RTGS: Metric-Scale RGB–Thermal Gaussian Splatting

Jian Zhang, Peiwei Lin, Runyu Chen, Hui Zheng, Yingying Wang, Huimin Guo,
Xiaotong Tu*, Yue Huang, and Xinghao Ding
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
School of Informatics, Xiamen University
journey@stu.xmu.edu.cn, {yhuang2010, xttu, dxh}@xmu.edu.cn

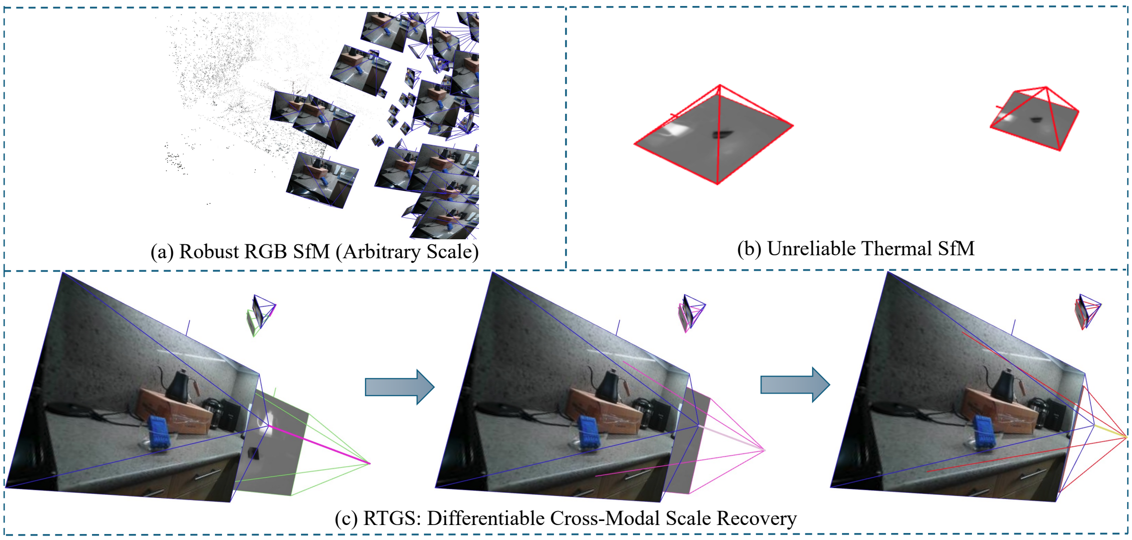

Fig. 1. (a) RGB SfM is robust but suffers from inherent scale ambiguity. (b) Thermal SfM is unreliable due to sparse and unstable textures. (c) RTGS leverages the RGB model as a geometric anchor to computationally recover metric scale, enabling precise thermal pose estimation.

***Abstract*—Metric-scale RGB-Thermal (RGB-T) 3D reconstruction is critical for applications like robotics and infrastructure inspection, requiring both geometric accuracy and multi-spectral information. However, since thermal SfM is often unreliable, effective fusion is hindered by the scale ambiguity inherent in monocular RGB Structure-from-Motion (SfM), which fails to align with the metric constraints of factory-calibrated sensor rigs. To address this, we propose a robust, differentiable framework for cross-modal scale recovery. Unlike methods relying on discrete search or explicit depth supervision, we formulate scale estimation as a continuous gradient-based optimization using a frozen RGB 3D Gaussian Splatting (3DGS) backbone as a geometric anchor. Specifically, we synthesize RGB images from candidate thermal viewpoints and bridge the modality gap using a structure-oriented objective (Normalized Gradient Fields). This allows us to exploit the structural correlation between RGB gradients and thermal signatures, backpropagating alignment errors directly to the global scale parameter despite drastic radiometric differences. Experiments on synthetic and real-world scenes demonstrate that our approach reliably recovers the metric scale from random initializations, enabling geometrically consistent RGB-T fusion and high-fidelity thermal novel view synthesis without altering the pre-trained RGB geometry.**

***Index Terms*—3D Gaussian splatting, RGB-thermal fusion, scale estimation, differentiable calibration, neural rendering**

* Corresponding author.

I. Introduction

RGB-Thermal (RGB-T) 3D reconstruction is pivotal for applications requiring both geometric precision and thermodynamic insights, such as building energy inspection [1], [2], precision agriculture [3], and search and rescue operations [4]. While RGB imaging offers detailed texture for geometry recovery, thermal imaging captures essential physical properties and remains effective in low-light environments. Consequently, creating a unified, metric-scale 3D model that seamlessly integrates these modalities is a highly compelling goal.

However, achieving RGB-T reconstruction faces a practical obstacle: cross-modal metric alignment. Due to sparse gradients and unstable textures in thermal data, performing Structure-from-Motion (SfM) on thermal streams is unreliable [5] (Fig. 1b). In contrast, monocular RGB SfM is robust but suffers from inherent scale ambiguity [6] (Fig. 1a). Recent zero-shot metric depth estimation methods [7] have shown remarkable progress, but rely on data-driven priors. They often suffer from domain gaps and scale drift in unseen environments, failing to meet the strict metric constraints of factory-calibrated sensor rigs. While a factory-calibrated stereo rig provides a metric transformation between sensors, this constraint is inconsistent with the scale produced by RGB

SfM. If this scale factor is not resolved, thermal views cannot be accurately anchored to the RGB geometry, resulting in misalignment and ghosting artifacts in the fused reconstruction.

Existing approaches [8]–[13] often circumvent this problem. Some methods rely on external physical measurements or manual intervention. For instance, ThermalNeRF [11] requires manual measurement of camera positions to resolve scale, while ThermoNeRF [12] depends on specific hardware features (e.g., FLIR MSX) or manual user alignment to correct parallax. Other works assume the scale is pre-solved or rely on heavy pipelines. ThermalGS [14] depends on complex 2D–3D correspondence matching which can be fragile in textureless thermal regions, and NTR-Gaussian [13] typically assumes a pre-existing 3D model is available for registration. These constraints limit the deployment of RGB-T reconstruction in autonomous, monocular video scenarios where human intervention or prior models are unavailable.

To address these limitations, we propose *RTGS*, a robust RGB-T Gaussian Splatting framework that computationally recovers the metric scale from monocular data without external measurements. We formulate scale estimation as a differentiable optimization problem (Fig. 1c). Our core insight is to utilize a frozen, pre-trained RGB 3DGS [15] model as a geometric anchor. By synthesizing RGB images from candidate thermal viewpoints and aligning them with thermal observations via a structure-oriented objective (Normalized Gradient Fields), we exploit the structural correlation between modalities, such as shared object boundaries, to backpropagate alignment errors directly to the global scale parameter. This allows us to establish a shared metric frame purely through optimization, bypassing radiometric discrepancies.

Once the metric scale is recovered, we follow a two-stage pipeline: we freeze the RGB geometry and adapt the model to the thermal modality by optimizing only thermal appearance parameters. By anchoring thermal observations to the robust RGB geometry, our method enables high-fidelity thermal novel view synthesis. In summary, our main contributions are:

- A differentiable calibration framework that resolves the scale ambiguity of monocular RGB SfM for RGB-T rigs without manual measurements or depth sensors.
- A structure-aware optimization strategy using a frozen RGB-3DGS backbone, which robustly aligns cross-modal data by exploiting gradient consistency.
- A complete pipeline for metric-scale RGB-T 3D reconstruction that achieves superior thermal rendering quality and alignment accuracy compared to baselines.

II. Related Work

RGB-Thermal Alignment and Scale Recovery. While factory-calibrated rigs provide metric extrinsics, they cannot be directly fused with monocular SfM trajectories, which inherently suffer from an unknown global scale [5], [6]. Consequently, the rigid inter-sensor transform is physically consistent only if the RGB trajectory is first anchored to the correct metric scale. To bridge this gap, existing approaches typically resort to strong priors or manual intervention. Hardware-dependent methods such as ThermoNeRF [12] and MrGS [16] rely on specific sensor features (e.g., FLIR MSX) or manual user alignment, which lacks automation and frequently ignores parallax effects in near-field scenes. ThermalNeRF [11] utilizes a calibrated rig but necessitates manual physical measurements of camera positions to resolve the scale ambiguity, limiting its autonomous application. Planar-assumption methods such as ThermalGaussian [8] warp thermal views via homography assuming a unit depth plane. While efficient, this approximation introduces severe geometric distortion in complex 3D environments. Matching-based methods such as ThermalGS [14] and NTR-Gaussian [13] attempt to register thermal images to RGB-derived meshes via 2D-3D correspondence (PnP) [17]. However, robust feature matching between texture-less thermal images and RGB geometry remains a fragile and computationally intensive challenge.

In contrast, we propose a fully computational framework that strictly respects factory calibration. We recover the metric RGB-T scene without external sensors or priors by utilizing the RGB geometry as a differentiable metric anchor.

Test-Time Camera Optimization. Our approach draws inspiration from analysis-by-synthesis pose refinement [18]. In the RGB domain, methods such as iNeRF [19] and GS-CPR [20] invert trained models to estimate camera poses by minimizing the photometric residual between rendered and observed images. Recent works such as iComMa [21] and GS-SLAM [22] have extended this to 3D Gaussian Splatting, leveraging fast rasterization for real-time optimization. However, these methods are predominantly designed for mono-modal (RGB-to-RGB) settings where pixel intensity consistency holds. We extend this paradigm to the cross-modal setting by replacing the photometric objective with a structure-oriented Normalized Gradient Fields (NGF) loss. This enables the transfer of differentiable pose optimization techniques to the challenging task of metric scale recovery, robustly aligning geometric gradients despite drastic radiometric discrepancies.

III. Method

An overview of our proposed framework is illustrated in Fig. 2. Our pipeline consists of three main stages: first, we train a standard RGB 3DGS model to reconstruct the scene geometry (Sec. III-B); second, we estimate the unknown global scale s to align thermal poses with the RGB reconstruction by minimizing a cross-modal gradient loss (Sec. III-C); finally, we freeze the geometry and adapt the appearance coefficients to synthesize high-fidelity thermal novel views (Sec. III-D).

A. Problem Formulation

Given N synchronized RGB–thermal image pairs $\mathcal{S} = \{(I_i^{\mathrm{rgb}}, I_i^{\mathrm{th}})\}_{i=1}^{N}$, we assume calibrated intrinsics $\mathbf{K}_{\mathrm{rgb}}, \mathbf{K}_{\mathrm{th}}$ (with distortion pre-corrected) and a rigid transform between sensors. We denote the calibrated RGB-to-thermal extrinsic as $\mathbf{T}_{\mathrm{rgb}\to\mathrm{th}} = (\mathbf{R}_{\mathrm{rt}}, \mathbf{t}_{\mathrm{rt}})$, where $\mathbf{t}_{\mathrm{rt}}$ is expressed in meters. RGB camera-to-world poses $\{\mathbf{T}_i^{\mathrm{rgb}} \in SE(3)\}$ are estimated by monocular SfM and are defined up to an unknown scale [6].

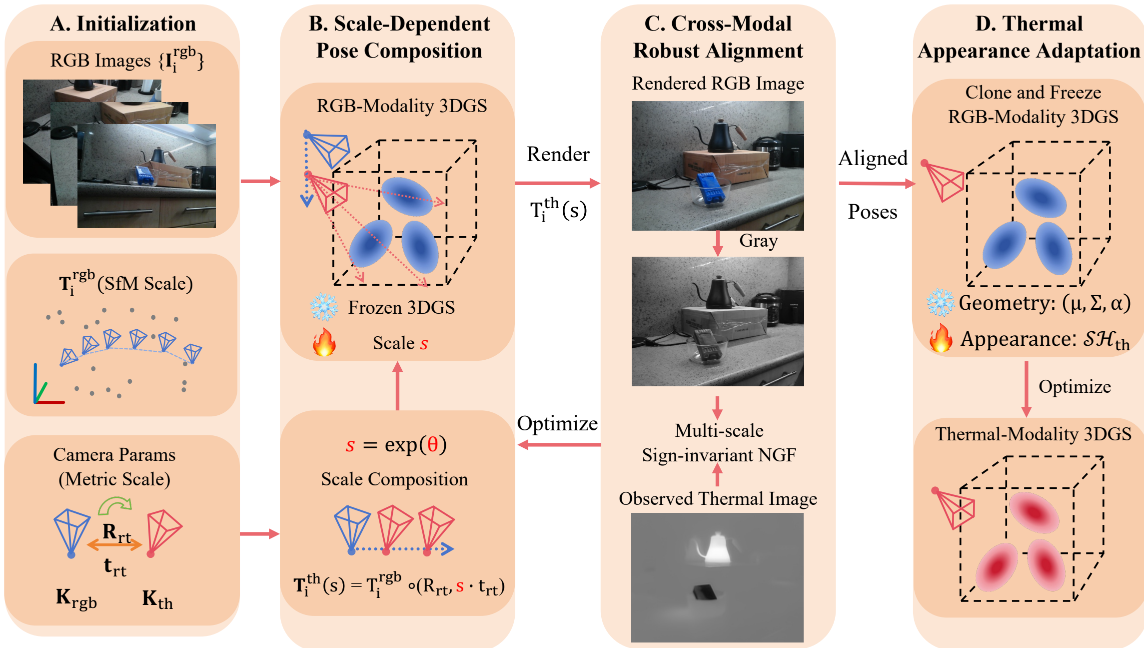

Fig. 2. Overview of our proposed method. We first train an RGB-3DGS model (Sec. III-B). To align the thermal sensors with the scale-ambiguous RGB reconstruction, we optimize a global scale factor s by minimizing a sign-invariant gradient loss between thermal images and grayscale RGB renderings (Sec. III-C). Finally, we adapt the model to the thermal modality by learning thermal spherical harmonics on the frozen geometry (Sec. III-D).

We treat the SfM reconstruction as the reference world frame; hence translations in $\mathbf{T}_i^{\text{rgb}}$ are measured in arbitrary *SfM units*. To express thermal poses in the same world frame, we introduce a single scalar $s > 0$ that converts the *metric* rig baseline to SfM units. Specifically, we scale only the baseline translation and compose it with the RGB pose:

$$\mathbf{T}_i^{\text{th}}(s) \;=\; \mathbf{T}_i^{\text{rgb}} \circ \big(\mathbf{R}_{\text{rt}},\, s\,\mathbf{t}_{\text{rt}}\big), \tag{1}$$

where $\circ$ denotes rigid-body composition in $SE(3)$ [6]. (Equivalently, $\bar{s} = 1/s$ maps SfM translations to metric units; here we use s for metric$\rightarrow$SfM conversion.) Our goal is to estimate $s^\star$ such that the thermal trajectory $\{\mathbf{T}_i^{\text{th}}(s^\star)\}$ becomes consistent with the SfM world frame, enabling consistent RGB–thermal reasoning up to a single global scale.

B. RGB 3D Gaussian Splatting Backbone

We adopt 3D Gaussian Splatting (3DGS) as a scene representation. We train an *RGB-only* 3DGS model using $\{I_i^{\text{rgb}}\}$ and $\{\mathbf{T}_i^{\text{rgb}}\}$ (up to an unknown scale). The learned scene is parameterized by a set of M 3D Gaussians, denoted as $\mathcal{G}_{\text{rgb}} = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k, \mathcal{SH}_k^{\text{rgb}}\}_{k=1}^M$, where $\boldsymbol{\mu}_k \in \mathbb{R}^3$, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{3\times3}$, and $\alpha_k \in [0,1]$ represent the position, anisotropic covariance, and opacity of the k-th Gaussian, respectively, and $\mathcal{SH}_k^{\text{rgb}}$ denotes the spherical harmonic (SH) coefficients for RGB appearance.

Geometry-frozen differentiable rendering. During scale estimation, we freeze the geometric parameters of $\mathcal{G}_{\text{rgb}}$ (i.e., $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k\}_{k=1}^M$) and use the model purely as a differentiable renderer. Given a camera pose $\mathbf{T}$ and intrinsics $\mathbf{K}$, we denote RGB rendering by

$$\hat{I}(\mathbf{T}, \mathbf{K}) \;=\; R(\mathcal{G}_{\text{rgb}}, \mathbf{T}, \mathbf{K}) \in \mathbb{R}^{H\times W\times 3}, \tag{2}$$

where $R(\cdot)$ is the 3DGS rasterizer.

Cross-modal observation model. To compare the RGB rendering with a thermal observation, we convert the rendered RGB image to grayscale:

$$\hat{I}_i^{\text{gray}}(s) \;=\; g\Big(\hat{I}(\mathbf{T}_i^{\text{th}}(s), \mathbf{K}_{\text{th}})\Big) \in \mathbb{R}^{H\times W}, \tag{3}$$

where $g(\cdot)$ is a fixed grayscale conversion.The thermal image I_i^{th} is normalized to $[0,1]$. Since I_i^{th} and $\hat{I}_i^{\text{gray}}(s)$ differ in radiometry (emission vs. reflection), we rely on radiometry-robust, structure-oriented losses for alignment.

C. Differentiable Scale Optimization

Radiometry-robust cross-modal loss. For each frame i, we render an RGB image from the thermal viewpoint $\mathbf{T}_i^{\text{th}}(s)$ with thermal intrinsics $\mathbf{K}_{\text{th}}$, convert it to grayscale $\hat{I}_i^{\text{gray}}(s)$ ((3)), and align it with the thermal observation I_i^{th}. The total alignment objective is the average of a multi-scale NGF loss (defined below in (7)) over a batch:

$$\mathcal{L}_{\text{xmod}}(s) \;=\; \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \ell_{\text{NGF-pyr}}\Big(\hat{I}_i^{\text{gray}}(s),\, I_i^{\text{th}}\Big), \tag{4}$$

where $\mathcal{I}$ is a set of frame indices for efficiency.

Sign-invariant NGF (single-scale). We first define the single-scale metric. Given two images A and B, let ∇ denote the spatial gradient operator (using standard 3×3 Sobel filters) and $\epsilon > 0$ a small constant. The normalized gradient of image $I \in \{A, B\}$ at pixel $\mathbf{p}$ is defined as:

$$\mathbf{n}_I(\mathbf{p}) = \frac{\nabla I(\mathbf{p})}{\sqrt{\|\nabla I(\mathbf{p})\|_2^2 + \epsilon^2}}. \tag{5}$$

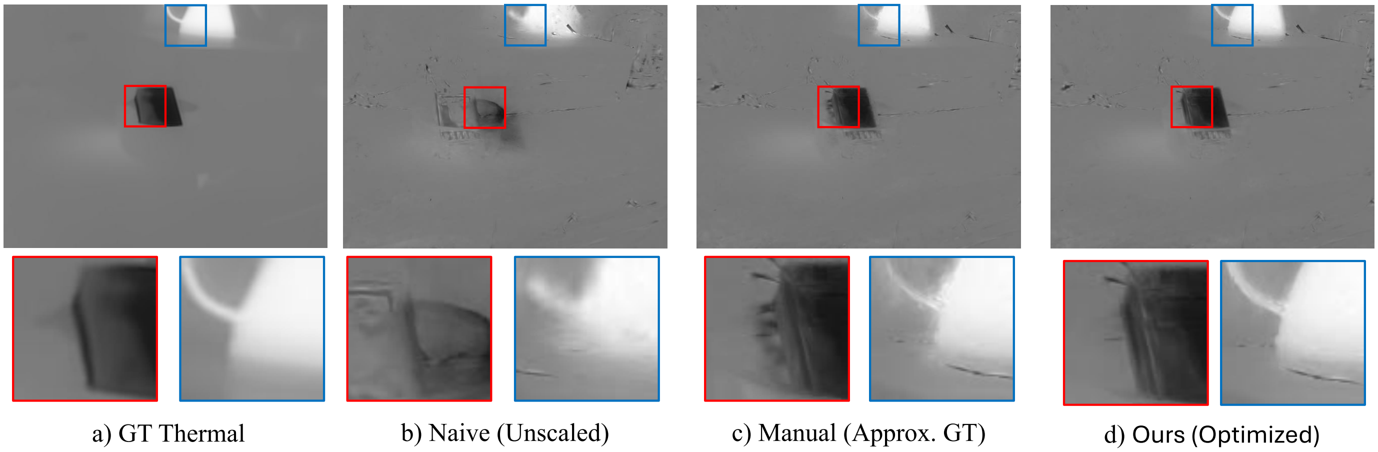

Fig. 3. **Qualitative comparison on a real-world sequence.** The top row displays the full thermal renderings, while the bottom row details the **black object boundary** (red box) and the **kettle spout** (blue box). **Naive (Unscaled):** Failing to recover the metric scale results in severe geometric overfitting, manifesting as high-frequency noise and ghosting artifacts. **Manual Calibration:** Despite using physical measurements, slight calibration errors still lead to visible blurring (e.g., the double edges in the red box). **RTGS (Ours):** By computationally optimizing the scale, our method eliminates ghosting and recovers sharp object boundaries, achieving superior fidelity consistent with the **Observed Thermal**.

Given that edge polarity can invert across RGB and thermal modalities (contrast reversal), we employ a sign-invariant NGF dissimilarity that ignores gradient direction:

$$\ell_{\mathrm{NGF}}(A,B)=\frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}\left(1-\big(\mathbf{n}_A(\mathbf{p})\cdot\mathbf{n}_B(\mathbf{p})\big)^2\right), \qquad (6)$$

where Ω denotes the set of pixels in the image domain. This term is minimized when gradients are parallel or anti-parallel.

Multi-scale pyramid loss. To improve convergence and suppress high-frequency noise, we optimize at multiple scales. Let $\mathcal{D}_l(\cdot)$ be a downsampling operator for pyramid level l (scale factor $1/2^l$), and let L be the number of levels. The multi-scale loss $\ell_{\text{NGF-pyr}}$ used in (4) is defined as:

$$\ell_{\text{NGF-pyr}}(A,B)=\frac{\sum_{l=0}^{L-1} w_l\,\ell_{\mathrm{NGF}}\big(\mathcal{D}_l(A),\,\mathcal{D}_l(B)\big)}{\sum_{l=0}^{L-1} w_l}, \qquad (7)$$

where w_l decays with l (we use $w_l = 0.5^l$) to prioritize coarse alignment. In practice, we set $L=3$ and $\epsilon = 10^{-6}$. Scale parameterization and optimization. We parameterize $s=\exp(\theta)$ and optimize θ using Adam:

$$\theta^\star \;=\; \arg\min_{\theta}\;\mathcal{L}_{\mathrm{xmod}}(\exp(\theta)), \qquad s^\star=\exp(\theta^\star). \qquad (8)$$

During optimization, we may use a random mini-batch of frames to reduce computation while keeping the RGB-3DGS parameters fixed. Gradients propagate to s through the scaled translation in (1) and the differentiable 3DGS renderer.

D. Thermal Appearance Adaptation

Once the optimal scale $s^\star$ is recovered, thermal poses $\{\mathbf{T}_i^{\mathrm{th}}(s^\star)\}$ are geometrically aligned with the RGB reconstruction. To synthesize high-fidelity thermal novel views while preventing geometry drift on low-texture thermal data, we adopt a geometry-frozen adaptation strategy.

Specifically, we clone the RGB-3DGS model and replace RGB appearance coefficients with a new set of thermal appearance coefficients $\mathcal{SH}_{\mathrm{th}}=\{\mathcal{SH}_k^{\mathrm{th}}\}$ (initialized to zero or small random noise), while keeping all geometry parameters fixed. We then optimize only $\mathcal{SH}_{\mathrm{th}}$ using thermal images $\{I_i^{\mathrm{th}}\}$ and the fixed thermal poses. Since this stage is within a single modality (thermal-to-thermal), standard photometric rendering losses (e.g., ℓ_1 with optional SSIM) can be employed. This strict geometry freezing prevents "floating" artifacts and preserves cross-modal geometric consistency.

IV. Experiments

Obtaining ground-truth metric scale in real-world RGB–thermal sequences is often impractical. We therefore evaluate our method on both synthetic and real-world data. On synthetic scenes, where the ground-truth scale s_{gt} is known, we directly measure scale-recovery accuracy and study robustness under controlled image and calibration perturbations. On real-world scenes, where s_{gt} is unavailable, we assess the downstream effect of the recovered scale using a geometry-frozen thermal reconstruction protocol and report image-quality metrics.

A. Synthetic Evaluation

To directly evaluate scale recovery with ground truth, we conduct controlled experiments on synthetic RGB–thermal data rendered in a metric coordinate system. The synthetic setup provides ground-truth metric poses for both modalities; we then inject a global scale mismatch into the RGB–thermal rig and test whether our optimization can recover the correct metric alignment using only RGB poses and thermal images.

Dataset and Rendering. We use six scenes from the NeRF-Synthetic [23] benchmark (Chair, Drums, Hotdog, Lego, Materials, Ship). For each scene, we render N=50 paired RGB–thermal views at 800×800 with Blender. The RGB and thermal

TABLE I

SCALE OPTIMIZATION ACCURACY ON SYNTHETIC SCENES. COMPARISON AGAINST INITIALIZATION AND PHOTOMETRIC BASELINES. ALL ERROR METRICS ARE REPORTED IN PERCENTAGE POINTS.

Method	Med Err↓	P90 Err↓	SR@2↑	Med Iter↓
Init	36.47	77.81	10	N/A
iNeRF (SSIM)	1.40	308.96	53	**25**
iNeRF (ℓ_1)	0.41	20.86	77	33
NGF (single-scale)	0.32	34.02	83	51
Ours (pyramid)	**0.19**	**1.22**	**93**	41

TABLE II

SENSITIVITY ANALYSIS. COMPARISON OF INERF (ℓ_1) VS. OURS UNDER DATA AND CALIBRATION PERTURBATIONS. WE REPORT MEDIAN RELATIVE ERROR (MED, %) AND SUCCESS RATE (SR@2%, %).

		iNeRF (ℓ_1)		Ours (NGF)	
Perturbation	Level	Med (%)	SR (%)	Med (%)	SR (%)
Reference	0	0.52	75.00	**0.35**	**91.67**
Thermal Blur σ	2 px	0.66	66.67	**0.34**	**91.67**
	4 px	1.04	66.67	**0.59**	**100.00**
AGC jitter (a,b)	Mild	0.42	66.67	**0.32**	**91.67**
Rig Δt (%)	1%	1.08	83.33	**0.76**	**100.00**
	2%	1.95	58.33	**1.78**	**58.33**
Intrinsics $\Delta f/f$	1%	0.89	75.00	**0.48**	**91.67**
	2%	1.87	50.00	**0.92**	**91.67**

cameras form a rigid rig with $\|\mathbf{t}_{rt}\| = 0.20\,\text{m}$ and a fixed relative rotation. Thermal images are synthesized as single-channel intensity maps using emission-based materials and ambient occlusion, producing physically plausible shading that is distinct from RGB textures.

Implementation Details. We first train an RGB-only 3DGS model using the `splatfacto` implementation [24] (30k iterations), and use it as a frozen renderer during scale recovery. In our synthetic setup, the RGB and thermal cameras have metric ground-truth poses and the rig translation $\mathbf{t}_{rt}$ is given in meters. To emulate the global scale ambiguity of monocular RGB SfM, we perturb the thermal camera trajectory by scaling its translation with a random factor $s_0 \sim \mathcal{U}(0.5, 2.0)$, while keeping rotations fixed. We then optimize a single global scale $s = \exp(\theta)$ to undo this perturbation and recover the metric trajectory, starting from $s \leftarrow s_0$ and running Adam (lr=5×10^{-3}) for 200 steps. Unless specified, we use all N=50 frames and aggregate results over 30 trials (6 scenes × 5 random seeds).

Baselines and Metrics. On synthetic scenes, we compare different alignment signals for scale optimization. We compare against: (i) Init: use the random initialization s_0 without optimization; (ii) iNeRF (SSIM): optimize s by minimizing an SSIM-based photometric loss between the rendered RGB view and the thermal observation (iNeRF-style); (iii) iNeRF (ℓ_1): optimize s using an ℓ_1 photometric loss (iNeRF-style); (iv) NGF (single-scale): replace photometric matching with our structure-based NGF objective at a single scale; (v) Ours (pyramid): our full multi-scale NGF pyramid optimization. Synthetic scenes are rendered in metric space; hence $s_{gt} \equiv 1$. We measure scale accuracy using Relative Error $\text{RelErr} = 100\%\,|\hat{s} - s_{gt}|/s_{gt}$. We report the median (Med), 90th percentile (P90), and success rate within 2% error (SR@2%).

Results and Robustness Analysis. Table I summarizes scale recovery performance. The iNeRF baselines (ℓ_1/SSIM) improve over initialization but exhibit heavy tails due to the modality gap—thermal emission does not consistently correlate with RGB intensity. In contrast, our structure-oriented NGF Pyramid achieves the best overall accuracy (0.19% median error) and reliability (93% SR@2%), indicating that gradient-field consistency provides a more stable signal for cross-modal alignment than raw photometric matching.

To further evaluate robustness, we apply controlled perturbations to thermal observations and camera parameters. As reported in Table II, our method remains resilient to thermal blur (σ=2∼4 px) and mild AGC jitter. For AGC jitter, we apply a linear contrast perturbation around the image mean with the contrast factor sampled from $[0.9, 1.1]$. We also perturb calibration by adding errors to the rig baseline (Δt) and focal length ($\Delta f/f$); overall, our NGF-based optimization tolerates 1% baseline error and 2% focal length error while keeping the median error below 1%.

B. Real-World Evaluation

To validate our approach in practical scenarios, we evaluate on 9 real-world scenes from the ThermalNeRF dataset [11]. These sequences were captured using a smartphone-connected FLIR One Pro sensor, which simultaneously captures aligned RGB and LWIR (long-wave infrared) images.

Experimental Protocol: Geometry-Frozen Appearance Fitting. Unlike synthetic tests where the geometry is noise-free, real-world cross-modal SfM inevitably contains pose and reconstruction errors. To strictly isolate the effect of metric scale recovery—and decouple it from 3DGS geometry optimization—we adopt a *Geometry-Frozen* protocol: we first train a standard RGB-3DGS model using COLMAP poses; we then align the thermal camera rig to the RGB model by estimating a global scale factor s with five strategies (defined below); finally, we freeze all geometric parameters (position, rotation, scaling, opacity) as well as the RGB SH coefficients, and optimize only a new set of thermal SH coefficients.

Baselines. We compare five strategies for estimating the global scale factor s that aligns the metric RGB–thermal rig to the RGB SfM trajectory: (i) Naive (Unscaled): apply the metric rig transform $\mathbf{T}_{rt}$ without any scale correction (s=1); (ii) Manual (Measured): compute s from the physically measured RGB–thermal baseline reported in [11] and treat it as a practical reference; (iii) MASt3R (Metric Pose) [25]: estimate metric-scale relative poses between RGB views, obtain a baseline in meters, and use it to convert the SfM trajectory to metric scale; (iv) Planar Reproj. (TG-style) [8]: assume a constant depth (d=1) and warp thermal observations onto RGB views via homography, providing an implicit proxy

TABLE III

REAL-WORLD EVALUATION (AVG.). THERMAL RENDERING QUALITY ON 9 THERMALNERF SCENES UNDER THE GEOMETRY-FROZEN PROTOCOL.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Manual (Measured)	31.22	0.94	0.34
Naive (Unscaled)	29.47	0.93	0.37
MASt3R (Metric Pose)	29.81	0.93	0.36
Planar (TG-style)	23.13	0.55	0.63
Ours (NGF)	**31.97**	**0.94**	**0.34**

under a strong planar assumption; (v) Ours (NGF Pyramid): recover s via our differentiable NGF pyramid optimization using only RGB geometry and unposed thermal images.

Results. We evaluate real-world performance using thermal rendering quality under the geometry-frozen protocol. As shown in Table III, Naive (Unscaled) suffers from parallax mismatch caused by incorrect metric alignment, leading to degraded PSNR/SSIM and higher LPIPS. MASt3R (Metric Pose) improves over Naive, but remains limited by metric-pose estimation noise and domain gaps. The Planar (TG-style) baseline performs poorly, indicating that fixed-depth homography warping is inadequate for scenes with non-planar structure and depth variation. In contrast, Ours (NGF) achieves the best overall performance, matching Manual (Measured) and even improving average PSNR (31.97 vs. 31.22 dB), suggesting that computational scale recovery can effectively replace manual measurements in practice.

V. CONCLUSION

We introduced RTGS, a metric-scale RGB–thermal (RGB–T) Gaussian splatting framework. RTGS resolves the scale ambiguity of monocular SfM via differentiable calibration using a frozen RGB-3DGS model as a geometric anchor, and optimizes the global scale with a structure-oriented NGF objective. Experiments on synthetic and real-world datasets show that RTGS recovers metric scale reliably and improves RGB–T alignment without external sensors or manual measurements. Limitations include scenes with extremely weak thermal gradients or large factory calibration errors; future work will jointly refine extrinsics with scale and extend to dynamic RGB–T reconstruction.

ACKNOWLEDGMENT

This work was supported by the Dreams Foundation of Jianghuai Advance Technology Center under Grant Nos. 2023-ZM01X018 and 2023-ZM01D002.

REFERENCES

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

[2] 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.

[3] 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.

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

[5] 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.

[6] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[7] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "Unidepth: Universal monocular metric depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 106–10 116.

[8] R. Lu, H. Chen, Z. Zhu, Y. Qin, M. Lu, L. Zhang, C. Yan, and A. Xue, "Thermalgaussian: Thermal 3d gaussian splatting," *arXiv preprint arXiv:2409.07200*, 2024.

[9] 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.

[10] C. Zou, Q. Ma, J. Wang, R. Lu, M. Lu, and Z. Qu, "Tga-gs: Thermal geometrically accurate gaussian splatting," *Applied Sciences*, vol. 15, no. 9, p. 4666, 2025.

[11] 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.

[12] M. Hassan, F. Forest, O. Fink, and M. Mielle, "Thermonerf: A multimodal neural radiance field for joint rgb-thermal novel view synthesis of building facades," *Advanced Engineering Informatics*, vol. 65, p. 103345, 2025.

[13] K. Yang, Y. Liu, Z. Cui, Y. Liu, M. Zhang, S. Yan, and Q. Wang, "Ntr-gaussian: Nighttime dynamic thermal reconstruction with 4d gaussian splatting based on thermodynamics," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 691–700.

[14] Y. Liu, X. Chen, S. Yan, Z. Cui, H. Xiao, Y. Liu, and M. Zhang, "Thermalgs: Dynamic 3d thermal reconstruction with gaussian splatting," *Remote Sensing*, vol. 17, no. 2, p. 335, 2025.

[15] 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.

[16] M. Kweon, J. Kim, U. Shin, and J. Park, "Mrgs: Multi-modal radiance fields with 3d gaussian splatting for rgb-thermal novel view synthesis," *arXiv preprint arXiv:2511.22997*, 2025.

[17] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, pp. 155–166, 2009.

[18] Z. Fan, W. Cong, K. Wen, K. Wang, J. Zhang, X. Ding, D. Xu, B. Ivanovic, M. Pavone, G. Pavlakos *et al.*, "Instantsplat: Sparse-view gaussian splatting in seconds," *arXiv preprint arXiv:2403.20309*, 2024.

[19] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, "inerf: Inverting neural radiance fields for pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1323–1330.

[20] C. Liu, S. Chen, Y. Bhalgat, S. Hu, M. Cheng, Z. Wang, V. A. Prisacariu, and T. Braud, "Gs-cpr: Efficient camera pose refinement via 3d gaussian splatting," *arXiv preprint arXiv:2408.11085*, 2024.

[21] Y. Sun, X. Wang, Y. Zhang, J. Zhang, C. Jiang, Y. Guo, and F. Wang, "icomma: Inverting 3d gaussian splatting for camera pose estimation via comparing and matching," *arXiv preprint arXiv:2312.09031*, 2023.

[22] 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.

[23] 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.

[24] 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.

[25] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *European Conference on Computer Vision*. Springer, 2024, pp. 71–91.