow \beta_i$.

Then, we obtain the pixel intensity $\widehat{C}^{\text{FIR}} \in \mathbb{R}$ by volume rendering (with Eq. (15)) and optimize the parameters of $\Theta_{f^{\text{obj}}}$, $\Theta_{\sigma^{\text{gas}}}$, $\Theta_{f^{\text{gas}}}$, and Θ_{β_i} through the difference between the rendered intensity and the input intensity $C^{\text{FIR}} \in \mathbb{R}$. Our loss function for FIR images is:

$$\mathcal{L}_F = (\widehat{C}^{\text{FIR}} - C^{\text{FIR}})^2. \tag{4}$$

IV. Volume Rendering for RGB-FIR Fusion

In this section, we derive the rendering equation in the presence of the gas in front of background objects to obtain the pixel intensity for Eq. (4).

A. Volume Rendering

We consider a general setting in which the light travels across the invisible area (usually gas) and has a significant

impact on the captured intensity. For a ray traveling from the nearest distance $t_n \in \mathbb{R}$ to the farthest distance $t_f \in \mathbb{R}$, we can derive the observed color by volume rendering [32] while integrating the components of objects and gases, as follows:

$$\widehat{C}^{\text{FIR}} = \int_{t_n}^{t_f} T^{\text{FIR}}(t) \left(\sigma^{\text{obj}}(t) f^{\text{obj}}(t) + \sigma^{\text{gas}}(t) f^{\text{gas}}(t)\right) dt, \tag{5}$$

$$T^{\text{FIR}}(t) = \exp\left(-\int_{t_n}^{t} \left(\sigma^{\text{obj}}(t) + \sigma^{\text{gas}}(t)\right) dt\right). \tag{6}$$

σ^{obj} is the density of visible objects obtained in RGB training (Sec. III-A) while σ^{gas} represents the density of gas captured by the FIR camera. f^{obj} and f^{gas} represent the intensity of FIR signals of objects and gases.

In the discretized version, we get the color $\hat{C}^{\text{FIR}}$ at the point $\mathbf{x}_i$ according to [33], as follows:

$$\widehat{C}_i^{\text{FIR}} = T_i^{\text{FIR}} \left(1 - e^{-\left(\sigma_i^{\text{obj}} + \sigma_i^{\text{gas}}\right)\delta_i}\right) \frac{\sigma_i^{\text{obj}} f_i^{\text{obj}} + \sigma_i^{\text{gas}} f_i^{\text{gas}}}{\sigma_i^{\text{obj}} + \sigma_i^{\text{gas}}}, \tag{7}$$

$$T_i^{\text{FIR}} = \exp\left(-\sum_{j=0}^{i-1} \left(\sigma_j^{\text{obj}} + \sigma_j^{\text{gas}}\right) \delta_j\right). \tag{8}$$

Separating the volume rendering equation into *object* and *gas* components, we get the rendering equation of a ray, as follows:

$$\widehat{C}^{\text{FIR}} = \sum_{i=0}^{n_s - 1} \widehat{C}_i^{\text{obj}} + \sum_{i=0}^{n_s - 1} \widehat{C}_i^{\text{gas}}, \tag{9}$$

$$\widehat{C}_i^{\text{obj}} = T_i^{\text{FIR}} \left(1 - e^{-\left(\sigma_i^{\text{obj}} + \sigma_i^{\text{gas}}\right)\delta_i}\right) \frac{\sigma_i^{\text{obj}} f_i^{\text{obj}}}{\sigma_i^{\text{obj}} + \sigma_i^{\text{gas}}}, \tag{10}$$

$$\widehat{C}_i^{\text{gas}} = T_i^{\text{FIR}} \left(1 - e^{-\left(\sigma_i^{\text{obj}} + \sigma_i^{\text{gas}}\right)\delta_i}\right) \frac{\sigma_i^{\text{gas}} f_i^{\text{gas}}}{\sigma_i^{\text{obj}} + \sigma_i^{\text{gas}}}. \tag{11}$$

To simplify our model, we assume that the object is opaque in the scene. Thus, the object density along the ray is close to zero except for a high peak on the object's surface. On the other hand, the gas is semi-transparent and has a low non-zero density. Based on these considerations, we formulate the subsequent assumptions:

$$\begin{cases} \sigma_{\text{gas}} > \sigma_{\text{obj}} \approx 0 & \text{(between surface and camera)} \\ \sigma_{\text{obj}} \gg \sigma_{\text{gas}} \approx 0 & \text{(on surface).} \end{cases} \tag{12}$$

Accordingly, Eqs. (10) and (11) can be reduced to:

$$\widehat{C}_i^{\text{obj}} = T_i^{\text{FIR}} \left(1 - e^{-\sigma_i^{\text{obj}} \delta_i}\right) f_i^{\text{obj}}, \tag{13}$$

$$\widehat{C}_i^{\text{gas}} = T_i^{\text{FIR}} \left(1 - e^{-\sigma_i^{\text{gas}} \delta_i}\right) f_i^{\text{gas}}. \tag{14}$$

B. Final Rendering Equation

As the FIR signal of objects and gases at the distance $t \in \mathbb{R}$ travels from the source to the camera, it is scattered and attenuated. [34] describes this phenomenon as $e^{-\beta(\lambda)\cdot t}$ where β is the attenuation coefficient and λ represents the dependency, the wavelength of the photons being measured.

Since it is difficult to model a function of β under explicit representations, we design the framework to learn β using the attenuation field as described in Sec. III-B. Assume β_i is an attenuation factor at each point $\mathbf{x}_i$, and $t_i \in \mathbb{R}$ is the distance between the point and camera, we obtain the final rendering equation, as follows:

$$\widehat{C}^{\text{FIR}} = \sum_{i=0}^{n_s - 1} e^{-\beta_i \cdot t_i} \cdot \widehat{C}_i^{\text{obj}} + \sum_{i=0}^{n_s - 1} e^{-\beta_i \cdot t_i} \cdot \widehat{C}_i^{\text{gas}}. \tag{15}$$

V. FEW SHOT METHOD

If the number of input images is small, NeRF-based approaches have been found to be insufficiently accurate [35], [36]. Therefore, we heuristically introduce the sliding multi-resolution hash encoding (SMHE) method, which is an extension of multi-resolution hash encoding [26]. We have also observed similar strategies in [36], [37].

A. Multi-resolution Hash Encoding

We first sample 3D points $\{\mathbf{x}_i\}$ along rays with corresponding direction vectors $\{\mathbf{d}_i\}$ starting from the camera center. We represent the encoding function as $\mathcal{H}(\mathbf{x}_i; \theta)$. Here, θ is trainable encoding parameters that are arranged into L levels. Each level l has a hash table that stores T_l feature vectors of dimensionality F at the vertices of a grid, the resolution of which is chosen to be a geometric progression.

Consider a single-level l, for the input point $\mathbf{x}_i$, we check which grid cell it is located in and look up eight feature vectors of the vertices of this cell in the hash table. Then, we linearly interpolate these vectors to be the feature vector of this level: $v_l \in \mathbb{R}^F$. Please refer to [26] for more details of the calculation. We repeat this process for L resolution levels and concatenate the feature vectors of each level to be the output of the encoding as the following:

$$\mathcal{H}(\mathbf{x}_i; \theta) = v_1 \oplus v_2 \oplus \ldots \oplus v_L \in \mathbb{R}^{L \cdot F}. \tag{16}$$

By mapping continuous input coordinates to discrete vertices in hash tables of different sizes, multi-resolution hash encoding stores and retrieves coarse to fine features of points in the space, thus enhancing spatial representations.

B. Sliding Multi-resolution Hash Encoding

When the number of input views is restricted, the trained fields tend to overfit to the input 2D images while not explaining geometric information in a 3D manner consistently. High-frequency encoding like hash encoding exacerbates this problem. Although high-frequency mapping enables faster convergence during training, it prevents the model from learning low-frequency features, especially in few-shot scenarios [35], [36].

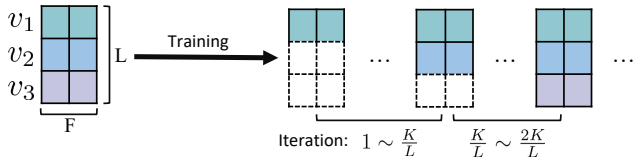

Fig. 3. Illustration of sliding multi-resolution hash encoding of 3 levels. During the training, we start with features from low levels and gradually apply features from higher levels.

Therefore, we apply a sliding multi-resolution hash encoding method to regularize the frequency mapping. Fig. 3 illustrates an example of our algorithm. Given $\mathcal{H}(\mathbf{x}_i;\theta)$ (Eq. (16)), we apply a linearly increasing mask $\mathcal{M} \in \mathbb{R}^{L\cdot F}$ to control each level of features based on the training steps:

$$\tilde{\mathcal{H}}(\mathbf{x}_i;\theta) = \mathcal{H}(\mathbf{x}_i;\theta) \odot \mathcal{M}(k,K,L,F) \tag{17}$$

$$\mathcal{M}_{(m)}(k,K,L,F) = \begin{cases} 1 & \text{if } m \in \left[1, \left\lceil \frac{k\cdot L}{K} \right\rceil \cdot F\right] \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where $\mathcal{M}_{(m)}(k,K,L,F)$ denotes the m-th element of $\mathcal{M}$, and k, K are the current and total iteration of training.

More specifically, we start with the low-frequency features at level 1 and linearly increase the frequency at each level as training progresses. It circumvents the unstable high-frequency features at the beginning and gradually applies high-frequency information during the training.

VI. EXPERIMENT

A. Experimental Setup

1) Datasets & evaluation metrics: Our experiment data comprises real-world and synthetic datasets, including RGB frames that capture a full view of the scene and FIR frames collected under few-shot settings. In the real-world datasets, every FIR frame is paired with a corresponding RGB frame, taken from the identical position and orientation to ensure consistency. RGB frames and FIR frames are of 1280×720 and 320×256 size respectively. The **Plaster** dataset takes a view of the indoor scene with no gas present to demonstrate the effectiveness of our method under gas-free conditions. For safety concerns, we use the **Glass** dataset as a substitution for the real-world dataset involving gases. In this dataset, a piece of heated glass is placed vertically between tables which is hardly seen from the RGB camera but observed by the FIR camera. For the synthetic datasets, named **Table** and **Bedroom**, we create RGB and FIR frames that simulate the presence of heated gas based on [38]. They are both of the resolution size 1920×1080.

We separate the experiment into two parts, full views and sparse views. In the full-view part, the proposed method receives complete views of RGB and FIR frames as input. For the sparse-view part, although full views of RGB frames are available, the FIR scenes adhere to the few-shot setting in [36], where the models are trained on 8 frames and evaluated on 25 frames. For our quantitative evaluations, we report PSNR, SSIM, and LPIPS scores. For the qualitative evaluation, we plot rendered FIR images and their corresponding ground truth images. Since FIR images are usually dark, we adjust the image contrast for better visualization.

2) Implementation Details: We implemented our work on Nerfstudio [39] and chose $L = 16$ and $F = 2$ for the setting of MHE and SMHE.

3) Comparing methods: We compare the proposed method with other view synthesis methods:

- Nerfacto [39]: includes state-of-the-art methods from recent NeRF's extensions.
- RawNeRF [40]: designed for dark images.
- FreeNeRF [36]: proposed frequency regularization for few-shot scenarios.
- With and without SMHE for an ablation study.

B. Evaluation of synthesis quality from full views

Table I shows the quantitative of novel view synthesis quality from full views. Our method outperforms other techniques by at most 19.4%, 2.3%, and 44.8% improvement in PSNR, SSIM, and LPIPS respectively.

The advantage of our method is most evident when comparing the reconstructed shapes to the ground truth, particularly in areas highlighted by a red rectangle in Fig. 4. In the Plaster dataset, our method reconstructs the shape of heated objects most similar to the ground truth. Although RawNeRF performs better in the Glass dataset, it captures the shape of the glass with internal noise (Fig. 4). In contrast, our method captures the shape of the glass more effectively. Because of the prior knowledge of the object information, the proposed method can concentrate on enhancing the distinct details and nuances of the scene in the FIR field.

All methods achieve high-performance metrics in the Table and Bedroom dataset. This success, however, owes much to the simplicity of FIR images, which feature a single color channel with low pixel values. Therefore, it does not necessarily reflect a correct understanding of the scene. In Fig. 5, we plot the depth map of each method. Nerfacto fails to accurately capture the shape of gases while RawNeRF struggles to reconstruct the objects in dark areas. In contrast, our method shows a smooth color transition from near to far, indicating a more accurate scene understanding.

C. Evaluation of view synthesis from sparse views

Table I shows the quantitative evaluation of view synthesis quality from sparse views. We also plot the synthesized images of the bedroom dataset in Fig. 6 for a qualitative evaluation. Our method outperforms FreeNeRF in all datasets with at most 13.8%, 3.2%, and 24.4% improvement in PSNR, SSIM, and LPIPS respectively. The advantage is also evident in Fig. 6, particularly in low-light conditions. A notable instance is the bed area within the image, where FreeNeRF faces difficulties in accurately discerning and reconstructing objects. In contrast, our method exhibits remarkable precision in reconstructing objects in dark areas. The increased accuracy stems from our approach's capability to leverage information from the RGB frames for scene reconstruction.

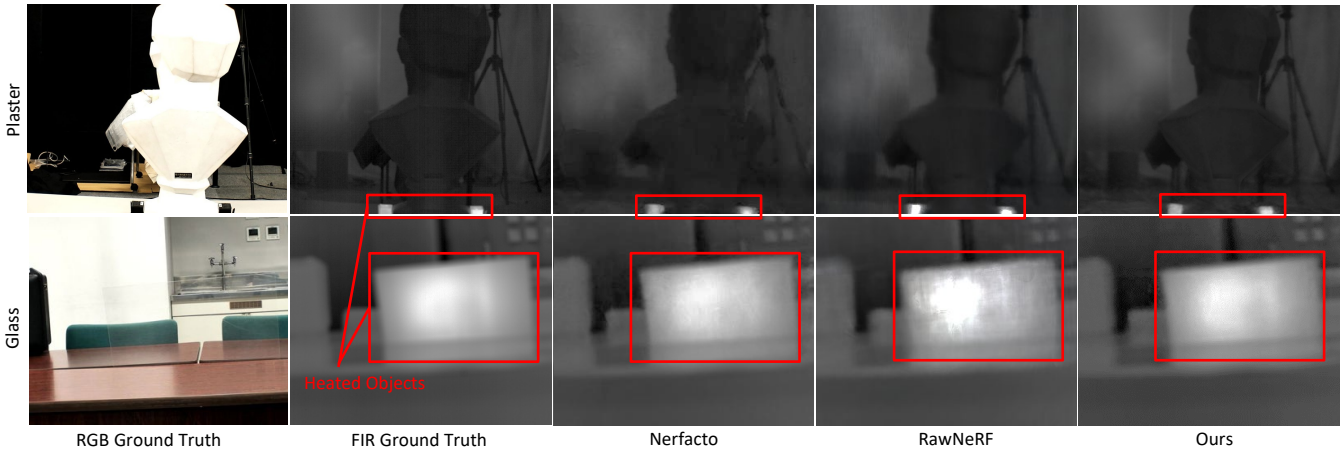

Fig. 4. Qualitative Evaluation for synthesis quality from full views of the Plaster and Glass dataset. In the plaster dataset, cube-shaped heated objects are placed on the table. In the Glass dataset, a piece of heated glass is placed vertically between tables. Compared with other methods that reconstruct heated objects with noises, ours captures and reconstructs the shape more accurately.

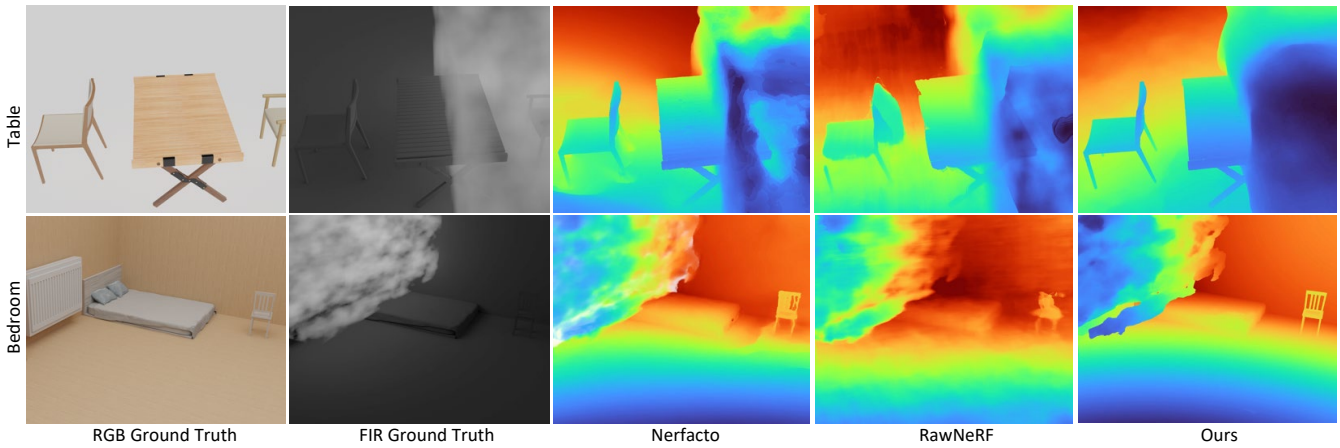

Fig. 5. Qualitative Evaluation for synthesis quality (Depth Map) from full views of the Table and Bedroom dataset. When rendering depth maps, Nerfacto demonstrates errors in capturing the shape of gases while RawNeRF struggles to learn the objects in dark areas. In contrast, the color transition of our depth map is smooth, showing the spatial arrangement of objects and gases based on distances.

Method	Plaster			Glass			Table			Bedroom		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
	Synthesis quality from full views											
Nerfacto [39]	33.55	0.933	0.181	23.51	0.830	0.334	44.03	0.988	0.026	47.94	0.991	0.027
RawNeRF [40]	33.23	0.949	0.156	**24.58**	**0.881**	**0.310**	41.99	**0.989**	0.055	42.16	0.990	0.029
Ours	**35.56**	**0.955**	**0.110**	23.55	0.835	0.321	**46.15**	**0.989**	**0.020**	**50.34**	**0.993**	**0.016**
	Synthesis quality from sparse views											
FreeNeRF [36]	21.60	0.570	0.489	22.81	0.839	0.241	25.92	0.927	0.206	28.10	0.930	0.131
Ours without $\tilde{\mathcal{H}}$	22.56	0.561	0.343	16.90	0.739	0.400	26.45	0.943	0.198	16.92	0.782	0.361
Ours with $\tilde{\mathcal{H}}$	**23.13**	**0.578**	**0.334**	**23.34**	**0.841**	**0.260**	**27.48**	**0.943**	**0.179**	**31.98**	**0.960**	**0.099**

TABLE I

QUANTITATIVE EVALUATION FOR NOVEL VIEW SYNTHESIS OF FIR SCENE

D. Visualization of FIR scene

Besides the outstanding performance of the proposed method in novel view synthesis, our method can recognize and isolate the invisible scene in the FIR field. We plot the accumulation map of σ^{gas} as $\sum_{i=0}^{N-1} T_i^{\text{FIR}} \left(1 - e^{-\sigma_i^{\text{gas}} \delta_i}\right)$ in the left column of Fig. 7. The purple color in the background indicates the regions with almost zero density while the red color represents the highest density of gas. The sharp contrast

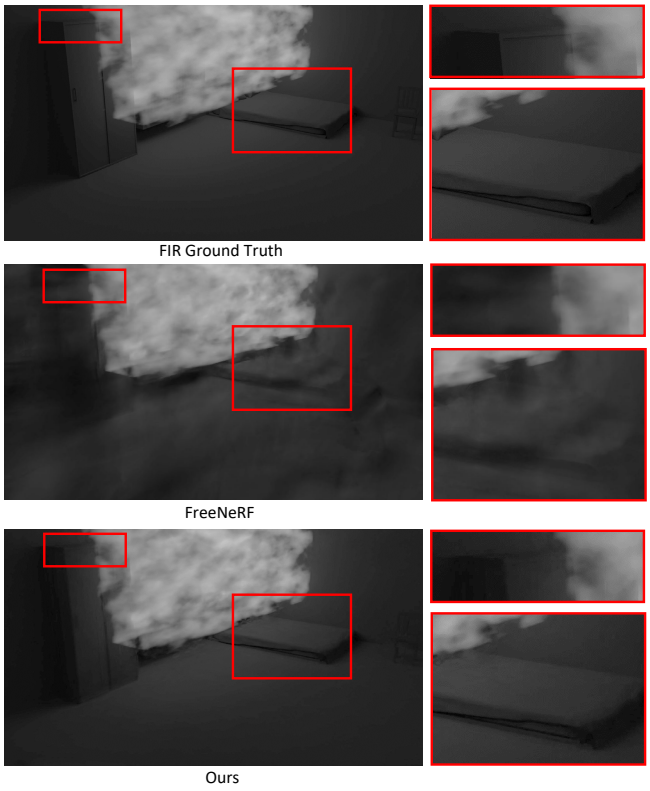

Fig. 6. Qualitative Evaluation of novel view synthesis from sparse views. FreeNeRF struggles to accurately determine the objects in dark areas. In contrast, our method can reconstruct the scene even in poorly lit conditions.

between colors shows the ability to isolate gas areas from the background, which is particularly useful in identifying gas leaks or sources of emissions.

E. Ablation Study of Sliding Multi-resolution Hash Encoding

An ablation study is carried out for the few-shot technique described in V-B. The last two rows of Table I show the result of the quantitative evaluation. In the few-shot setting, MHE tends to suffer from overfitting, which is particularly obvious in the Glass and Bedroom datasets. To further investigate this issue, we have tracked the progression of PNSR for both training and evaluation images within the Bedroom dataset, as depicted in Figure 8. Although the training metric increases rapidly with MHE, the evaluation metric struggles to demonstrate similar improvements, with the PSNR hovering around 15. On the contrary, when employing SMHE, the PSNR of the evaluation image exhibits periodic improvements, indicating a more effective adaptation to the evaluation data. As for some drop points in the evaluation metric, it is because we incorporate encoding from higher levels for training at that iteration.

VII. CONCLUSIONS

In this paper, we propose INF of RGB and FIR for invisible scenes. To our best knowledge, this work is the first attempt to reconstruct invisible scenes in the FIR field using INRs. The proposed system exhibits good reconstruction

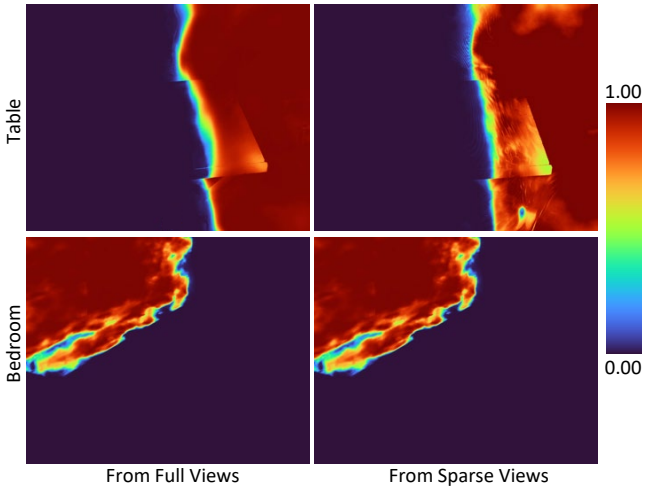

Fig. 7. Visualization of gas accumulation map. Original FIR images are shown in Fig. 5. With the purple and red representing the low and high-density values, our method can isolate the gas components for both full and sparse-view scenarios, offering a new dimension of scene understanding.

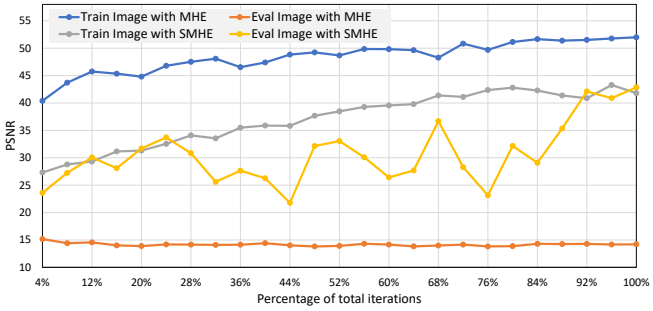

Fig. 8. PSNR with respect to the training progress. Although PSNR of the training image increases quickly with MHE, the evaluation quality struggles to improve. In contrast, SMHE demonstrates a gradual improvement in evaluation metrics throughout the training process.

quality on both real-world and synthetic datasets. Also, the proposed rendering formula is proven effective for capturing the distribution of gases. Furthermore, our method does not require complete scene views. It maintains competitive performance with just a few FIR frame shots. Thus, our method could suit various scenarios.

Future work will aim to explore our framework for dynamic scenes. We will closely investigate and try to develop rendering equations to capture the movement of gases, such as diffusion and turbulent flows. We also plan to implement an extrinsic calibration of color and FIR cameras in the same framework for robust data input. Furthermore, we plan to broaden the application of INF to other sensors, including radar, x-ray, and event cameras. Sensor fusion shall become more unified by leveraging the implicitness of the system.

ACKNOWLEDGEMENT

This work was partially supported by JST PRESTO Grant Number JPMJPR22C4, JSPS KAKENHI Grant Number JP22H00545, and NEDO project JPNP20006 in Japan.

References

[1] F. Vatansever and M. R. Hamblin, "Far infrared radiation (fir): Its biological effects and medical applications: Ferne infrarotstrahlung: Biologische effekte und medizinische anwendungen," *Photonics & lasers in medicine*, vol. 1, no. 4, pp. 255–266, 2012.

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[3] X. Wei, Y. Zhang, Z. Li, Y. Fu, and X. Xue, "Deepsfm: Structure from motion via deep bundle adjustment," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 230–247.

[4] D. Wang, X. Cui, X. Chen, Z. Zou, T. Shi, S. Salcudean, Z. J. Wang, and R. Ward, "Multi-view 3d reconstruction with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5722–5731.

[5] 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.

[6] S. Zhou, S. Xie, R. Ishikawa, K. Sakurada, M. Onishi, and T. Oishi, "Inf: Implicit neural fusion for lidar and camera," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 10 918–10 925.

[7] Q. Herau, N. Piasco, M. Bennehar, L. Roldão, D. Tsishkou, C. Migniot, P. Vasseur, and C. Demonceaux, "Moisst: Multi-modal optimization of implicit scene for spatiotemporal calibration," *arXiv preprint arXiv:2303.03056*, 2023.

[8] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5380–5389.

[9] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 407–419, 2019.

[10] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[11] Y. Chen, L. Wan, Z. Li, Q. Jing, and Z. Sun, "Neural feature search for rgb-infrared person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 587–597.

[12] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 013–15 022.

[13] H. Lu, X. Zou, and P. Zhang, "Learning progressive modality-shared transformers for effective visible-infrared person re-identification," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1835–1843.

[14] Y. Zhang, Z. Sun, B. Wu, J. Yang, and H. Kong, "Building a stereo and wide-view hybrid rgb/fir imaging system for autonomous vehicle," *IEEE Sensors Journal*, vol. 22, no. 2, pp. 1638–1651, 2021.

[15] Y. Zhang, J. Xie, J. M. Álvarez, C.-Z. Xu, J. Yang, and H. Kong, "Capitalizing on rgb-fir hybrid imaging for road detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13 819–13 834, 2021.

[16] S. Speth, A. Goncalves, B. Rigault, S. Suzuki, M. Bouazizi, Y. Matsuo, and H. Prendinger, "Deep learning with rgb and thermal images onboard a drone for monitoring operations," *Journal of Field Robotics*, vol. 39, no. 6, pp. 840–868, 2022.

[17] A. Banuls, A. Mandow, R. Vázquez-Martín, J. Morales, and A. García-Cerezo, "Object detection from thermal infrared and visible light cameras in search and rescue scenes," in *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2020, pp. 380–386.

[18] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

[19] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, "Learning representations and generative models for 3d point clouds," in *International conference on machine learning*. PMLR, 2018, pp. 40–49.

[20] Y. Liao, S. Donne, and A. Geiger, "Deep marching cubes: Learning explicit surface representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2916–2925.

[21] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 52–67.

[22] 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.

[23] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470.

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5855–5864.

[25] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Zip-nerf: Anti-aliased grid-based neural radiance fields," *ICCV*, 2023.

[26] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.

[27] Q. Herau, N. Piasco, M. Bennehar, L. Roldo, D. Tsishkou, C. Migniot, P. Vasseur, and C. Demonceaux, "Soac: Spatio-temporal overlap-aware multi-sensor calibration using neural radiance fields," *arXiv preprint arXiv:2311.15803*, 2023.

[28] A. Carlson, M. S. Ramanagopal, N. Tseng, M. Johnson-Roberson, R. Vasudevan, and K. A. Skinner, "Cloner: Camera-lidar fusion for occupancy grid-aided neural representations," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2812–2819, 2023.

[29] M. Li, J. He, Y. Wang, and H. Wang, "End-to-end rgb-d slam with multi-mlps dense neural implicit representations," *IEEE Robotics and Automation Letters*, 2023.

[30] H. Zhu, Y. Sun, C. Liu, L. Xia, J. Luo, N. Qiao, R. Nevatia, and C.-H. Kuo, "Multimodal neural radiance field," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9393–9399.

[31] 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.

[32] N. Max, "Optical models for direct volume rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 2, pp. 99–108, 1995.

[33] D. Levy, A. Peleg, N. Pearl, D. Rosenbaum, D. Akkaynak, S. Korman, and T. Treibitz, "Seathru-nerf: Neural radiance fields in scattering media," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 56–65.

[34] R. Highnam and M. Brady, "Model-based image enhancement of far infrared images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 4, pp. 410–415, 1997.

[35] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.

[36] J. Yang, M. Pavone, and Y. Wang, "Freenerf: Improving few-shot neural rendering with free frequency regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8254–8263.

[37] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu, "Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3295–3306.

[38] Institute of Space Systems, University of Stuttgart, "Blender radar," https://www.iss.uni-stuttgart.de/download/blender-radar.

[39] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, *et al.*, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–12.

[40] 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.