ThermalGaussian++: Improving Alignment and Resolution for ThermalGaussian

Rongfeng Lu, Chi Zhu, Quan Chen, Le Zhang, Ming Lu, Tingyu Wang,
Haofan Ren, Yitian Xue, Yunfei Guo and Chenggang Yan

***Abstract*—Thermography is especially valuable for the military and other users of surveillance cameras. Some recent methods based on Neural Radiance Fields (NeRF) have been proposed to reconstruct thermal scenes in 3D from a set of thermal and RGB images. However, unlike NeRF, 3D Gaussian splatting (3DGS) prevails due to its rapid training and real-time rendering. In this work, we propose ThermalGaussian, the first thermal 3DGS approach capable of rendering high-quality images in RGB and thermal modalities. We first calibrate the RGB camera and the thermal camera to ensure that both modalities are accurately aligned. Subsequently, we use the registered images to learn the multimodal 3D Gaussians. To prevent the overfitting of any single modality, we introduce several multimodal regularization constraints. We also develop smoothing constraints tailored to the physical characteristics of the thermal modality. Besides, we contribute a real-world dataset named RGBT-Scenes, captured by a handheld thermal-infrared camera, facilitating future research on thermal scene reconstruction. Based on ThermalGaussian, we further introduce ThermalGaussian++ to improve the alignment and resolution of ThermalGaussian. To improve multimodal alignment, we design a multimodal pose optimization module. This module enables direct processing of non-aligned multimodal image pairs, reducing the need for professional calibration before each use. To improve thermal resolution, we also propose a multimodal joint super-resolution reconstruction module, which enhances the quality of low-resolution thermal fields. Additionally, we contribute a new dataset: RGBT-Scenes++, which offers higher-resolution thermal images. We conduct comprehensive experiments demonstrating that ThermalGaussian++ achieves photorealistic thermal rendering and improves RGB rendering quality. It significantly enhances both alignment and resolution, enabling better practical deployment. In addition, our multimodal regularization constraints reduce the model's storage requirements. The code and datasets will be released.**

***Index Terms*—3DGS, Thermal Imaging, View Synthesis, Multimodal 3D Reconstruction, Temperature Field Reconstruction.**

I. Introduction

THERMAL imaging is widely used in fields such as military [1], healthcare [2], industry [3], agriculture [4], building inspection [5], and search and rescue [6] because it converts temperature information—an important physical modality not visible to the human eye—into interpretable images. 3D reconstruction technology, which involves lifting multi-view 2D images into 3D scenes, is foundational for key technologies such as the metaverse, digital twins, autonomous driving, and robotics. Any image with valuable 2D applications can be lifted into 3D to view the captured scene from a new view and in greater detail. Thermal images are no exception [7], [8].

Previous 3D thermal scene reconstruction [9]–[12] typically involves a two-stage process. In the first stage, RGB images and traditional multi-view geometry methods [13] are used to achieve a 3D geometric reconstruction of the scene. In the second stage, thermal images are mapped onto the reconstructed 3D scene. However, these methods not only fail to fully exploit thermal information but are also constrained by the limitations of traditional 3D reconstruction techniques, which impede their ability to render high-quality images from a new view. This significantly hinders the practical application of thermal reconstruction.

Neural Radiance Fields (NeRF) [14] can render photorealistic images from a new view, thus revolutionizing novel-view synthesis and 3D reconstruction. Recently, several NeRF-based methods [15], [16] have been proposed to reconstruct thermal scenes in 3D using thermal images. However, NeRF's slow rendering speed and implicit scene representation limit its practical applications. In contrast, 3D Gaussian Splitting (3DGS) [17] not only maintains photorealistic rendering quality from new views but also significantly improves rendering speed. Additionally, the explicit 3D scene representation of 3DGS, akin to point clouds, facilitates integration with downstream tasks, establishing it as a leading approach in research. Building on 3DGS, we propose ThermalGaussian, a multimodal Gaussian technique that renders high-quality RGB and thermal images from new views.

The absence of open-source datasets dedicated to thermal scene reconstruction significantly impedes progress in this domain. Several researchers [15], [16] have recognized this issue and have released some datasets. However, these datasets suffer from problems such as lack of paired RGB–thermal images, inconsistencies in thermal information from different views, watermarked images. To address these issues, we contribute a real-world dataset named RGBT-Scenes.

Unlike RGB images, thermal images possess unique low-texture and ghosting characteristics [18] that hinder accurate camera pose estimation using Structure-from-Motion (SfM) [19], as illustrated in Fig. 1b. Consequently, thermal images cannot directly replace RGB images for running 3DGS. To address this issue, we first register the RGB and thermal images and then fuse them (Fig. 1c), or use Multi-Spectral

Rongfeng Lu, Chi Zhu, Quan Chen, Le Zhang, Ming Lu, Tingyu Wang, Haofan Ren, Yunfei Guo and Chenggang Yan with the Hangzhou Dianzi University, China 310018 (e-mail: rongfeng-lu@hdu.edu.cn; chizhu@hdu.edu.cn; chenquan@alu.hdu.edu.cn; lezhang@hdu.edu.cn; lu199192@gmail.com; tingyu.wang@hdu.edu.cn; rhfkris@gmail.com; gyf@hdu.edu.cn; cgyan@hdu.edu.cn)

Yitian Xue with the Zhejiang University, China 310058 (e-mail: xueyt@zhejianglab.org)

Le Zhang is the Corresponding Author.

Dynamic Imaging (MSX) [20] (Fig. 1d) to localize the thermal image camera. Additionally, we design a thermal loss to adapt to the unique characteristics of thermal images.

Introducing a new modality, such as thermal imaging, into 3D reconstruction should enable the model to understand the scene from a more comprehensive perspective. However, ThermoNeRF [15] reduces the RGB rendering quality after implementing thermal reconstruction. In contrast, our method not only improves thermal rendering quality but also enhances RGB rendering quality by 1 dB. Furthermore, to prevent overfitting of any single modality during multimodal Gaussian training, we introduce a multimodal regularization coefficient. This approach significantly reduces model storage requirements and accelerates rendering speed.

To further advance the practical deployment of thermal scene reconstruction, we extend ThermalGaussian with enhanced multimodal alignment and resolution. In multimodal data acquisition, signals originate from different hardware devices and are naturally misaligned. Aligning such multimodal data is typically complex and requires professional expertise, which raises the barrier for non-experts and limits real-world applications. To address this challenge, we design a multimodal pose optimization module. This module automatically processes non-aligned inputs during reconstruction, eliminating the need for manual registration. It significantly reduces the difficulty of using ThermalGaussian, enabling non-experts to perform high-fidelity multimodal 3D reconstruction with any device setup.

High-resolution thermal cameras are prohibitively expensive, which increases the equipment cost for deployment. To overcome this issue, we propose a multimodal super-resolution reconstruction module. This module allows low-cost, low-resolution thermal cameras to achieve reconstruction results comparable to those from expensive high-resolution cameras. In addition, the lack of publicly available high-resolution thermal datasets has hindered research in this field. To fill this gap, we contribute RGBT-Scenes++, a real-world dataset captured with a high-end thermal camera that provides higher-resolution thermal images.

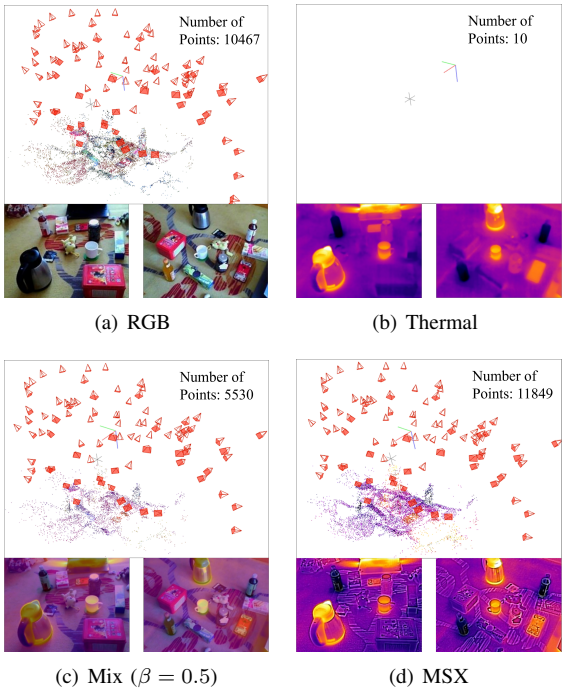
Unlike RGB-based 3D reconstruction, which mainly serves visualization and novel view synthesis, thermal 3D reconstruction also requires analysis of temperature distributions. In practice, users need to know the exact temperature values at different surface points of the reconstructed scene. To support this, we design an interactive 3D temperature measurement tool built upon our reconstruction results.

In summary, the main contributions are as follows:

(1) We propose ThermalGaussian, the first multimodal 3DGS capable of simultaneously rendering photorealistic thermal and RGB images of a scene.

(2) We propose a series of strategies for multimodal Gaussian reconstruction, including multimodal initialization, three different thermal Gaussians, constraints specific to thermal modalities, and multimodal regularization.

(3) We introduce two real-world multimodal datasets for 3D reconstruction and novel-view synthesis. RGBT-Scenes consists of paired RGB and thermal images captured from multiple viewpoints across various scenes. RGBT-Scenes++ extends this with higher-resolution thermal images acquired using more expensive devices.



(a) RGB (b) Thermal

(c) Mix ($\beta = 0.5$) (d) MSX

Fig. 1: Top: camera poses and point cloud generated by SfM. Bottom: input images for SfM.

(4) We propose ThermalGaussian++, which improves the handling of non-aligned and low-resolution thermal inputs, and integrates a real-time interactive interface for temperature measurement. These enhancements accelerate the practical deployment of 3D thermal reconstruction.

(5) Finally, experimental results show that our multimodal method not only improves the rendering quality of both thermal and RGB images but also reduces storage space by 90% compared to training each modality separately, while also improving rendering speed.

II. Related Work

A. Thermal imaging

Thermal imaging is a powerful non-contact technology that captures the surface temperatures of objects and creates intuitive thermographic visualizations. According to the theory of blackbody radiation [21], all objects with temperatures above absolute zero emit energy in the form of electromagnetic waves, a phenomenon known as thermal radiation. Planck's law indicates that as the temperature of an object increases, the wavelength of its emitted thermal radiation decreases and its radiative power density increases. Typically, in everyday settings, the electromagnetic waves emitted due to thermal radiation fall within the infrared and visible light spectrum. For example, the peak wavelength of human body temperature ($37°C$) is approximately 9.4 μm, located in the mid-infrared band. Using an infrared camera, we can capture the radiative power density of the infrared radiation emitted from an object's surface. By applying Planck's law, the surface temperature of the object can be computed. Subsequently, a colormap is used to correlate different temperatures with

distinct colors, and the surface temperatures are rendered into a pseudocolored thermal image recognizable by humans.

Thermal imaging has extensive military applications and, as the cost of thermal imaging equipment continues to decrease, its use in civilian contexts is becoming increasingly widespread, including in industrial anomaly detection [22], building energy audits [23], medical diagnostics [24], search and rescue operations [25], and laboratory testing [26]. Additionally, in these applications, converting two-dimensional temperature data into three-dimensional information can significantly enhance the utility of temperature information across various industries. However, unique characteristics of thermal images prevent them from being used for high-fidelity 3D reconstructions in the same way as color images. For instance, due to the phenomena of heat conduction and radiation, temperature does not change abruptly like color but rather gradually transitions, resulting in a distinctive ghosting effect [18] in thermal images. Furthermore, since most non-heat-emitting objects and the surrounding environment typically reach thermal equilibrium, rendering the temperature of most objects similar to the ambient temperature, thermal images often exhibit low-texture characteristics. These unique properties of thermal images pose challenges in using tools like COLMAP for location positioning and in reconstructing detailed surface features.

B. Thermal 3D reconstruction

The advent of the groundbreaking KinectFusion [13] marked the beginning of an era of high-precision, dense 3D reconstruction. Subsequent developments, such as Voxel Hashing [27], InfiniTAM [28], BundleFusion [29], and ROSEFusion [30], have optimized 3D reconstruction algorithms in terms of accuracy, efficiency, and unconstrained camera movement. These methods utilize camera intrinsics to project depth maps into 3D space, continually ascertain camera pose through point cloud registration algorithms, and update the 3D scene using TSDF-Fusion [31]. This process is repeated to achieve dense 3D reconstruction results. As thermal information is crucial in various applications, [9]–[12]efforts have been made to integrate thermal imaging with these methods, leading to the development of surface temperature field reconstruction algorithms. However, these traditional multi-view geometry-based methods do not perform as well in reconstructing details and rendering new viewpoints as the more recent deep learning-based approaches.

NeRF [14] has emerged as a significant milestone in the fields of computer graphics and 3D reconstruction due to its impressive ability to render highly realistic images from new viewpoints. To address NeRF's limitations in real-time performance and dense view requirements, several approaches have proposed improved representations. For example, SNeRG [32] introduces a Sparse Neural Radiance Grid to enable real-time rendering by baking trained NeRFs into a compact voxel-based format, while RGBDNeRF [33] leverages geometric priors from sparse RGB-D images to achieve high-quality synthesis from fewer input views. These methods improve NeRF's usability, but still rely on implicit volumetric representations that limit explicit 3D geometry extraction. In our concurrent work, ThermoNeRF [15] and Thermal-NeRF [16], like us, are exploring the integration of thermal images with deep learning-based algorithms for 3D reconstruction. Although these approaches successfully generate images from new perspectives, they are limited by the implicit scene representation of NeRF [14]. This limitation prevents them from providing a true three-dimensional representation, such as point clouds, which is crucial for seamless integration with other applications. To address these limitations and enhance the fidelity and rendering speed of 3D thermal reconstruction, we have developed a surface temperature field reconstruction algorithm based on 3D Gaussians.

C. 3DGS and multimodality

3DGS [17] represents a revolutionary technology in the fields of 3D reconstruction. Distinct from methods like NeRF, 3DGS employs millions of explicit Gaussians, fundamentally altering its approach. This technology merges the advantages of neural network-based optimization with structured data representation, enabling photorealistic rendering from new views, significantly enhancing real-time rendering capabilities, and introducing the ability to manipulate and edit 3D scenes. These features make 3DGS highly compatible with a broad range of downstream applications, establishing it as the baseline for next-generation 3D reconstruction technologies [34]. For example, GaussNav [35] applies 3DGS to embodied visual navigation tasks, constructing Gaussian-based scene representations that preserve both semantic and textural information, significantly boosting instance-level navigation performance. In addition, Gamba [36] demonstrates the feasibility of combining 3DGS with efficient architectures like Mamba for single-view 3D reconstruction at millisecond speeds, further extending 3DGS's applicability. Although 3DGS is constrained and trained using only RGB modality images, it ultimately generates millions of Gaussians, resembling a point cloud. This characteristic makes 3DGS particularly suitable for multimodal fusion with other devices that directly capture scene point clouds, such as depth cameras and LiDAR. Studies [37]–[39] have effectively integrated depth cameras to implement 3DGS-based simultaneous localization and mapping. Studies [40], [41] have effectively combined depth images [42] estimated from a pre-trained Monocular Depth estimation model [43] with the RGB modality, resulting in improved rendering quality and more accurate geometry.

In real-world scenarios, data includes not only color and geometric modalities but also other important physical information such as temperature, pressure, and magnetic fields. To build a more realistic digital twin in virtual space, these physical properties should also be reconstructed digitally. In this work, we are the first to incorporate the temperature modality into 3DGS [17] to explore multimodal 3D reconstruction. We find that adding a new modality not only enables the reconstructed digital twin to support rendering in this new modality but also allows the model to better understand the scene from a more comprehensive perspective. As a result, the rendering quality of the original RGB images is improved. For example, integrating the thermal modality significantly enhances novel-view rendering performance in low-light scenes. Moreover,

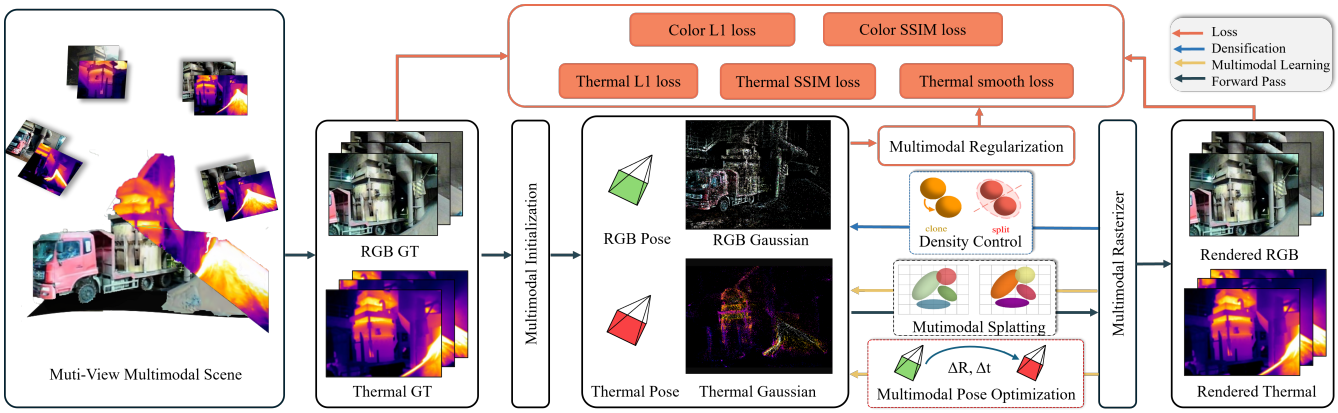


Fig. 2: **ThermalGaussian Overview.** We simultaneously construct Gaussians for RGB and thermal modalities using the point cloud obtained from multimodal initialization. Each modality's Gaussians are used to render images in their respective modality. However, the losses from different modalities are combined to jointly constrain the optimization of both sets of Gaussians. To address the difficulty of aligning multi-modal inputs under large viewpoint differences, we use the color camera poses estimated by COLMAP and optimize the thermal camera poses through a learning-based approach. Additionally, we establish a multimodal regularization based on the number of Gaussians in each modality, which dynamically adjusts the training coefficients for both modalities.

our proposed multimodal regularization greatly reduces the model's storage requirements.

III. ThermalGaussian

Fig. 2 shows the overview of the proposed ThermalGaussian, which is based on the 3DGS [17], aiming to extend its capability to simultaneously render images of color and temperature. In this section, we first briefly introduce the background of the 3DGS. Then, we provide a detailed description of our method's specific implementation details, including multimodal initialization, three types of multimodal thermal Gaussians, thermal loss, and multimodal regularization.

A. Preliminary: 3D Gaussian Splatting

3DGS [17] represents a 3D reconstruction scene using a large number of anisotropic 3D Gaussians. This representation not only provides differentiability, which offers advantages in learning-based methods, but also enables explicit spatial expression, enhancing the editability and controllability of 3D scenes. Furthermore, it allows for rapid and efficient rasterization rendering through splatting. Initially, a set of unordered images of objects to be reconstructed is processed using SfM to obtain the camera poses and sparse point clouds. 3DGS then initializes these sparse point clouds as the position μ of a 3D Gaussian:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{1}$$

where Σ represents the covariance matrix of the 3D Gaussian, and x denotes any point in the 3D scene. Σ is defined using a scaling matrix $\boldsymbol{S}$ and a rotation matrix $\boldsymbol{R}$:

$$\Sigma = \boldsymbol{R}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{R}^T \tag{2}$$

The 3D Gaussian $G(x)$ is projected onto the imaging plane using the camera's intrinsic parameters, transforming it into a 2D Gaussian. Subsequently, the image is rendered through alpha-blending:

$$C\left(x'\right) = \sum_{k \in N} c_k \alpha_k \prod_{j=1}^{k-1} \left(1 - \alpha_j\right) \tag{3}$$

where x' represents the queried pixel position, N denotes the number of 2D Gaussians corresponding to this pixel, α denotes the opacity of each Gaussian and the color c on each Gaussian is modeled spherical harmonics. All attributes of the 3D Gaussians are learnable and optimized directly in an end-to-end manner during training.

B. Multimodal Initialization

Previously, methods for calibration RGB and thermal images [44] often involve designing specialized, non-standard metallic calibration boards with uniformly sized circular holes. The calibration relies on the temperature difference between the board and the background to compute thermal features, enabling calibration. However, the high complexity and stringent requirements for producing these calibration boards make them difficult to obtain and lack a universal standard. We find that a standard chessboard pattern, as shown in Fig 3, commonly used for RGB camera calibration, can effectively be used for calibrating both thermal and color cameras, with a mean reprojection error of less than 0.5 pixels. Initially, we heat the calibration board using devices like an infrared heater; black regions, absorbing heat faster due to their material

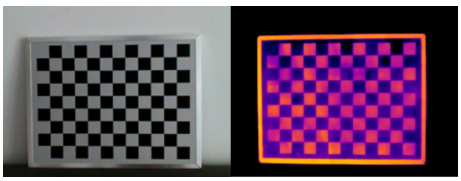


Fig. 3: Standard checkerboard for thermal camera calibration

properties, exhibit relatively higher temperatures. We capture color and thermal images simultaneously before thermal equilibrium, which occurs when two systems reach a balanced state with equal temperatures, halting heat flow. Subsequently, conventional camera calibration [45] is performed.

Using the calibrated intrinsic parameters $\boldsymbol{K}_{\text{RGB}}$ for the color camera, $\boldsymbol{K}_{\text{Th}}$ for the thermal camera, and the rotation $\boldsymbol{R}$ and translation $\boldsymbol{t}$ from the temperature camera to the color camera, we computed the corresponding positions $(u_{\text{Th}}, v_{\text{Th}})$ on the thermal image mapped to the registered positions on the color image:

$$\begin{bmatrix} u_{\text{RGB}} \\ v_{\text{RGB}} \\ 1 \end{bmatrix} = \boldsymbol{K}_{\text{RGB}} \left(\boldsymbol{R} \cdot \boldsymbol{K}_{\text{Th}}^{-1} \begin{bmatrix} u_{\text{Th}} \\ v_{\text{Th}} \end{bmatrix} + \boldsymbol{t} \right) \tag{4}$$

As shown in Fig.1(b), directly using thermal images, which exhibit low texture and ghosting characteristics, makes it difficult to successfully run SfM [46]. Therefore, to obtain the thermal camera poses, we test three different multimodal SfM strategies. The first utilizes registered high-texture RGB images directly for camera pose estimation. These poses serve simultaneously for both the RGB and thermal cameras. However, practical scenarios that require thermal scene reconstruction often occur under dim lighting conditions or in scenes lacking distinct color features. Therefore, relying solely on color images may impede the precise camera pose estimation necessary for thermal scene reconstruction. The second approach, illustrated in Fig.1(c), involves blending registered color and thermal images using the following formula:

$$I_{\text{mix}} = \beta I_{\text{Th}} + (1 - \beta) I_{\text{RGB}} \tag{5}$$

where in the above equation, β represents the opacity of the thermal image. This method produces blended images containing both rich color and thermal information, catering to various practical applications of thermal scene reconstruction. The third strategy, depicted in Fig.1(d), maps high-frequency color variations from the color images onto the thermal images. This approach mitigates the lack of feature points caused by thermal images' low texture and ghosting characteristics.

C. Thermal Gaussian

We utilize three different multimodal training strategies to construct the thermal Gaussian.

Multimodal Fine-Tuning Gaussians (MFTG): Inspired by the fine-tuning approach used in large-scale models, our first multimodal training strategy is training a basic Gaussian with RGB images and then refining this Gaussian with thermal images to generate thermal Gaussian. This is a two-stage process. In the first stage, similar to 3DGS, we utilize multimodal camera poses and initial point clouds obtained from multimodal initialization as inputs. The training is supervised using RGB images, with $\mathcal{L}_1$ combined with a D-SSIM term:

$$\mathcal{L}_{\text{RGB}} = (1 - \lambda)\mathcal{L}_1 + \lambda \mathcal{L}_{\text{D-SSIM}} \tag{6}$$

This stage enables us to render high-quality RGB modality images from a new view and establish a basic 3D Gaussian with preliminary geometry. In the second stage, we fine-tune this basic Gaussian model with thermal images and multimodal camera poses obtained from initialization. Since the first stage constraints are based on texture-rich color images rather than thermal images, which results in a better geometry. Therefore, training on this geometry yields better results than training thermal Gaussian directly from the initial point cloud derived from multimodal initialization.

Multiple Single-Modal Gaussians (MSMG): The training of MFTG initially utilizes RGB modal information followed by thermal modal information. Although both modalities are employed, they are not used simultaneously. Since only thermal images are utilized for supervision in stage two, the information from the color modality was not fully leveraged. Therefore, in MSMG (as shown in Fig. 2), we constrain the training with information from both color and thermal modalities simultaneously. We train two single-modal Gaussians initialized by point clouds from multimodal initialization. The thermal Gaussian renders thermal images, while the RGB Gaussian renders RGB images. Subsequently, these rendered images of both modalities are compared separately with the ground truth of their respective inputs using loss functions:

$$\mathcal{L} = \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{thermal}} \tag{7}$$

The details of $\mathcal{L}_{thermal}$ constraint will be elaborated below. Each Gaussian model is influenced not only by its corresponding input modality but also by others. Experimental results indicate that joint constraints across multiple modalities enhance the training outcomes for both color and thermal modalities. Moreover, since these modalities jointly optimize the entire scene from different perspectives, redundant points are pruned to some extent, reducing the number of points in the point cloud and lowering the model's storage requirements.

One Multi-Modal Gaussian (OMMG): OMMG extends MSMG by not only employing dual-modal loss constraints in Eq. (7) but also integrating multiple modalities onto a single Gaussian. This integration ensures that information from diverse modalities is unified within a single geometric structure. Specifically, we construct a multimodal Gaussian comprising positional coordinates x, scaling matrix $\mathbf{S}$, rotation matrix $\mathbf{R}$, opacity α, spherical harmonics c for RGB representation, and spherical harmonics t for thermal representation. RGB rendering is achieved using Eq. 3, while thermal rendering follows the equation below:

$$T\left(x'\right) = \sum_{k \in N} t_k \alpha_k \prod_{j=1}^{k-1} \left(1 - \alpha_j\right) \tag{8}$$

D. Thermal Loss

The loss function for RGB modality images is directly given by Eq. 6. The same loss function can also be applied to thermal modality images. However, because thermal images exhibit unique low-texture and ghosting characteristics, we design a specific thermal loss function to better accommodate these features.

The RGB modality may exhibit abrupt changes. However, because all objects above absolute zero continuously engage in heat transfer and thermal radiation, eventually reaching

thermal equilibrium with their surroundings, significant abrupt changes are typically not observed in thermal images. Additionally, most regions of objects in thermal equilibrium have similar temperatures, resulting in smoother thermal images. Therefore, we introduce a smoothness term for regularization:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{4M} \sum_{i,j} \left(|T_{i\pm1,j} - T_{i,j}| + |T_{i,j\pm1} - T_{i,j}| \right) \quad (9)$$

where $T_{i,j}$ represents rendered thermal values at pixel position (i, j). M denotes the number of rendering pixels. Similarly to the color modality, we also incorporate $\mathcal{L}_1$ and $\mathcal{L}_{\text{D-SSIM}}$ Thus, our final temperature loss is:

$$\mathcal{L}_{\text{thermal}} = (1 - \lambda)\mathcal{L}_1 + \lambda \mathcal{L}_{\text{SSIM}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}} \quad (10)$$

where λ_{smooth} is the coefficient of $\mathcal{L}_{\text{smooth}}$.

E. Multimodal Regularization

When training multiple modalities of the same object simultaneously, it is often undesirable for one modality to dominate at the expense of others. Therefore, a regularization strategy is needed to dynamically adjust the weight of each modality's loss during training.

In the training of MSMG, we observe that the weight of a modality align linearly with the Gaussian it ultimately generates. A higher weight for a modality results in more Gaussian generated by that modality, while fewer Gaussian are generated by another modality. Hence, we design multimodal regularization coefficient γ_{MSMG} based on the number of Gaussian generated by each modality during training.

$$\gamma_{MSMG} = \frac{N_{thermal}}{N_{thermal} + N_{RGB}} \quad (11)$$

where $N_{thermal}$ represents the number of Gaussian for the thermal modality during training.

When the number of one modality's Gaussian increases, we increase the training weight of the other modality. This dynamic balancing of weights ultimately prevents overfitting to any single modality.

The final design of this loss is:

$$\mathcal{L} = \gamma_{MSMG} \mathcal{L}_{\text{RGB}} + (1 - \gamma_{MSMG}) \mathcal{L}_{\text{thermal}} \quad (12)$$

When the number of Thermal Gaussians increases, γ_{MSMG} increases accordingly, raising the RGB loss weight and suppressing further growth of the Thermal branch; meanwhile, the resulting increase in RGB Gaussians causes γ_{MSMG} to decrease, which in turn suppresses the RGB branch. This bidirectional mutual suppression mechanism prevents either modality from expanding massively, significantly reducing the total number of Gaussians and leading to memory savings, while reconstruction quality remains nearly unchanged.

In OMMG, all modalities share the same geometry. That is, different modalities operate on the same set of Gaussians. Therefore, unlike Eq. (11), we cannot regulate multimodal training by adjusting the number of Gaussians for each modality. Instead, during OMMG training, we introduce a new multimodal regularization coefficient based on the training loss of each modality. We observe that when reconstructing a scene using 3DGS, the loss of the thermal images and the RGB images usually converges to a ratio of approximately 1:2. This is because RGB images contain more detailed textures, resulting in a higher loss. However, each modality is equally important for 3D reconstruction. To avoid the model overfitting to one modality due to imbalanced weights, we use the loss values of both modalities to construct a regularization term, denoted as γ_{OMMG}, to balance their contributions.

$$\gamma_{OMMG} = \frac{\mathcal{L}_{\text{RGB}}}{\mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{thermal}}} \quad (13)$$

Therefore, the final loss is defined as:

$$\mathcal{L} = \gamma_{OMMG} \mathcal{L}_{\text{RGB}} + (1 - \gamma_{OMMG}) \mathcal{L}_{\text{thermal}} \quad (14)$$

With our designed multimodal regularization, the model dynamically adjusts the weights of different modalities during training. A higher loss for a modality indicates that it is currently less sufficiently optimized and should therefore receive a larger training weight, preventing any single modality from dominating the training process for an extended period. As a result, our method significantly reduces the number of Gaussians while maintaining rendering quality.

IV. THERMALGAUSSIAN++

A. Multi-Modal Camera Pose Optimization

To reduce the usage barrier of the multimodal 3D reconstruction method ThermalGaussian, we enhance its capability to directly process unaligned image inputs from various devices. This eliminates the need for complex, expert-driven multi-camera calibration and alignment before each data collection session.

High-fidelity 3D reconstruction is highly sensitive to camera pose accuracy. RGB images contain rich textures and can obtain reliable camera poses using COLMAP. In contrast, thermal images often lack sufficient texture, which makes direct pose estimation with COLMAP unreliable. Simply reusing RGB camera poses for thermal images therefore frequently introduces ghosting artifacts during reconstruction. To enable high-fidelity thermal scene reconstruction, we introduce a multimodal pose optimization module into ThermalGaussian. We use the RGB camera, which is rigidly attached to the thermal camera, to provide an initial pose for the thermal modality. During reconstruction, we introduce a learnable multimodal relative pose parameter, which is jointly optimized to estimate the correct thermal camera pose for each viewpoint.

To enable multimodal 3D reconstruction, we design the cross-modal relative pose as a learnable parameter, allowing the system to automatically adjust the spatial relationship between different sensor modalities. Each cross-modal relative pose is represented as a transformation from the reference camera coordinate system to the target camera coordinate system: $\mathbf{T}_{\text{rel}} = [\mathbf{R} \mid \mathbf{t}] \in \text{SE}(3)$, where $\mathbf{R} \in \text{SO}(3)$ denotes the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ denotes the translation vector. Since the translation $\mathbf{t}$ lies in Euclidean space, it can be directly optimized as a set of learnable parameters. In contrast, rotations are constrained on the $\text{SO}(3)$ manifold. Therefore,

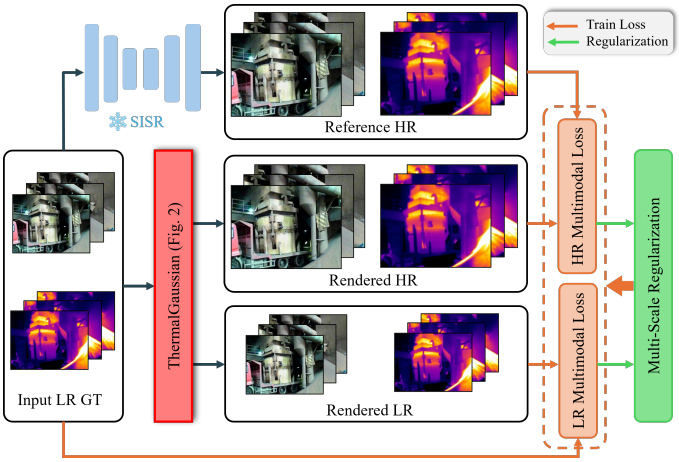


Fig. 4: We extend ThermalGaussian to reconstruct high-quality 3D multi-modal scenes using only low-resolution inputs.

we adopt the axis–angle representation $\boldsymbol{\phi} = \alpha\boldsymbol{\omega} \in \mathbb{R}^3$, where $\boldsymbol{\omega}$ is the unit rotation axis and α is the rotation angle. The corresponding rotation matrix is recovered using Rodrigues' formula: $\mathbf{R} = \mathbf{I} + \frac{\sin\alpha}{\alpha}\boldsymbol{\phi}^\wedge + \frac{1-\cos\alpha}{\alpha^2}(\boldsymbol{\phi}^\wedge)^2$, where $(\cdot)^\wedge$ denotes the skew-symmetric operator.

In our method, RGB camera poses are used as initial estimates for thermal camera poses. Rather than directly optimizing thermal camera poses independently, we learn the relative pose variation between the RGB and thermal modalities by modeling the global evolution of the reconstructed 3D Gaussian point cloud. Specifically, when the initial thermal pose carries an offset, the projected positions of the Gaussian points in the thermal image plane will be incorrect, producing a large rendering loss. This loss in turn drives the global point cloud to move toward the position where its projection aligns with the thermal ground truth, thereby implicitly compensating for the cross-modal pose error without ever directly manipulating extrinsic matrices. Our cross-modal pose optimization is directly related to 3DGS point cloud learning, and therefore inherits its well-established robustness and stability. 3DGS drives Gaussian optimization through pixel-level $\mathcal{L}_1$ and $\mathcal{L}_{\text{SSIM}}$ losses, which provide dense and stable gradient signals even in low-texture thermal scenes, ensuring stable pose optimization.

While NeRF−− [47] shows that single-modality camera poses and scene geometry can be jointly optimized during training, it often fails when camera rotations are large (e.g., beyond ±20°) or in full 360° capture scenarios, making it unsuitable for heterogeneous multimodal settings. A similar issue is addressed in Z-Splat [48], which mitigates missing cone along the view axis by fusing sonar and RGB data. In contrast, our method derives relative camera pose variations through point-cloud-level optimization, yielding a more stable alignment across diverse acquisition settings and heterogeneous sensor configurations. This implies that even when the initial cross-modal alignment contains a large offset, the system can still converge through dense pixel-level loss signals, without requiring any additional stabilization mechanism.

TABLE I: Comparison of our collected datasets with others.

Dataset	Mode T / RGB	Bimodal Calibration	Multiview Consistency	Thermal Resolution	Content Richness
Thermal-NeRF [16]	✓ / ×	-	×	382×288	Simple
ThermalNeRF [50]	✓ / ✓	×	×	160×120	Moderate
ThermoNeRF [15]	✓ / ✓	✓	×	80×60	Moderate
Thermal3DGS [51]	✓ / ×	-	×	640×512	Moderate
Veta-GS [52]	✓ / ×	-	×	640×512	Moderate
Ours-conf [53]	✓ / ✓	✓	✓	240×180	Rich
Ours	✓ / ✓	×	✓	640×480	Rich

B. Multimodal Super-Resolution Reconstruction

High-resolution thermal imaging devices are expensive. Moreover, ThermalGaussian currently performs well only when the rendering resolution matches the input resolution. Its performance deteriorates on higher-resolution outputs. To address this limitation and enable low-cost, low-resolution thermal cameras to reconstruct high-quality thermal fields and render high-resolution thermal maps, we propose a high-resolution multi-modal Gaussian approach in the ThermalGaussian++ version, as illustrated in Fig. 4. This method takes low-resolution images as input but produces reconstructions comparable to those from high-resolution inputs.

We first apply single-image super-resolution (SISR) models [49] to both the low-resolution RGB and thermal images, generating high-resolution reference images. However, since these super-resolution images are not captured from real scenes, they inevitably introduce noise. Using them directly as ground truth (GT) for training may degrade the quality of 3D reconstruction. To mitigate the impact of such noise, we incorporate the original low-resolution inputs and introduce multi-modal constraints during the super-resolution training process. This strategy enhances the reconstruction quality across all modalities, including both thermal and RGB. Specifically, for each single modality, we compute the loss using the super-resolution image and the rendered high-resolution image from the 3D model. Additionally, we compute the loss between the low-resolution ground truth and the rendered low-resolution image. For the thermal modality, the loss function for high-resolution reconstruction is defined as follows:

$$\mathcal{L}_{\text{thermal}} = \gamma_{scale}^{thermal}\mathcal{L}_{\text{thermal}}^{high} + (1-\gamma_{scale}^{thermal})\mathcal{L}_{\text{thermal}}^{low} \tag{15}$$

where $\mathcal{L}_{\text{thermal}}^{high}$ and $\mathcal{L}_{\text{thermal}}^{low}$ represent the loss functions for thermal data at high and low resolutions, respectively. The detailed formulation of the loss functions at different resolutions can be found in Eq. (10). The coefficient $\gamma_{thermal}$ denotes the multi-scale regularization weight used in the multi-modal supervision framework. It balances the influence of different scales during training and is computed as follows:

$$\gamma_{scale}^{thermal} = \frac{\mathcal{L}_{\text{thermal}}^{high}}{\mathcal{L}_{\text{thermal}}^{high} + \mathcal{L}_{\text{thermal}}^{low}} \tag{16}$$

Similarly, for the RGB modality, we construct a multi-scale loss in the same manner. The loss functions for both RGB and thermal modalities are then jointly incorporated into Eq. (12) to obtain the final loss for multi-modal super-resolution reconstruction.

V. Self-Collected Thermal Dataset

We introduce two new datasets. The first, named RGBT-Scenes, consists of aligned collections of thermal and RGB images captured from various viewpoints of a scene. The images are collected using the commercial-grade handheld thermal-infrared camera FLIR E6 PRO [54], which can simultaneously capture RGB and thermal images. The basic specifications of this camera include a resolution of 240×180, a field of view of 33°×25°, a temperature range from -20°C to 550°C, and a temperature accuracy of ±2% of the reading. Our dataset includes over 1,000 RGB and thermal images from 10 different scenes. These scenes encompass both indoor and outdoor environments, various object sizes (from large structures to everyday items), different temperature variations (ranging from a 300°C difference to a 4°C difference), and include both 360-degree and forward-facing scenarios.

In practice, using custom-built multimodal devices often requires labor-intensive alignment before each use, which significantly limits the real-world applicability of related algorithms. To address this issue, we not only propose a corresponding reconstruction method but also introduce a high-resolution, non-aligned multimodal 3D reconstruction dataset, aiming to accelerate the deployment of such systems in practical scenarios. This dataset, named RGBT-Scenes++, is collected using the FLIR A700, which captures clearer and higher-resolution thermal images. Unlike existing datasets that generate 640×480 thermal images through internal software upscaling, RGBT-Scenes++ provides true 640×480 thermal images directly from the sensor. The dataset covers a wide range of scenarios, including nighttime scenes, human subjects, indoor and outdoor environments, solar panel inspections, glass that blocks heat but allows visible light, and dark bags that block visible light but transmit thermal radiation.

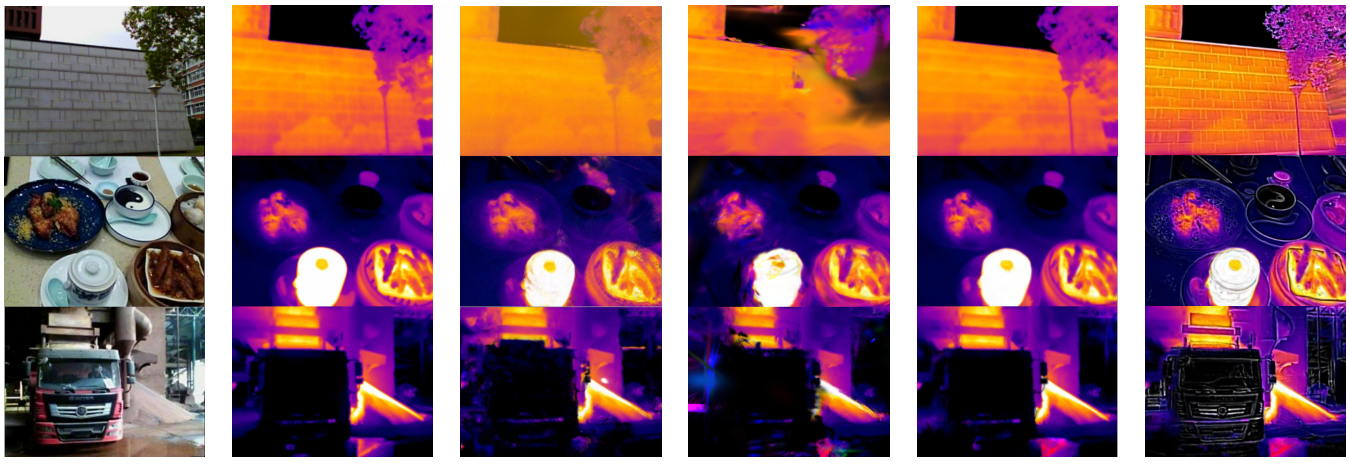
We provide the raw images captured by the thermal camera, as well as the RGB images, thermal images, MSX images, and camera pose data. In Table I, we compare our datasets with those of concurrent work, Thermal-NeRF [16], ThermalNeRF [50] and ThermoNeRF [15]. Our dataset includes both RGB and thermal images. For each scene, the temperature range used for rendering thermal images is fixed to ensure consistent color representation of the same 3D point across different viewpoints. Compared to other datasets, our datasets provide more accurate and view-consistent thermal measurements, higher-resolution thermal details, and a broader variety of scene types. Detailed descriptions of each scene are provided in the supplementary.

VI. Experiments

A. Implementation Details

Our method is an improvement upon the 3DGS framework, with all experimental settings (e.g., λ) remaining consistent with the reference 3DGS. The specific hyperparameter λ_{smooth} is set to 0.6. Each comparative experiment was trained for 30K iterations. For the relative pose learning module, we use a separate Adam optimizer with an initial learning rate of 0.001. The learning rate is decayed according to a milestone-based schedule, where decay is applied every 300 iterations up to 30K iterations, with a decay factor of 0.9. All experiments are conducted on a single NVIDIA 3090 GPU. The resolution of the rendered RGB images and thermal images is 640×480.

For aligned multimodal inputs, the camera poses of test views can be estimated by including all training and test images in COLMAP. In contrast, for unaligned multimodal input settings, test-view poses require additional processing to be accurately determined. Similar to NeRF– [47], after training, we freeze both the scene representation and the poses of the training views. Then, we input the test-view images and optimize only their poses, without updating the scene model. This optimization uses the same loss function as in training, but focuses solely on aligning the test views with the reconstructed scene. Finally, we evaluate the quality of novel view synthesis by comparing the rendered images—generated using the reconstructed scene and the optimized test-view poses—with the ground-truth test images.



(a) RGB GT (b) Thermal GT (c) ThermoNerf (d) 3DGS (e) Ours$_{OMMG}$ (f) Ours (MSX)

Fig. 5: Qualitative comparisons of rendered thermal images from novel viewpoints using our method, previous approaches [15], [17], and the ground truth. Training results on MSX images are also shown, which are easier to apply in practice.

TABLE II: Quantitative evaluation of thermal image using our method compared to previous work from test views. "×" indicates a failure to localize using only thermal images in the scene, making it impossible to succeed with 3DGS. 3DGS+MI represents the results obtained by directly training 3DGS after Multimodal Initialization.

Metric	Method	Dimsum	Daily Stuff	Ebike	Road Block	Truck	Rotary Kiln	Building	Iron Ingot	Parterre	Land Scape	Avg.
PSNR ↑	3DGS	25.38	×	×	×	20.97	23.79	23.75	×	×	×	×
	ThermoNeRF	24.27	17.34	19.70	17.17	23.53	26.40	23.31	22.97	17.88	18.79	21.13
	3DGS+MI	26.35	18.77	20.89	26.75	26.17	26.59	25.76	29.57	22.09	20.17	24.31
	Ours$_{MFTG}$	**26.94**	20.52	22.51	24.96	25.02	26.91	26.11	**30.41**	23.55	20.03	24.70
	Ours$_{MSMG}$	26.73	21.35	23.23	26.52	**26.27**	**27.15**	**26.83**	30.06	25.01	20.61	25.38
	Ours$_{OMMG}$	26.46	**22.28**	**23.31**	**27.17**	25.88	26.33	26.72	29.86	**26.16**	**22.27**	**25.64**
SSIM ↑	3DGS	0.860	×	×	×	0.717	0.872	0.810	×	×	×	×
	ThermoNeRF	0.747	0.759	0.694	0.781	0.750	0.916	0.804	0.717	0.709	0.774	0.765
	3DGS+MI	0.889	0.789	0.806	0.917	0.872	0.922	0.872	0.887	0.843	0.794	0.859
	Ours$_{MFTG}$	**0.890**	0.798	0.845	0.906	**0.880**	0.920	0.886	0.895	0.859	0.808	0.869
	Ours$_{MSMG}$	0.891	0.829	0.857	0.909	0.879	**0.926**	**0.897**	**0.898**	0.860	0.832	0.878
	Ours$_{OMMG}$	0.886	**0.835**	**0.862**	**0.919**	0.874	0.922	0.888	0.896	**0.883**	**0.850**	**0.882**
LPIPS ↓	3DGS	0.157	×	×	×	0.281	0.193	0.299	×	×	×	×
	ThermoNeRF	0.312	0.494	0.290	0.293	0.291	0.170	0.234	0.152	0.309	0.264	0.280
	3DGS+MI	0.124	0.274	0.313	0.204	0.139	0.125	0.211	0.093	0.252	0.328	0.206
	Ours$_{MFTG}$	**0.121**	0.258	0.235	0.210	**0.133**	0.129	0.199	0.091	0.232	0.317	0.192
	Ours$_{MSMG}$	0.124	**0.208**	0.220	0.213	**0.133**	0.130	0.189	**0.086**	0.227	0.293	0.182
	Ours$_{OMMG}$	0.129	0.210	**0.203**	**0.198**	0.136	**0.124**	**0.177**	0.091	**0.181**	**0.248**	**0.170**

TABLE III: Quantitative evaluation of RGB image using our method compared to 3DGS and ThermoNeRF.

Metric	Method	Dimsum	Daily Stuff	Ebike	Road Block	Truck	Rotary Kiln	Building	Iron Ingot	Parterre	Land Scape	Avg.
PSNR ↑	3DGS	23.91	20.43	26.77	27.80	22.30	20.79	20.95	23.96	24.91	20.20	23.20
	ThermoNeRF	19.74	16.79	17.75	18.32	18.77	18.89	17.12	15.07	23.13	19.13	18.46
	Ours$_{MSMG}$	**24.42**	21.71	**27.34**	**28.22**	23.57	22.23	23.08	**25.69**	**25.57**	20.91	24.27
	Ours$_{OMMG}$	24.38	**21.76**	26.85	28.12	**24.17**	**23.14**	**24.19**	24.55	25.48	**21.71**	**24.34**
SSIM ↑	3DGS	0.847	0.748	0.901	0.910	0.807	0.772	0.791	0.872	0.859	0.696	0.820
	ThermoNeRF	0.688	0.639	0.540	0.619	0.688	0.600	0.460	0.293	0.756	0.583	0.586
	Ours$_{MSMG}$	**0.858**	0.793	**0.917**	0.916	0.833	0.811	0.844	**0.891**	**0.874**	0.715	0.845
	Ours$_{OMMG}$	**0.858**	**0.797**	0.905	**0.920**	**0.840**	**0.822**	**0.849**	0.884	0.855	**0.739**	**0.847**
LPIPS ↓	3DGS	**0.194**	0.299	0.171	0.201	0.232	0.217	0.228	0.188	0.183	0.280	0.219
	ThermoNeRF	0.228	0.465	0.244	0.548	0.311	0.207	0.291	0.301	**0.167**	0.275	0.303
	Ours$_{MSMG}$	**0.194**	0.262	**0.156**	0.221	0.217	0.190	**0.168**	**0.172**	0.184	0.275	0.204
	Ours$_{OMMG}$	**0.194**	**0.253**	0.169	**0.199**	**0.211**	**0.184**	0.170	0.186	0.195	**0.268**	**0.203**

B. Aligned Multi-Modal View Synthesis

1) Thermal View Synthesis: Similar to 3DGS, we employ image quality assessment metrics including Peak Signal-to-Noise Ratio (PSNR) [55], Structural Similarity Metric (SSIM) [56], and Learned Perceptual Image Patch Similarity (LPIPS) [57] to evaluate the quality of reconstructed thermal and RGB images from new views.

As shown in Table II, even in scenes with pronounced thermal variations, specifically targeting low-texture thermal characteristics, direct application of thermal data proves challenging for 3DGS. In very few successful cases, inadequate precision in thermal camera positioning has compromised the quality of thermal reconstructions. 3DGS+MI denotes training the original 3DGS using thermal images instead of RGB images after obtaining accurate thermal poses through our multimodal initialization. Compared to 3DGS, 3DGS+MI adapts to a wider range of scenarios and achieves higher reconstruction quality. Given the higher reconstruction quality of 3DGS [17] compared to NerfStudio [58], 3DGS+MI and our method naturally outperforms ThermoNeRF. Our three thermal Gaussian methods outperform 3DGS+MI across all scenes in PSNR, SSIM, and LPIPS. Among them, ours$_{OMMG}$ shows an average PSNR improvement of 1.3 dB. As shown in Fig. 5, our method's qualitative rendering of thermal images is clearly superior. Additionally, as depicted in Fig. 5f, we enhance thermal image readability by training with MSX images using thermal Gaussian. This hierarchical and easily recognizable thermal Gaussian further promotes the application of thermal scene reconstruction.

2) RGB View Synthesis: Our method not only achieves high-quality thermal image rendering but also significantly enhances RGB image rendering quality. As shown quantitatively in Table III, our multimodal constraints improve RGB rendering quality in nearly all scenarios, with an average PSNR improvement of 1.1 dB compared to the original 3DGS, demonstrating the effectiveness of cross-modal supervision in boosting overall reconstruction fidelity. This improvement is particularly evident in scenarios where the RGB modality struggles to identify the environment, while the thermal modality can recognize it clearly. As shown in the top of Fig.6,

where distinguishing between foreground and background is challenging in the RGB modality but straightforward in the thermal modality due to temperature differences, constraints from the thermal modality aid in the accurate learning of the RGB modality. Additionally, as depicted in the bottom of Fig.6, the assistance from thermal images enables accurate color rendering in low-light scenes for the RGB modality. Our results demonstrate that, under multimodal constraints, when one modality fails, our approach leverages accurate information from the other modality to enhance the model's understanding of the scene, thus facilitating the correct learning of the failing modality. This enables our method to advance 3D reconstruction in low-light scenes and enhances the robustness of 3D reconstruction techniques to some extent.

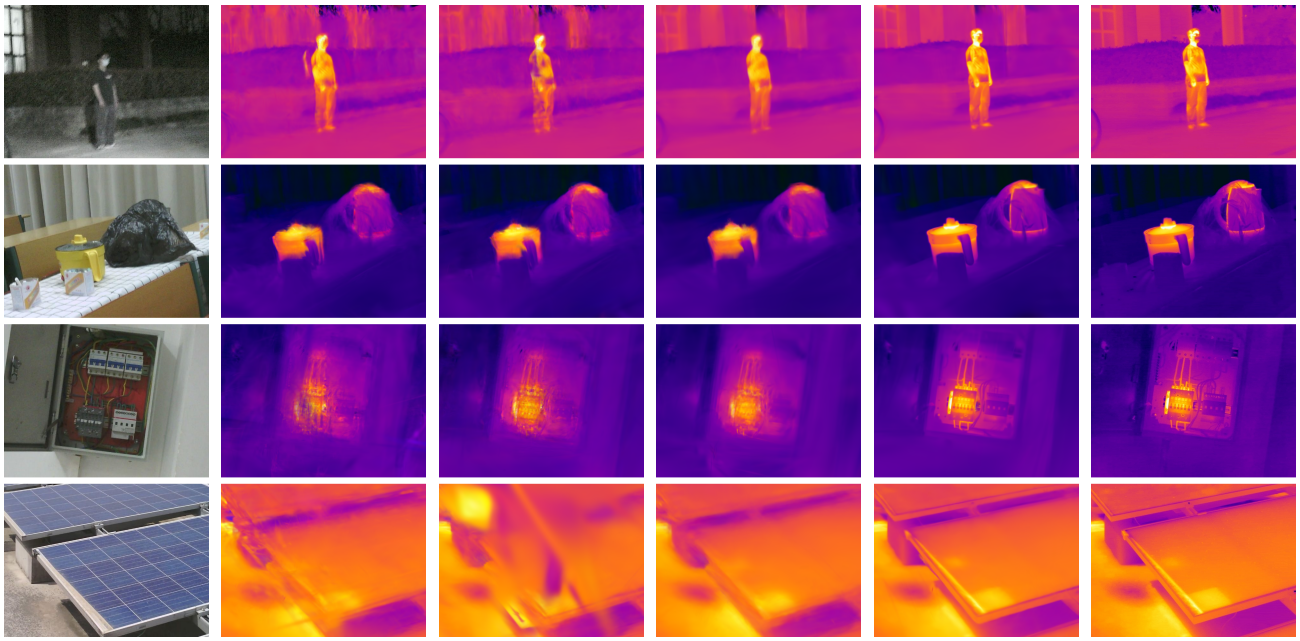
C. Unaligned Multi-Modal View Synthesis

Due to the weak texture in thermal images, it is difficult to estimate the pose of thermal cameras using structure-from-motion tools like COLMAP [46]. When RGB and thermal images are well aligned, directly using the RGB camera pose as a substitute for the thermal camera pose is effective [53]. However, in non-aligned settings, this substitution often leads to blurry and ghosting artifacts in the reconstructed results, as shown in Fig. 7 (b), (c), and (d). To address this, we introduce a multimodal pose optimization strategy that jointly refines the thermal camera pose during thermal field reconstruction. As shown in Fig. 7 (e), our method clearly outperforms the others, especially in regions with fine structural details and textures.

In the quantitative experiments, we compare our method

(a) 3DGS (b) Ours$_{OMMG}$ (c) RGB GT (d) Thermal GT

Fig. 6: We present qualitative RGB image comparisons between our method and 3DGS.



(a) RGB GT (b) 3DGS (c) Thermal3D-GS (d) Ours-conf (e) Ours (f) Thermal GT

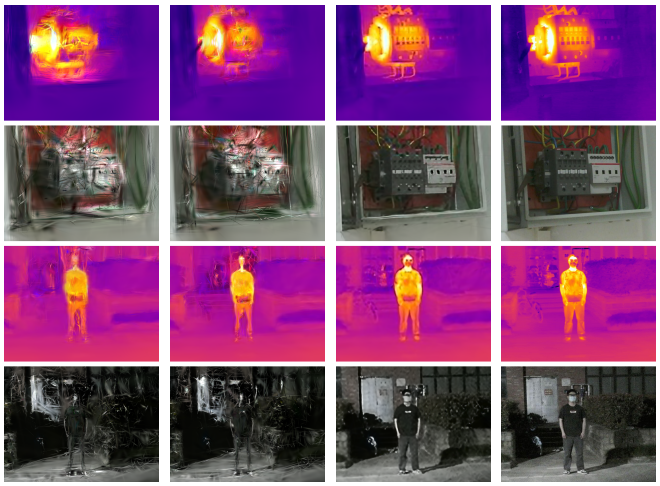
Fig. 7: Qualitative comparison of rendered thermal images from novel viewpoints under non-aligned RGB–thermal inputs.

TABLE IV: Quantitative comparison of rendered thermal image from novel viewpoints under non-aligned RGB–thermal inputs.

Metric	Method	Appliances	Human	Refreshments	Switch	Plastic	Glass	Chair	Laptop	PV Panel1	PV Panel2	Avg.
PSNR ↑	3DGS [17]	27.60	21.77	25.31	22.98	24.18	22.46	23.70	23.39	22.49	19.60	23.34
	ThermoNeRF [15]	24.38	23.61	24.60	20.94	23.49	17.59	18.73	23.68	21.73	20.38	21.91
	Therml3D-GS [51]	28.41	22.65	25.88	23.10	24.35	22.98	23.96	25.89	20.69	19.68	23.76
	INGP [59]	22.74	19.06	19.90	16.59	17.48	16.58	19.61	17.52	19.33	17.70	18.65
	Veta-GS [52]	31.12	23.42	27.09	25.69	25.66	24.41	24.19	27.31	23.09	20.53	25.25
	Ours-conf [53]	28.30	23.81	25.98	23.38	24.87	22.41	24.30	25.73	23.45	21.22	24.34
	Ours	**36.23**	**30.68**	**31.94**	**27.03**	**32.04**	**28.51**	**27.11**	**27.72**	**27.14**	**25.46**	**29.39**
SSIM ↑	3DGS [17]	0.932	0.825	0.941	0.843	0.874	0.856	0.849	0.851	0.809	0.782	0.856
	ThermoNeRF [15]	0.932	0.836	0.912	0.862	0.858	0.750	0.776	0.854	0.822	0.821	0.842
	Therml3D-GS [51]	0.944	0.843	0.944	0.843	0.876	0.863	0.853	0.882	0.813	0.784	0.864
	INGP [59]	0.796	0.798	0.868	0.681	0.725	0.646	0.758	0.675	0.769	0.702	0.742
	Veta-GS [52]	0.957	0.851	0.948	0.876	0.888	0.872	0.859	**0.890**	0.849	0.792	0.879
	Ours-conf [53]	0.947	0.869	0.946	0.860	0.888	0.858	0.862	0.887	0.853	0.813	0.878
	Ours	**0.966**	**0.920**	**0.967**	**0.914**	**0.932**	**0.914**	**0.887**	**0.890**	**0.898**	**0.882**	**0.917**
LPIPS ↓	3DGS [17]	0.068	0.320	0.151	0.290	0.210	0.157	0.329	0.158	0.337	0.389	0.241
	ThermoNeRF [15]	0.164	0.346	0.322	0.440	0.372	0.391	0.431	0.314	0.371	0.429	0.358
	Therml3D-GS [51]	0.064	0.307	0.149	0.288	0.208	0.147	0.328	0.141	0.334	0.379	0.234
	INGP [59]	0.174	0.292	0.219	0.315	0.301	0.358	0.354	0.312	0.371	0.434	0.313
	Veta-GS [52]	0.037	0.252	0.130	0.232	0.147	0.112	**0.282**	**0.087**	0.280	0.358	0.192
	Ours-conf [53]	0.064	0.302	0.152	0.291	0.203	0.158	0.343	0.139	0.294	0.355	0.230
	Ours	**0.034**	**0.227**	**0.098**	**0.218**	**0.129**	**0.096**	0.320	0.119	**0.244**	**0.284**	**0.176**

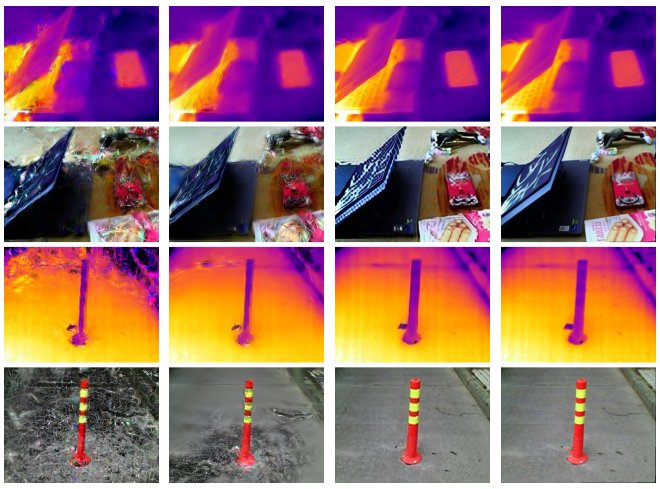
TABLE V: Quantitative comparison of rendered RGB and thermal image under non-aligned RGB–thermal inputs.

Methods	Thermal			RGB		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS [17]	23.34	0.856	0.241	25.56	0.842	0.216
ThermoNeRf [15]	21.91	0.842	0.358	16.99	0.561	0.632
Ours-conf [53]	24.34	0.878	0.230	26.17	0.854	0.242
Ours	**29.39**	**0.917**	**0.176**	**26.20**	**0.854**	**0.244**



(a) Ours-conf (b) Ours w/o SR (c) Ours (d) GT

Fig. 8: Qualitative comparison on RGBT-Scenes++ with ×8 downsampled, non-aligned RGB–thermal inputs.



(a) 3DGS (b) Ours-conf (c) Ours (d) GT

Fig. 9: Qualitative comparison on RGBT-Scenes with ×8 downsampled, aligned RGB–thermal inputs.

TABLE VI: Quantitative comparison on RGBT-Scenes with ×8 downsampled, aligned RGB–thermal inputs.

Methods	Thermal			RGB		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS [17]	16.26	0.539	0.512	12.65	0.252	0.606
SRGS [60]	22.28	0.800	0.275	18.44	0.572	0.442
Ours-conf [53]	20.07	0.742	0.347	14.42	0.409	0.529
Ours	**23.24**	**0.811**	**0.256**	**20.02**	**0.643**	**0.399**

with 3DGS [17], ThermoNeRF [15], Thermal3D-GS [51], INGP [59] , Veta-GS [52] and Our-conf [53]. All these methods, including ours, use camera poses estimated from texture-rich RGB images via COLMAP [46]. During training, our method refines the poses of the thermal modality using a multimodal pose optimization module. As a result, our method consistently outperforms previous approaches across all test scenes. On average, it achieves a 5dB improvement in PSNR over the previous state-of-the-art thermal 3D reconstruction methods, as shown in Table IV.

Moreover, our method not only improves the thermal modality, where camera poses are optimized, but also enhances the rendering quality of the RGB modality, even though the RGB camera poses remain fixed, as shown in Table V.

TABLE VII: Quantitative comparison on RGBT-Scenes++ with ×4 downsampled, non-aligned RGB–thermal inputs.

Methods	Thermal			RGB		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours-conf [53]	24.08	0.865	0.237	19.64	0.624	0.378
Ours w/o SR	27.73	0.896	0.197	19.72	0.627	0.376
Ours	**28.53**	**0.900**	**0.184**	**23.43**	**0.760**	**0.276**

This result further demonstrates that, during multimodal joint reconstruction, improvements in one modality can enhance the model's overall understanding of the scene, thereby facilitating improvements in the other modality as well.

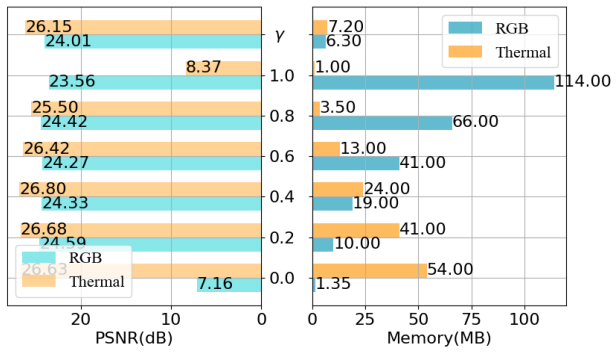
D. High-Resolution View Synthesis

High-resolution thermal cameras are extremely expensive, which limits the accessibility of thermal 3D reconstruction. To lower the entry barrier, we extend ThermalGaussian with a multimodal super-resolution module, allowing it to reconstruct high-quality thermal fields from low-resolution inputs that closely match the results from high-resolution inputs. We evaluate this SR extension on both of our datasets using ×4 and ×8 downsampled thermal images as input. As shown in Fig. 8 and Fig. 9, our SR-enhanced model still achieves high-quality reconstruction under low-resolution inputs. In both the RGB and thermal modalities, the results are significantly improved compared to those produced by the baseline without the multimodal SR module.

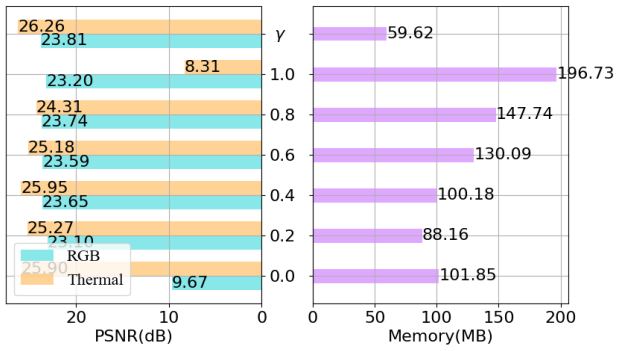
Table VI compares our method on RGBT-Scenes against 3DGS, our conference version (Ours-conf), and SRGS [60], a recent method for low-resolution reconstruction. In this experiment, we feed only single-modality inputs to each method. Results show that our multimodal SR module outperforms both the non-SR methods (3DGS and Ours-conf) and the single-modality SR approach (SRGS). This improvement indicates that joint constraints from multiple modalities help the model better understand and reconstruct 3D geometry. Tables VII and VIII further compare the performance under different input resolutions on RGBT-Scenes++, which is a more realistic and challenging dataset with non-aligned multimodal inputs. "Ours w/o SR" refers to our method equipped with non-aligned reconstruction but without the SR module. The quantitative results demonstrate that our multimodal SR module consistently improves performance at all resolutions, suggesting that our multimodal super-resolution module achieves strong and reliable reconstruction quality. The improvements are particularly notable at lower input resolutions. As shown in Table VIII, the thermal modality achieves a 2 dB increase in PSNR, while the RGB modality improves by nearly 10 dB.

TABLE VIII: Quantitative comparison on RGBT-Scenes++ with ×8 downsampled, non-aligned RGB–thermal inputs.

Methods	Thermal			RGB		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours-conf [53]	21.92	0.826	0.271	14.91	0.436	0.508
Ours w/o SR	24.85	0.854	0.235	14.97	0.439	0.506
Ours	**26.80**	**0.878**	**0.211**	**21.40**	**0.660**	**0.418**



(a) Ours$_{MSMG}$



(b) Ours$_{OMMG}$

Fig. 10: Comparison Between Dynamic Multi-Modal Regularization term and Fixed coefficient.

E. Ablation Study

We separate different contributions and algorithm choices to test their effectiveness. As shown in Table II and Table IX, after incorporating multimodal initialization, allows 3DGS to achieve thermal reconstruction across various environments. Our multimodal thermal Gaussian models, MSMG and OMMG, not only render both thermal and RGB images simultaneously but also improve rendering quality for both modalities in all scenes, with an average increase of over 1.2 dB. We also observed that multimodal constraints mitigate the generation of excessive redundant Gaussians. Later, we introduced a regularization term to dynamically adjust the coefficients of both modalities. As shown in Table IX, directly training RGB modality Gaussians with 3DGS results in an average storage requirement of 159 MB. On the other hand, directly training thermal Gaussians with MI requires an average of 65 MB. The RGB Gaussians for MSMG+MR average only 9 MB in storage, with thermal Gaussians averaging the same. Our method requires only 8% ($\frac{9+9}{159+65} = 0.08$) of the storage space compared to directly using 3DGS. Due to the reduction in the number of Gaussians, the rendering speed has also significantly increased. Inspired by the multimodal regularization used in MSMG, we designed a new multimodal regularization for OMMG. Although its effect is less pronounced than that on MSMG, it still reduces storage by 48% while maintaining rendering quality. Additionally, the rendering quality for both modalities has also improved. MFTG, MSMG+MR, and OMMG excel in different aspects: training speed, storage efficiency, and rendering quality. In Fig.10, we compare our multimodal regularization γ_{MSMG}, γ_{OMMG}

TABLE IX: **Ablation Study.** We conducted ablation experiments by gradually adding each component to the baseline 3DGS model. We then performed a comprehensive comparison across various dimensions, including rendering capability, the quality of rendered color and thermal images, training time, model memory usage, and the number of Gaussians. "-" indicates that the model lacks the specified capability or metric, and "×" denotes a reconstruction failure. Since multimodal regularization relies on the Gaussians from multiple modalities, it only applies to Ours$_{MSMG}$.

Methods	Mode	Thermal			RGB			Train	FPS	Mem.
	T / RGB	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS			
3DGS	✓ / -	×	×	×	-	-	-	×	×	×
	- / ✓	-	-	-	23.20	0.820	0.219	507s	231	159MB
+MI	✓ / -	24.31	0.859	0.206	-	-	-	367s	277	65MB
+MI +$\mathcal{L}_{\text{smooth}}$	✓ / -	24.65	0.867	0.198	-	-	-	603s	292	61MB
Ours$_{MFTG}$	✓ / -	24.70	0.871	0.191	-	-	-	**491s**	316	51MB
Ours$_{MSMG}$	✓ / ✓	25.38	**0.883**	0.180	24.27	**0.845**	0.204	873s	330 / 298	18MB+66MB
Ours$_{OMMG}$	✓ / ✓	**25.64**	**0.883**	**0.169**	**24.34**	0.800	**0.203**	838s	271 / 242	136MB
Ours$_{MSMG}$+MR	✓ / ✓	25.09	0.880	0.189	24.21	0.840	0.235	760s	**390 / 420**	**9MB+9MB**
Ours$_{OMMG}$+MR	✓ / ✓	25.42	0.882	0.174	23.62	0.824	0.244	768s	490 / 588	70MB

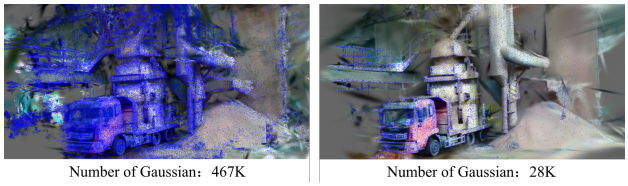


Fig. 11: Qualitative Comparison of Multi-Modal Regularization Effectiveness via Gaussian Distributions. Left: 3DGS; Right: Ours$_{MSMG}$+MR

with manually adjusting the thermal constraint coefficients in the truck scene. The comparison shows that our multimodal regularization approach reduces storage space for both RGB and thermal modalities while maintaining high image quality. In Fig.11, we visually present the Gaussian distributions of the original 3DGS method and our method with multimodal regularization. The results clearly show that our regularization significantly reduces redundant Gaussians in both modalities, leading to substantial memory savings for both RGB and thermal branches while preserving reconstruction quality.

We also extend ThermalGaussian to handle reconstruction from non-aligned RGB and thermal inputs. As shown in Table IV and Fig. 7, we compare Ours-conf with our enhanced version on the RGBT-Scenes++ dataset. Quantitatively, our method achieves a 5 dB improvement in PSNR over the baseline without multimodal pose optimization. Qualitatively, the rendered novel-view images show significantly improved texture details, demonstrating the effectiveness of our pose optimization under non-aligned real-world inputs. To reduce hardware costs, we further extend our framework to support low-resolution multimodal thermal field reconstruction. In Fig. 8b, we show the novel-view rendering results under non-aligned, low-resolution inputs with only pose optimization. In Fig. 8c, we add the multimodal SR module on top of pose optimization. The visual comparison clearly shows the substantial improvement brought by the SR module. Tables VII and VIII provide corresponding quantitative comparisons, further confirming the effectiveness of our method under low-resolution and non-aligned conditions.

VII. Applications

In practical temperature analysis, we aim not only to visualize the 3D temperature distribution of a reconstructed object but also to obtain the specific temperature value at any location on the object. To this end, we designed an interactive temperature measurement interface based on a KNN retrieval model. This system allows users to view the overall 3D temperature field and interactively query the temperature value at any given point within it. Conventional thermal cameras first measure the temperature, then convert it into color values using a colormap and the upper and lower temperature limits. Our approach is the reverse: we start with color information from each point and infer the corresponding temperature using the scene's temperature range. Since different colormaps render temperatures into different colors, we build separate retrieval datasets for each colormap. For a specific colormap, we first capture multi-view images of a scene with a clear temperature gradient. The temperature range of the scene remains fixed. The captured thermal images should ideally contain colors corresponding to both the upper and lower limits, as well as as many intermediate temperature colors as possible. Using FLIR software, we extract the mapping between color and temperature for each pixel in these 2D images. To estimate the temperature at any point in the 3D field, we first retrieve its color and then map it to a temperature using the pre-established color-temperature relationship. A straightforward method is to directly look up the temperature in a color-temperature table. However, this approach is slow and requires re-generating the table every time the temperature range changes, making it impractical. Therefore, we propose a fast retrieval method that can compute the corresponding temperature for any color under arbitrary temperature limits and scenes. We first convert the temperature value into a normalized percentage based on the color-temperature table using the following equation:

$$\rho = \frac{x - \theta_{low}}{\theta_{high} - \theta_{low}} \tag{17}$$

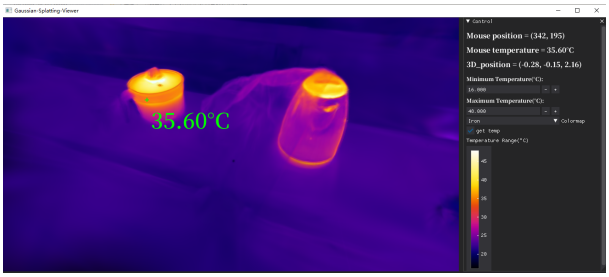


Fig. 12: A real-time interactive interface for temperature measurement on the reconstructed 3D thermal model.

where x denote the specific temperature value, θ_{high} and θ_{low} represent the upper and lower temperature limits of the scene, respectively. We first convert the temperature x into a normalized percentage ρ within this range. Then, we map each ρ value to its corresponding color and use K-Nearest Neighbors (KNN) to generate a Serialized Machine Learning Model Object. During actual use, the user clicks a point on the 3D model through a graphical user interface. The system retrieves the color value of that point. This color is then passed to the KNN-based Serialized Model, which returns the corresponding percentage ρ. Finally, we compute the temperature at that point using the following equation:

$$\hat{x} = \theta_{low} + \hat{\rho}(\theta_{high} - \theta_{low}) \tag{18}$$

VIII. Conclusion

To advance multimodal 3D reconstruction and thermal 3D reconstruction, we introduce two new datasets: RGBT-Scenes, which provides aligned RGB–thermal image pairs, and RGBT-Scenes++, which offers higher-resolution, non-aligned RGB and thermal images that more accurately reflect real-world conditions. We are the first to implement thermal reconstruction based on 3DGS. Our method not only enables simultaneous rendering of thermal and RGB images, but also significantly improves the rendering quality in both modalities while substantially reducing the model's storage requirements. To support practical deployment, our method directly processes non-aligned multimodal inputs—raw data typically captured in real-world applications—and achieves high-fidelity reconstruction. Furthermore, it supports low-resolution thermal inputs and produces results comparable to those obtained from high-resolution sensors, effectively lowering hardware costs. Lastly, we provide a interactive tool for accurate temperature measurement on the reconstructed 3D thermal models.

References

[1] 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.

[2] B. B. Lahiri, S. Bagavathiappan, T. Jayakumar, and J. Philip, "Medical applications of infrared thermography: a review," *Infrared physics & technology*, vol. 55, no. 4, pp. 221–235, 2012.

[3] A. Glowacz, "Fault diagnosis of electric impact drills using thermal imaging," *Measurement*, vol. 171, p. 108815, 2021.

[4] Z. Zhou, Y. Majeed, G. D. Naranjo, and E. M. Gambacorta, "Assessment for crop water stress with infrared thermal imagery in precision agriculture: A review and future prospects for deep learning applications," *Computers and Electronics in Agriculture*, vol. 182, p. 106019, 2021.

[5] Y. El Masri and T. Rakha, "A scoping review of non-destructive testing (ndt) techniques in building performance diagnostic inspections," *Construction and Building Materials*, vol. 265, p. 120542, 2020.

[6] S. Yeom, "Thermal image tracking for search and rescue missions with a drone," *Drones*, vol. 8, no. 2, p. 53, 2024.

[7] M. Abreu de Souza, D. C. Alka Cordeiro, J. d. Oliveira, M. F. A. d. Oliveira, and B. L. Bonafini, "3d multi-modality medical imaging: Combining anatomical and infrared thermal images for 3d reconstruction," *Sensors*, vol. 23, no. 3, p. 1610, 2023.

[8] Y. Liu, R. Wu, S. Yan, X. Cheng, J. Zhu, Y. Liu, and M. Zhang, "Atloc: Aerial thermal images localization via view synthesis," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[9] J. Rangel, S. Soldan, and A. Kroll, "3d thermal imaging: Fusion of thermography and depth cameras," in *International Conference on Quantitative InfraRed Thermography*, vol. 3, 2014.

[10] S. Zhao, Z. Fang, and S. Wen, "A real-time handheld 3d temperature field reconstruction system," in *2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE, 2017, pp. 289–294.

[11] A. R. O. Müller, *Close range 3D thermography: real-time reconstruction of high fidelity 3D thermograms*. kassel university press GmbH, 2019, vol. 7.

[12] H. Li, S. Wang, Z. Bai, H. Wang, S. Li, and S. Wen, "Research on 3d reconstruction of binocular vision based on thermal infrared," *Sensors*, vol. 23, no. 17, p. 7372, 2023.

[13] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 2011, pp. 127–136.

[14] 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.

[15] M. Hassan, F. Forest, O. Fink, and M. Mielle, "Thermonerf: Multimodal neural radiance fields for thermal novel view synthesis," *arXiv preprint arXiv:2403.12154*, 2024.

[16] T. Ye, Q. Wu, J. Deng, G. Liu, L. Liu, S. Xia, L. Pang, W. Yu, and L. Pei, "Thermal-nerf: Neural radiance fields from an infrared camera," *arXiv preprint arXiv:2403.10340*, 2024.

[17] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[18] 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.

[19] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] S. Abdullah, "Low-energy radar for handheld ir-cameras: Using mmwave radar as a complement for multi spectral dynamic imaging in handheld cameras," 2023.

[21] M. J. Riedl, *Optical design fundamentals for infrared systems*. SPIE press, 2001, vol. 48.

[22] I. Campione, F. Lucchi, N. Santopuoli, and L. Seccia, "3d thermal imaging system with decoupled acquisition for industrial and cultural heritage applications," *Applied Sciences*, vol. 10, no. 3, p. 828, 2020.

[23] A. Schischmanow, D. Dahlke, D. Baumbach, I. Ernst, and M. Linkiewicz, "Seamless navigation, 3d reconstruction, thermographic and semantic mapping for building inspection," *Sensors*, vol. 22, no. 13, p. 4745, 2022.

[24] R. F. van Doremalen, J. J. van Netten, J. G. van Baal, M. M. Vollenbroek-Hutten, and F. van der Heijden, "Infrared 3d thermography for inflammation detection in diabetic foot disease: A proof of concept," *Journal of diabetes science and technology*, vol. 14, no. 1, pp. 46–54, 2020.

[25] C. Cruz Ulloa, G. Prieto Sánchez, A. Barrientos, and J. Del Cerro, "Autonomous thermal vision robotic system for victims recognition in search and rescue missions," *Sensors*, vol. 21, no. 21, p. 7346, 2021.

[26] G. Cardone, A. Ianiro, G. Dello Ioio, and A. Passaro, "Temperature maps measurements on 3d surfaces with infrared thermography," *Experiments in fluids*, vol. 52, pp. 375–385, 2012.

[27] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 6, pp. 1–11, 2013.

[28] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. H. S. Torr, and D. W. Murray, "Very High Frame Rate Volumetric Integration of Depth Images on Mobile Device," *IEEE Transactions on Visualization and*

Computer Graphics (Proceedings International Symposium on Mixed and Augmented Reality 2015, vol. 22, no. 11, 2015.
[29] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, p. 1, 2017.
[30] J. Zhang, C. Zhu, L. Zheng, and K. Xu, "Rosefusion: random optimization for online dense reconstruction under fast camera motion," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–17, 2021.
[31] 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.
[32] P. Hedman, P. P. Srinivasan, B. Mildenhall, J. T. Barron, and P. Debevec, "Baking neural radiance fields for real-time view synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5875–5884.
[33] Y.-J. Yuan, Y.-K. Lai, Y.-H. Huang, L. Kobbelt, and L. Gao, "Neural radiance fields from sparse rgb-d images for high-quality view synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8713–8728, 2022.
[34] G. Chen and W. Wang, "A survey on 3d gaussian splatting," *arXiv preprint arXiv:2401.03890*, 2024.
[35] X. Lei, M. Wang, W. Zhou, and H. Li, "Gaussnav: Gaussian splatting for visual navigation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
[36] Q. Shen, Z. Wu, X. Yi, P. Zhou, H. Zhang, S. Yan, and X. Wang, "Gamba: Marry gaussian splatting with mamba for single-view 3d reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
[37] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian splatting slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 039–18 048.
[38] C. Yan, D. Qu, D. Xu, B. Zhao, Z. Wang, D. Wang, and X. Li, "Gs-slam: Dense visual slam with 3d gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 595–19 604.
[39] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, "Splatam: Splat track & map 3d gaussians for dense rgb-d slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 357–21 366.
[40] J. Li, J. Zhang, X. Bai, J. Zheng, X. Ning, J. Zhou, and L. Gu, "Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 775–20 785.
[41] J. Chung, J. Oh, and K. M. Lee, "Depth-regularized optimization for 3d gaussian splatting in few-shot images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 811–820.
[42] C. Yan, Z. Li, Y. Zhang, Y. Liu, X. Ji, and Y. Zhang, "Depth image denoising using nuclear norm and learning graph model," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 4, pp. 1–17, 2020.
[43] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.
[44] J. Zhang, Y. Liu, M. Wen, Y. Yue, H. Zhang, and D. Wang, "L 2 v 2 t 2 calib: Automatic and unified extrinsic calibration toolbox for different 3d lidar, visual camera and thermal camera," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–7.
[45] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proceedings of the seventh ieee international conference on computer vision*, vol. 1. Ieee, 1999, pp. 666–673.
[46] 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.
[47] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu, "Nerf–: Neural radiance fields without known camera parameters," 2021.
[48] Z. Qu, O. Vengurlekar, M. Qadri, K. Zhang, M. Kaess, C. Metzler, S. Jayasuriya, and A. Pediredla, "Z-splat: Z-axis gaussian splatting for camera-sonar fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
[49] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.
[50] Y. Y. Lin, X.-Y. Pan, S. Fridovich-Keil, and G. Wetzstein, "Thermalnerf: Thermal radiance fields," in *2024 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2024, pp. 1–12.
[51] Q. Chen, S. Shu, and X. Bai, "Thermal3d-gs: Physics-induced 3d gaussians for thermal infrared novel-view synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 253–269.
[52] M. Nam, W. Park, M. Kim, H. Hur, and S. Lee, "Veta-gs: View-dependent deformable 3d gaussian splatting for thermal infrared novel-view synthesis," in *2025 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2025, pp. 965–970.
[53] R. Lu, H. Chen, Z. Zhu, Y. Qin, M. Lu, C. Yan *et al.*, "Thermalgaussian: Thermal 3d gaussian splatting," in *The Thirteenth International Conference on Learning Representations*.
[54] Teledyne FLIR. (2024) FLIR E6 Pro Thermal Imaging Camera. [Online]. Available: https://www.flir.com/products/e6-pro/?vertical=condition%20monitoring&segment=solutions
[55] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.
[56] 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.
[57] 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.
[58] 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.
[59] 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.
[60] X. Feng, Y. He, Y. Wang, Y. Yang, W. Li, Y. Chen, Z. Kuang, J. Fan, Y. Jun *et al.*, "Srgs: Super-resolution 3d gaussian splatting," *arXiv preprint arXiv:2404.10318*, 2024.

Rongfeng Lu received the BS degree in electrical engineering and automation from Shenyang Jianzhu University, Shenyang, China, in 2020. He is currently working toward the Ph.D. in the Department of Automation, Hangzhou Dianzi University, Hangzhou, China. His research interests include intelligent information processing, visual relocalization, 3D reconstruction, and multimodal information fusion.

Chi Zhu received the B.S. degree from Nanjing University of Information Science and Technology, Nanjing, China, in 2020. He is currently a second-year graduate student pursuing a Master's degree in Control Engineering at Hangzhou Dianzi University, Hangzhou, China. His research interests include 3D reconstruction and multi-modal pose estimation.

Quan Chen received the B.S. degree from the Hangzhou Dianzi University, Zhejiang, China, in 2020. He is currently pursuing the Ph.D. degree with Hangzhou Dianzi University, Zhejiang. His research interests include image retrieval, image super-resolution, and object detection. He regularly review for major computer vision conferences (CVPR, ICCV, ECCV, AAAI, and ACMMM) and related journals (IEEE TIP/TCSVT/TGRS/TIM).

Le Zhang received the Ph.D. degree from Hangzhou Dianzi University, Hangzhou, China,in 2023. She is a Senior Laboratory Scientist in intelligent detection and machine learning. She spent one year as a Visiting Scholar at McMaster University,Hamilton, ON, Canada. She is a Key Member in the construction and operation of several labs,including Zhejiang Provincial Key Laboratory of IoT Perception and Information Fusion, the Provincial Special Finance Laboratory of Advanced Control and Intelligent Information Integration.

Ming Lu received a PhD in Information and Communication Engineering from Tsinghua University, Beijing, China, in 2019. He is currently a staff researcher at Intel Labs China. His research interests include 3D vision and computer graphics. He is especially interested in AI + Chips, Neural Field, and Large AI Models.

Tingyu Wang was an assistant professor at the School of Information and Communication Engineering, Hangzhou Dianzi University, Hangzhou, China. He received his Ph.D. degree from the Lab of Intelligent Information Processing, Hangzhou Dianzi University. His research interests include deep learning, image retrieval, and image super-resolution.

Haofan Ren received the master's degree in automation from Hangzhou Dianzi University, Hangzhou, China, in 2025. He is currently working at XGRIDS as 3D reconstruction algorithm engineer. His research insterests include intelligent information processing, machine learning, image processing, and computational biology.

Yitian Xue received his B.S.M. degree from Tulane University in 2014 and his M.S.I.T. degree in Information Systems from Carnegie Mellon University in 2020. He is a researcher at Zhejiang Lab and is pursuing a Ph.D. degree in Computer Science at Zhejiang University, China. His research interests include generative digital humans, interactive media, and AI for science.

Yunfei Guo received the B.S. degree in electrical engineering from Yanshan University, Qinhuangdao, China, in 2002, and the Ph.D. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2007. From 2014 to 2015, he was a Visiting Professor with McMaster University, Hamilton, ON, Canada. He is currently a Professor with the School of Automation, Hangzhou Dianzi University, Hangzhou. His research interests include estimation, target detection, target tracking, and multimodal information fusion.

Chenggang Yan received the BS degree in control science and engineering from Shandong University, Shandong, China, in 2008 and the Ph.D. in computer science from the Chinese Academy of Sciences University, Beijing, China, in 2013. He is currently a professor in the Department of Automation, Hangzhou Dianzi University. His research interests include computational photography and pattern recognition and intelligent system.