PHYSICS-INSPIRED RGB-T 3D GAUSSIAN SPLATTING WITH MULTIMODAL INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-Thermal (RGB-T) 3D reconstruction combines the detail richness of RGB with the all-weather capability of thermal imaging, but faces challenges in jointly balancing both modalities and modeling their physical correlations. Prior RGB-T methods either degrade RGB reconstruction quality or overlook cross-modal physical relationships, leading to artifacts in distant regions and near object boundaries. To address these challenges, we propose PRT-3D, a novel physics-inspired RGB-T 3D Gaussian Splatting (3DGS) reconstruction framework. Unlike previous data-driven methods, PRT-3D draws inspiration from the physical principles governing thermal imaging to design cross-modal interaction networks that enhance reconstruction. Specifically, we introduce a Light Transmission Network (LTN) that, inspired by atmospheric transmission attenuation, leverages the depth consistency of the RGB modality to improve thermal transmission distance estimation. Furthermore, we design a Thermal Conduction Network (TCN) that, inspired by heat diffusion effects, exploits the edge consistency of RGB images to mitigate thermal boundary blurring. Experiments demonstrate that our method significantly improves thermal reconstruction quality over state-of-the-art approaches (e.g., improving thermal PSNR over ThermalGaussian by 3.72 dB on average), while maintaining RGB quality comparable to or surpassing previous best results, and requiring even less GPU memory on a single RTX 3090 GPU.

1 INTRODUCTION

RGB image-based 3D reconstruction Mildenhall et al. (2021); Kerbl et al. (2023); Wu et al. (2024) has attracted widespread attention. The two most representative approaches are neural radiance fields (NeRF) and 3D Gaussian Splatting (3DGS). While RGB-based 3D reconstruction methods offer high clarity and rich detail, they are highly susceptible to environmental conditions such as weather and lighting. For instance, adverse weather and low-light (dark) environment can significantly degrade reconstruction quality.

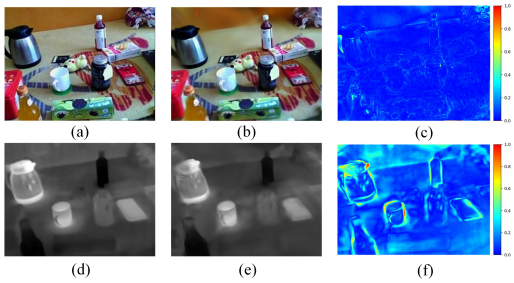

Figure 1: Reconstruction error analysis of ThermalGaussian. (a) Ground truth RGB image. (b) Rendered RGB image. (c) Visualization of RGB Reconstruction errors. (d) Ground truth Thermal image. (e) Rendered Thermal image. (f) Visualization of Thermal Reconstruction errors.

In contrast, thermal image-based 3D reconstruction Ye et al. (2024); Liu et al. (2025) offers all-weather imaging and strong penetration capabilities, enabling clear imaging at night and under adverse weather conditions. However, thermal image-based reconstruction is also prone to issues such as floaters and boundary blurring, which can lead to a decline in reconstruction quality.

RGB-Thermal (RGB-T) 3D reconstruction offers a new solution by leveraging the advantages of both RGB and thermal modalities while mitigating their respective limitations. However, balancing the reconstruction of both modalities and modeling their physical correlations introduces new

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

challenges. To the best of our knowledge, only two works, ThermoNeRF Hassan et al. (2025) and ThermalGaussian Lu et al. (2025), have focused on RGB-T 3D reconstruction. However, ThermoNeRF suffers from reduced RGB reconstruction quality after incorporating the thermal modality. ThermalGaussian achieves competitive RGB reconstruction quality, but both methods do not fully consider the physics of the thermal modality nor the underlying physical relationships between RGB and thermal modality, leaving considerable room for improvement in thermal reconstruction.

Our preliminary analysis shows that ThermalGaussian Lu et al. (2025), the strongest baseline, suffers from large thermal errors especially at long ranges and object boundaries (Figure 1), likely due to missing physics such as atmospheric attenuation and heat diffusion. We also observe depth and edge inconsistencies between RGB and thermal outputs (Figure 2), indicating weak cross-modal interaction. These findings motivate a physics-inspired method that leverages cross-modal geometric consistency to improve thermal reconstruction quality without compromising the already strong RGB performance.

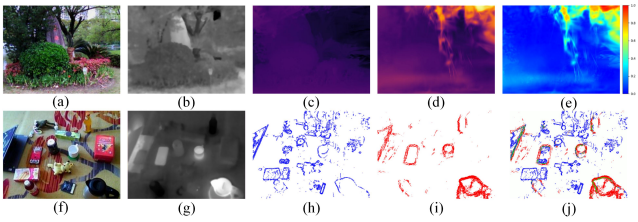

Figure 2: Modality consistency analysis of ThermalGaussian. (a,f) Rendered RGB image. (b,g) Rendered thermal image. (c) RGB depth map. (d) Thermal depth map. (e) Visualization of cross-modal depth discrepancies. (h) RGB edge map. (i) Thermal edge map. (j) Visualization of cross-modal edge discrepancies. See *Supplementary Material (SM)* for more examples.

Based on the above analysis, we propose PRT-3D, a physics-inspired RGB-T 3DGS reconstruction framework. Unlike previous approaches that are purely data-driven and treat the two modalities as relatively independent, our method draws inspiration from the physical principles governing thermal imaging to design cross-modal interaction networks that bridge the two modalities. Specifically, we introduce a Light Transmission Network (LTN) and a Thermal Conduction Network (TCN). Inspired by the atmospheric transmission attenuation of infrared radiation, the LTN leverages the "depth consistency" between visible and thermal modalities to enhance thermal transmission distance estimation. Inspired by the heat diffusion process that blurs thermal boundaries, the TCN exploits the "edge consistency" of the visible modality to improve thermal edge sharpness.

Experiments demonstrate that our method significantly improves thermal reconstruction quality over state-of-the-art approaches (e.g., improving thermal PSNR over ThermoNeRF and ThermalGaussian by 8.63 dB and 3.72 dB respectively), while maintaining RGB reconstruction quality comparable to or even surpassing strongest baseline ThermalGaussian. Furthermore, our method reduces memory consumption by 19% compared to ThermalGaussian and training time by 68% compared to ThermoNeRF.

2 Related Works

RGB 3D Reconstruction. RGB 3D reconstruction has evolved from explicit SfM/MVS pipelines Schonberger & Frahm (2016); Yao et al. (2018) to NeRF Mildenhall et al. (2021) and its variants Barron et al. (2021); Müller et al. (2022); Wang et al. (2021), and has recently converged on 3D Gaussian Splatting (3DGS) Kerbl et al. (2023), which achieves real-time, high-quality rendering via learnable anisotropic Gaussians and differentiable rasterization Wu et al. (2024); Niedermayr et al. (2024). However, these methods assume visible-light imaging and cannot be directly applied to RGB–T reconstruction, where thermal-specific physical effects must be modeled.

Thermal 3D Reconstruction. The lack of texture in infrared imagery long hindered conventional SfM/MVS pipelines Vidas et al. (2013). NeRF-based methods Ye et al. (2024); Carmichael et al. (2025) jointly optimize radiance and temperature fields and improve geometric consistency under sparse views, but suffer from low training and inference efficiency. More recently, the state-of-the-art 3DGS-based method Liu et al. (2025) leverages anisotropic Gaussian primitives for real-time,

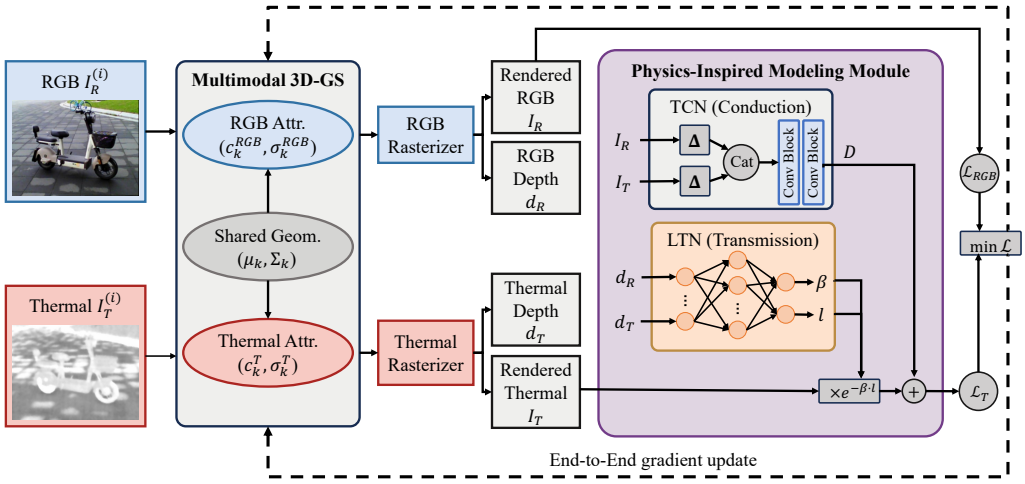

Figure 3: Pipeline of the proposed PRT-3D method.

high-fidelity thermal radiance field reconstruction. However, existing pure-thermal 3DGS solutions do not directly extend to RGB–T multimodal reconstruction, as they account for neither modality balancing nor cross-modal geometric consistency.

RGB-T 3D Reconstruction. To the best of our knowledge, only two works, ThermoNeRF Hassan et al. (2025) and ThermalGaussian Lu et al. (2025), have focused on RGB-T 3D reconstruction. ThermoNeRF leverages paired RGB and thermal images to synthesize novel views and temperature data for building facades, but its RGB quality degrades after incorporating the thermal modality. ThermalGaussian enables real-time, high-quality rendering of both modalities via multimodal 3D Gaussian Splatting. However, both methods remain largely data-driven, considering neither the physics of thermal imaging nor the underlying physical correlations between the RGB and thermal modalities.

3 Preliminary Analysis

Reconstruction Error Analysis. To understand the limitations of the strongest baseline, ThermalGaussian Lu et al. (2025), we analyze error maps between its renderings and the ground truth on 100 randomly selected outputs (Figure 1; more in *SM*). A clear asymmetry emerges: RGB errors are small and uniformly distributed, whereas thermal errors are substantially larger and concentrate in two characteristic regions—distant areas and object boundaries. Crucially, this error pattern is not random: it matches the signatures of two physical phenomena that the reconstruction pipeline ignores, namely atmospheric transmission attenuation, which weakens thermal radiation with increasing distance, and heat diffusion, which transfers heat from hotter to cooler regions and blurs thermal boundaries. This suggests that the current SOTA method remains essentially data-driven, and motivates physics-inspired mechanisms for high-quality thermal reconstruction.

Modality Consistency Analysis. The above asymmetry raises a natural question: *can the stronger RGB modality be exploited to enhance thermal reconstruction?* Since the two aligned modalities capture the same scene, object depths and edges in the thermal view should be geometrically consistent with those in the visible view. To test whether ThermalGaussian exploits this shared geometry, we compute the L1 discrepancy between its RGB and thermal outputs on 100 samples, over both rendered depth maps and Laplacian edge maps (Figure 2; more in *SM*). The results reveal noticeable cross-modal inconsistencies in both depth and edge structure, showing that the shared geometry is left underexploited. Together, these two analyses motivate our design: physics-inspired modality-interaction subnetworks that inject the reliable geometry of the RGB modality into thermal reconstruction, precisely where the physical degradations occur.

4 METHOD

4.1 PROBLEM FORMULATION

Given a set of aligned RGB and thermal images $\{I_R^{(i)}, I_T^{(i)}\}_{i=1}^N$, our objective is to jointly reconstruct the 3D representations X_R and X_T such that their renderings match the observations, where $f_{\theta_R^{(i)}}$ and $f_{\theta_T^{(i)}}$ denote the renderer with view-specific parameters for each modality. Unlike RGB imaging, real-world thermal imaging is governed by unique physical phenomena such as atmospheric transmission attenuation and thermal conduction, which standard 3DGS does not account for, causing artifacts and blurring in reconstructed thermal images. We therefore introduce a transformation $P_\varphi(\cdot)$ that models the effect of these physical processes on thermal images, with physical parameters φ, and formulate the optimization objective as:

$$\min_{X_R, X_T} \sum_{i=1}^{N} \left| f_{\theta_R^{(i)}}(X_R) - I_R^{(i)} \right| + \left| P_\varphi \left(f_{\theta_T^{(i)}}(X_T) \right) - I_T^{(i)} \right|. \tag{1}$$

Since φ is difficult to measure directly and estimating it solely from the thermal modality may introduce biases, we train a neural network g that exploits the geometric correlations between the two aligned modalities:

$$\tilde{\varphi} = g\left(I_R^{(i)}, I_T^{(i)}\right), \quad i = 1, 2, \ldots, N. \tag{2}$$

In our framework, P_φ encompasses atmospheric attenuation, parameterized by the attenuation coefficient β and transmission distance l, and thermal conduction compensation, characterized by a residual D, i.e., $\varphi = \{\beta, l, D\}$; their formulations and estimation networks are detailed in Sections *Light Transmission Network* and *Thermal Conduction Network*. The entire pipeline is optimized end-to-end, jointly updating X_R, X_T, and g.

4.2 OVERALL PIPELINE

The pipeline of PRT-3D is shown in Figure 3. The framework leverages a multimodal 3D Gaussian Splatting representation that maintains shared geometric parameters while decoupling RGB and thermal appearance attributes. Following rasterization, the rendered images and depth maps are processed by a Physics-Inspired Modeling Module. This module draws inspiration from thermal physics via a Thermal Conduction Network to compensate for heat diffusion effects and a Light Transmission Network to estimate depth-dependent atmospheric attenuation. The final output fuses these physical priors with the rendered features, enabling the entire system to be optimized end-to-end through joint reconstruction supervision.

4.3 MULTIMODAL 3D GAUSSIAN SPLATTING

To unify scene geometry while capturing distinct modal properties, we employ a multimodal 3DGS framework: the scene is represented by a set of 3D Gaussians whose geometry is shared across modalities to ensure structural consistency, while appearance attributes are modeled separately. Following Kerbl et al. (2023), each Gaussian is defined by its center $\mu_k \in \mathbb{R}^3$ and anisotropic covariance $\Sigma_k \in \mathbb{R}^{3\times3}$. On top of this shared spatial structure (μ, Σ), each Gaussian carries modality-specific opacities (σ^{RGB}, σ^T) and spectral features (c^{RGB}, c^T), allowing surfaces with consistent geometry but different transparency across spectra (e.g., glass).

Rendering follows the standard splatting technique. For modality $m \in \{RGB, T\}$, the pixel intensity is computed by α-blending the depth-ordered Gaussians L overlapping the pixel:

$$C = \sum_{i \in L} c_i^{(m)} \alpha_i^{(m)} \prod_{j=1}^{i-1} \left(1 - \alpha_j^{(m)}\right), \tag{3}$$

where $\alpha_i^{(m)} = \sigma_i^{(m)} G_i^{2D}$ combines the modality-specific opacity with the projected 2D Gaussian, so that both transparency and appearance are optimized per modality. The RGB and thermal rasterizers independently produce the rendered images (I_R, I_T) and depth maps (d_R, d_T), which serve as inputs to our physics-inspired modules.

Table 1: Thermal (top) and RGB (bottom) reconstruction results on various scenes. Best results are in **bold**. **Dsum**: Dimsum; **Bldg.**: Building; **Parr.**: Parterre.

	Metric	Meth.	Dsum	Daily Stuff	Ebike	Road Block	Truck	Rotary Kiln	Bldg.	Iron Ingot	Parr.	Land Scape	Dark Scene	Avg.
Thermal	PSNR ↑	TNF	24.27	17.34	19.70	17.17	23.53	26.40	23.31	22.97	17.88	18.79	11.80	20.28
		TGS	26.46	22.28	23.31	27.17	25.88	26.33	26.72	29.86	26.16	22.27	20.65	25.19
		Ours	**29.97**	**25.47**	**27.91**	**31.82**	**29.56**	**30.47**	**30.12**	**33.31**	**29.88**	**24.38**	**25.12**	**28.91**
	SSIM ↑	TNF	0.747	0.759	0.694	0.781	0.750	0.916	0.804	0.717	0.709	0.774	0.666	0.756
		TGS	0.886	0.835	0.862	0.919	0.874	0.922	0.888	0.896	0.883	0.850	0.825	0.877
		Ours	**0.944**	**0.922**	**0.946**	**0.977**	**0.903**	**0.954**	**0.947**	**0.919**	**0.946**	**0.901**	**0.937**	**0.936**
	LPIPS ↓	TNF	0.312	0.494	0.290	0.293	0.291	0.170	0.234	0.152	0.309	0.264	0.598	0.309
		TGS	0.129	0.210	0.203	0.198	**0.136**	0.124	0.177	0.091	**0.181**	**0.248**	0.291	0.181
		Ours	**0.113**	**0.203**	**0.194**	**0.186**	0.150	**0.115**	**0.148**	**0.082**	0.185	0.255	**0.280**	**0.174**
RGB	PSNR ↑	TNF	19.74	16.79	17.75	18.32	18.77	18.89	17.12	15.07	23.13	19.13	18.82	18.49
		TGS	24.38	21.76	26.85	28.12	24.17	**23.14**	**24.19**	24.55	25.48	**21.71**	29.21	24.78
		Ours	**24.58**	**22.09**	**27.52**	**28.94**	**24.21**	21.71	23.36	**25.72**	**25.68**	20.47	**30.00**	**24.94**
	SSIM ↑	TNF	0.688	0.639	0.540	0.619	0.688	0.600	0.460	0.293	0.756	0.583	0.419	0.571
		TGS	0.858	0.797	0.905	0.920	0.840	**0.822**	0.849	0.884	0.855	**0.739**	0.878	0.850
		Ours	**0.861**	**0.803**	**0.921**	**0.922**	**0.843**	0.799	**0.853**	**0.895**	**0.872**	0.709	**0.891**	**0.852**
	LPIPS ↓	TNF	0.228	0.465	0.244	0.548	0.311	0.207	0.291	0.301	**0.167**	0.275	0.632	0.333
		TGS	0.194	0.253	0.169	**0.199**	**0.211**	**0.184**	0.170	0.186	0.195	**0.268**	0.284	**0.210**
		Ours	**0.193**	**0.250**	**0.153**	0.217	0.218	0.196	**0.158**	**0.175**	0.187	0.280	**0.281**	**0.210**

4.4 Light Transmission Network

Thermal imaging is intrinsically subject to atmospheric attenuation, where radiation intensity decays exponentially with propagation distance due to molecular (water vapor, CO_2, etc.) absorption and scattering. Standard 3D Gaussian Splatting (3D-GS) neglects this physical phenomenon, often resulting in incorrect temperature estimates and surface artifacts. To address this, we explicitly model the atmospheric transmission process using the Bouguer-Lambert-Beer law Kiteto & Mecha (2024):

$$T = T_0 e^{-\beta l}, \tag{4}$$

where T denotes the transmitted thermal radiation intensity, T_0 is the initial intensity at the source, l is the transmission distance, and β is the atmospheric attenuation coefficient. In practice, both β and l are difficult to measure directly and can vary with spatial location. We therefore design a neural network, termed the Light Transmission Network (LTN), to estimate both β and l.

A natural question arises: *what inputs should the LTN receive?* To answer this, we analyze the estimation of (β, l) through the lens of maximum a posteriori (MAP) inference. Given the thermal depth observation d_T and RGB depth prior d_R, the posterior distribution of (β, l) can be factorized by Bayes' rule as (see *SM* for full derivation):

$$p(l, \beta \mid d_T, d_R) \propto p(d_T \mid l, \beta)\, p(\beta)\, p(l \mid d_R). \tag{5}$$

Maximizing this posterior is equivalent to minimizing the negative log-posterior:

$$\arg\min_{l,\beta} - \underbrace{\log p(d_T \mid l, \beta)}_{\text{data likelihood}} - \underbrace{\log p(\beta)}_{\text{medium prior}} - \underbrace{\log p(l \mid d_R)}_{\text{depth prior}}. \tag{6}$$

The key insight lies in the depth prior term $p(l \mid d_R)$. Since the RGB and thermal images are spatially aligned and capture the same scene, the transmission distance l should be geometrically consistent with the depth l' estimated from the RGB branch. This "transmission depth consistency" prior can be modeled as:

$$p(l \mid d_R) \propto \exp\left(-\frac{|l - l'|^2}{2\sigma^2}\right), \tag{7}$$

which reveals that accurate estimation of (β, l) fundamentally benefits from incorporating the RGB depth as a cross-modal geometric reference. This analysis motivates our design choice of feeding both the thermal depth map d_T and the RGB depth map d_R as inputs to the LTN:

$$(\beta, l) = f_{LTN}(d_T, d_R), \tag{8}$$

enabling the network to implicitly learn the depth consistency constraint during end-to-end optimization. Finally, based on the estimated β and l, we reconstruct the thermal image accounting for atmospheric attenuation via Equation 4.

4.5 THERMAL CONDUCTION NETWORK

Another important physical phenomenon in the thermal imaging process is thermal conduction: heat flows from regions of higher temperature to regions of lower temperature, as described by the heat equation Widder (1976):

$$\partial \mathrm{T}/\partial \mathrm{t} = \kappa \Delta \mathrm{T}, \tag{9}$$

where κ denotes the thermal diffusivity and the 2D Laplacian Δ measures the spatial curvature of the temperature field. Physically, temperature evolution is driven by local gradients: regions locally hotter than their surroundings ($\Delta T < 0$) cool down, while cooler regions ($\Delta T > 0$) heat up. Thermal conduction thus naturally smooths temperature gradients, leading to edge blurring and loss of high-frequency detail in thermal images. To compensate for this diffusion-induced degradation, we aim to estimate a residual $D = I_C - I_T$ that recovers the sharp thermal image I_C from the rendered thermal image I_T.

We determine the inputs of the compensation network following the same MAP-based design principle as in Section *Light Transmission Network*: by Bayes' theorem, the posterior of D given the thermal image I_T and the edge maps of both modalities factorizes into a data likelihood and an edge prior (see *SM* for the derivation). The critical insight comes from the edge prior $p(D \mid \Delta I_T, \Delta I_R)$. Since both modalities capture the same geometric boundaries, the edges of the restored thermal image $(I_T + D)$ should spatially align with the RGB edges ΔI_R. This "edge consistency" prior can be modeled as:

$$p(D \mid \Delta I_T, \Delta I_R) \propto \exp\left(-\frac{|\Delta(I_T + D) - \Delta I_R|^2}{2(\sigma')^2}\right), \tag{10}$$

which reveals that the estimation of D requires access to both the thermal and RGB edge structures. This motivates the design of our Thermal Conduction Network (TCN), which takes the rendered thermal image I_T, its Laplacian edge map ΔI_T, and the aligned RGB edge map ΔI_R as inputs:

$$D = I_C - I_T = f_{TCN}(I_T, \Delta I_T, \Delta I_R). \tag{11}$$

This enables the network to implicitly learn the edge consistency constraint during end-to-end optimization. The recovered sharp thermal image is then obtained as $I_C = I_T + D$.

4.6 NETWORKS VS. EXPLICIT LOSSES.

Alternatively, the consistency priors in Equations 7 and 10 could be imposed as explicit L_2 losses. We instead encode them through learned networks for two reasons: (1) *spatial adaptivity*—the constraint strength is modulated per location, relaxing where strict cross-modal alignment does not hold (e.g., semi-transparent materials) while applying stronger correction where thermal diffusion is severe; and (2) *joint optimization in the native parameter space*—(β, l) are optimized in their coupled nonlinear space dictated by the Beer-Lambert law, and D directly in image space, rather than rigidly constraining depths or second-order derivatives. This design choice is empirically validated in our ablation study, where the explicit losses degrade thermal reconstruction while our networks improve it.

4.7 LOSS FUNCTIONS

Combining the two physical corrections, the final thermal prediction is $\hat{I}_T = I_T\, e^{-\beta l} + D$ (Figure 3). The overall objective comprises the thermal reconstruction loss $\mathcal{L}_{\mathrm{T}}$, computed between $\hat{I}_T$ and the ground-truth thermal images, and the RGB reconstruction loss $\mathcal{L}_{\mathrm{RGB}}$, computed between the rendered I_R and the ground-truth RGB images. For each modality $m \in \{\mathrm{RGB}, \mathrm{T}\}$, the objective function $\mathcal{L}_{(m)}$ incorporates both a pixel-wise $\mathcal{L}_1$ loss to measure intensity discrepancies and a $\mathcal{L}_{\text{D-SSIM}}$ loss to capture perceptual similarity differences:

$$\mathcal{L}_{(m)} = (1 - \lambda)\mathcal{L}_1 + \lambda \mathcal{L}_{\text{D-SSIM}}, \quad m \in \{\mathrm{RGB}, \mathrm{T}\}, \tag{12}$$

where the hyperparameter λ is determined empirically. The total objective function $\mathcal{L}$ is defined as the sum of the modality-specific losses:

$$\mathcal{L} = \mathcal{L}_{\mathrm{RGB}} + \mathcal{L}_{\mathrm{T}}. \tag{13}$$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate on the 10 scenes of *RGBT-Scenes* Lu et al. (2025) ($\sim$100 paired views per scene, with $\sim$10 held out) and the low-light *Dark Scene* Lu et al. (2026). We additionally evaluate on *ThermoScenes* Hassan et al. (2025), comprising 22 scenes and 2,942 paired views.

Baselines include ThermoNeRF Hassan et al. (2025), ThermalGaussian Lu et al. (2025) (OMMG), and four single-modality methods trained separately on RGB and thermal: NeRF Mildenhall et al. (2021), 3DGS Kerbl et al. (2023), Pixel-GS Zhang et al. (2024), and GSDF Yu et al. (2024). All methods use their official configurations with matched training budgets of 30K iterations at 640×480 resolution on a single NVIDIA RTX 3090 GPU. We report PSNR↑, SSIM↑, and LPIPS↓ for both modalities; implementation and metric details are provided in the *SM*.

5.2 3D RECONSTRUCTION IN THERMAL VIEW

Quantitative Analysis. We evaluate the thermal reconstruction performance of our method, PRT-3D, against two baselines: ThermoNeRF (TNF) and ThermalGaussian (TGS). As reported in Table 1, PRT-3D consistently outperforms both baselines across nearly all scenes. In terms of PSNR, our method achieves a remarkable average of 28.91 dB, surpassing TGS (25.19 dB) and TNF (20.28 dB) by significant margins of 3.72 dB and 8.63 dB, respectively. This substantial improvement highlights the effectiveness of our physics-inspired cross-modal interaction in accurately modeling thermal radiance. Similar improvements are observed in SSIM (0.936 vs. 0.877 / 0.756) and LPIPS (0.174 vs. 0.181 / 0.309), indicating superior structural preservation and perceptual quality. Notably, in the challenging *Dark Scene* that lacks sufficient visual features, our method still maintains the highest accuracy (25.12 dB), demonstrating robust performance even in adverse lighting conditions.

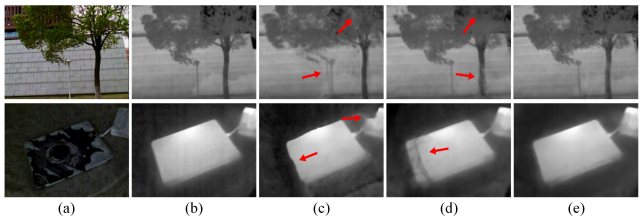

Figure 4: Visual comparisons of different methods in thermal view. (a) Ground truth RGB images. (b) Ground truth thermal images. (c) ThermoNerf. (d) ThermalGaussian. (e) Ours. Top row: Bright scene. Bottom row: Dark scene. See *SM* for more examples.

Qualitative Analysis. Visual comparisons in Figure 4 further validate our advantages. Baseline methods frequently suffer from blurring and thermal blooming artifacts. For instance, in the bottom row scene, strong radiation causes severe boundary ambiguity in TGS and TNF (indicated by red arrows). In contrast, our method produces sharp, clean edges, benefiting from the TCN's ability to leverage RGB edge consistency to compensate for heat diffusion. Furthermore, in the top row scene, our approach successfully recovers distant details that are washed out in baseline results, verifying the effectiveness of the LTN in compensating for atmospheric attenuation.

5.3 3D RECONSTRUCTION IN RGB VIEW

Quantitative Analysis. A key advantage of our method is that the physics-inspired modules significantly boost thermal reconstruction without compromising the already strong RGB performance achieved by TGS. As shown in Table 1, PRT-3D maintains RGB reconstruction quality fully comparable to TGS, achieving an average PSNR of 24.94 dB (vs. TGS 24.78 dB), SSIM of 0.852 (vs. 0.850), and identical LPIPS of 0.210. Both methods substantially outperform TNF (18.49 dB / 0.571 / 0.333), which suffers from degraded RGB quality after incorporating the thermal modality. In contrast, our physics-inspired modules enforce cross-modal consistency (depth consistency via the LTN and edge consistency via the TCN), which strengthens the shared geometric structure underlying both modalities rather than altering modality-specific appearance attributes. As a result, the thermal modality benefits directly from the physics-informed compensation, while the RGB modality is preserved or even slightly improved through the enhanced geometric foundation. Notably, in the *Dark*

Scene, where visual feature extraction is difficult, our method achieves 30.00 dB, slightly surpassing TGS (29.21 dB), suggesting that cross-modal synergy can even benefit RGB reconstruction under adverse conditions.

Qualitative Analysis. Figure 5 presents a visual comparison of RGB novel view synthesis. Our method effectively eliminates artifacts and preserves fine structural details compared to the baselines. In the top row scene, TNF suffers from severe geometric distortions, and TGS exhibits blurring in the background structures (highlighted by the red arrow). In contrast, our method recovers sharp details by leveraging the shared geometry refined through cross-modal interaction. Similarly, in the bottom row scene, TNF produces significant floating artifacts due to geometric ambiguity. Our PRT-3D suppresses these artifacts completely, producing a clean and photo-realistic rendering that closely matches the Ground Truth.

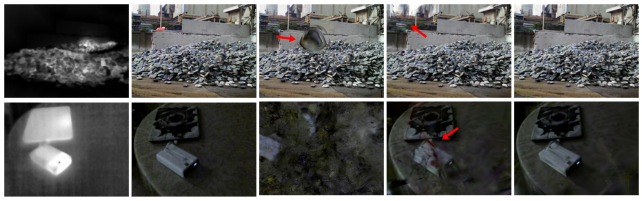

Figure 5: Visual comparisons of different methods in RGB view. (a) Ground truth thermal images. (b) Ground truth RGB images. (c) ThermoNerf. (d) ThermalGaussian. (e) Ours. Top row: Bright scene. Bottom row: Dark scene. See *SM* for more examples.

5.4 Analysis of Speed and Memory

PRT-3D demonstrates superior parameter efficiency (Table 2), consuming only 111 MB of memory—significantly lower than TGS (138 MB) and TNF (232 MB). While the inclusion of physics modules increases training time compared to TGS, it translates to a substantial quality leap (e.g., +3.72 dB in Thermal PSNR). Notably, the extra cost is incurred mainly during training; at inference, the full model still supports real-time rendering (24.81 FPS). The results highlight the flexibility of our framework, which offers a controllable trade-off between real-time rendering capabilities and state-of-the-art physical reconstruction fidelity.

Table 2: Training time, rendering speed, and memory usage.

Method	Time	FPS	Memory
TNF	3736s	8.03	232MB
TGS	806s	37.22	138MB
PRT-3D (None)	635s	47.24	109MB
PRT-3D (LTN)	916s	32.75	116MB
PRT-3D (TCN)	1099s	27.30	115MB
PRT-3D (LTN+TCN)	1209s	24.81	111MB

5.5 Ablation Study

To verify the effectiveness of our proposed LTN and TCN modules, we conducted an ablation study. The results shown in *SM* indicate that both modules contribute individually to performance gains, while their combination yields the best results. These results confirm that modeling physics-inspired cross-modal consistency (depth consistency via LTN and edge consistency via TCN) serves as an effective geometric regularizer, enhancing the reconstruction quality of both modalities. Per-scene detailed analysis is provided in the *SM*.

5.6 Explicit Loss vs. Network-Based Constraints

A natural alternative to LTN and TCN is to directly add the consistency constraints derived from our MAP analysis as explicit losses: $\mathcal{L}_{dist} = \|l - l'\|^2$ for depth consistency (cf. Equation 7) and $\mathcal{L}_{edge} = \|\Delta(I_T + D) - \Delta I_R\|^2$ for edge consistency (cf. Equation 10). As shown in *SM*, this approach not only fails to improve thermal reconstruction over the baseline, but degrades it by 0.24 dB when both losses are combined. In contrast, LTN+TCN consistently improves thermal PSNR over the baseline and clearly outperforms the explicit-loss counterparts. This confirms that the network-based design is essential: explicit losses impose spatially uniform, rigid constraints that create optimization conflicts (particularly when RGB estimates are unreliable during early training), while the networks can adaptively modulate constraint strength and jointly optimize coupled physical parameters in their nonlinear space.

5.7 Cross-Modal Consistency Verification

To verify that PRT-3D resolves the cross-modal inconsistencies identified in Section *Preliminary Analysis*, we measure the L1 discrepancy between RGB and thermal outputs over depth maps and Laplacian edge maps, and report the relative reduction of PRT-3D over TGS (details in *SM*). Grouping scenes by their dominant physical characteristics, we find the improvement concentrates exactly where the corresponding phenomenon dominates: in *open scenes* with distant objects (e.g., LandScape), where atmospheric attenuation is most severe, PRT-3D reduces the depth discrepancy by **51.0%**, confirming that LTN compensates radiation decay at long range; in *high-contrast scenes* with strong thermal boundaries (e.g., Rotary Kiln), where heat diffusion is most pronounced, it reduces the edge discrepancy by **18.3%**, validating that the TCN sharpens diffusion-blurred boundaries via RGB edge structure. This closes the loop with our preliminary analysis: the inconsistencies that motivated our design are precisely the ones our physics-inspired modules eliminate.

5.8 Comparison with Single-Modality Methods

To isolate the value of cross-modal interaction, we compare PRT-3D with the single-modality baselines introduced above, trained under identical settings (*SM*). As Table 3 shows, PRT-3D achieves the best result on five of the six metric–modality combinations and is on par on the sixth (thermal LPIPS, 0.174 vs. 0.172): on thermal it surpasses the strongest specialist, GSDF, by +0.33 dB, while on RGB it outperforms all specialists (+0.65 dB over the strongest, Pixel-GS), with +1.27/+1.14 dB (thermal/RGB) gains over the best specialists in the low-light *Dark Scene*. A single RGB-T model thus matches or surpasses every single-modality specialist on its own modality. This is the physical correlation between the two modalities at work—the very coupling our physics-inspired interaction is built to model: information flows to where it is scarce, lifting the information-poor thermal modality while turning the interaction into a net gain, rather than a cost, for the information-rich RGB.

Table 3: Comparison with single-modality methods (average over all scenes). Best in **bold**, second best underlined.

Method	Thermal			RGB		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF (T)	22.38	0.854	0.217	–	–	–
3DGS (T)	26.03	0.913	0.237	–	–	–
Pixel-GS (T)	26.65	0.909	0.210	–	–	–
GSDF (T)	28.58	0.932	**0.172**	–	–	–
NeRF (RGB)	–	–	–	19.52	0.649	0.307
3DGS (RGB)	–	–	–	23.80	0.833	0.217
Pixel-GS (RGB)	–	–	–	24.29	0.833	0.220
GSDF (RGB)	–	–	–	24.19	0.815	0.240
PRT-3D (Ours)	**28.91**	**0.936**	0.174	**24.94**	**0.852**	**0.210**

5.9 Cross-Dataset Generalization

To assess transferability, we further evaluate PRT-3D on ThermoScenes Hassan et al. (2025), the dataset introduced by ThermoNeRF and progressively extended to 22 scenes covering building facades, and everyday objects, with temperatures from $-16.2\,^{\circ}$C to $87.3\,^{\circ}$C—a substantial domain shift from RGBT-Scenes in scene scale, sensor, and thermal dynamics. Using identical hyperparameters without any per-dataset tuning, PRT-3D outperforms both RGB-T baselines on all metrics of both modalities over the full 22 scenes (thermal: 32.33 dB, +0.69 dB over TGS and +3.59 dB over TNF; RGB: 22.64 dB, +0.26 dB and +4.09 dB) See *SM* for more details. Notably, these margins hold on TNF's home dataset, where it is far stronger than on RGBT-Scenes.

6 Conclusion

In this work, we propose PRT-3D, a physics-inspired RGB-T 3DGS reconstruction framework. By drawing inspiration from atmospheric attenuation and heat conduction, we design two cross-modal interaction networks (LTN and TCN) that leverage depth consistency and edge consistency between RGB and thermal modalities to enhance thermal reconstruction through end-to-end optimization. Experiments demonstrate that PRT-3D significantly improves thermal reconstruction quality over state-of-the-art methods while maintaining comparable RGB quality and reducing memory consumption.

AI USE STATEMENT

The conception, experimental design and execution, and writing of this paper were carried out entirely by the human authors. Generative AI tools were used only for grammar checking and language polishing of author-written text. We take full responsibility for the final content of this work.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5855–5864, 2021.

Spencer Carmichael, Manohar Bhat, Mani Ramanagopal, Austin Buchan, Ram Vasudevan, and Katherine A Skinner. Trnerf: Restoring blurry, rolling shutter, and noisy thermal images with neural radiance fields. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7980–7990. IEEE, 2025.

Mariam Hassan, Florent Forest, Olga Fink, and Malcolm Mielle. Thermonerf: A multimodal neural radiance field for joint rgb-thermal novel view synthesis of building facades. *Advanced Engineering Informatics*, 65:103345, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Moses Kayanda Kiteto and Cleophas Achisa Mecha. Insight into the bouguer-beer-lambert law: a review. *Sustainable Chemical Engineering*, pp. 567–587, 2024.

Yuxiang Liu, Xi Chen, Shen Yan, Zeyu Cui, Huaxin Xiao, Yu Liu, and Maojun Zhang. Thermalgs: Dynamic 3d thermal reconstruction with gaussian splatting. *Remote Sensing*, 17(2):335, 2025.

Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. Thermalgaussian: Thermal 3d gaussian splatting. In *ICLR*, 2025.

Rongfeng Lu, Hangyu Chen, Zunjie Zhu, Yuhang Qin, Ming Lu, Le Zhang, Chenggang Yan, and Anke Xue. Rgbt-scenes-extended. `https://drive.google.com/drive/folders/1A6kdIjDe7kw-iKQkzjHNw0wgk_3V7hcp`, 2026. Accessed: 2026-01-28.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10349–10358, 2024.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Stephen Vidas, Peyman Moghadam, and Michael Bosse. 3d thermal mapping of building interiors using an rgb-d and thermal camera. In *2013 IEEE international conference on robotics and automation*, pp. 2311–2318. IEEE, 2013.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021.

David Vernon Widder. *The heat equation*, volume 67. Academic Press, 1976.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20310–20320, 2024.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.

Tianxiang Ye, Qi Wu, Junyuan Deng, Guoqing Liu, Liu Liu, Songpengcheng Xia, Liang Pang, Wenxian Yu, and Ling Pei. Thermal-nerf: Neural radiance fields from an infrared camera. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1046–1053. IEEE, 2024.

Mulin Yu, Tao Lu, Linning Xu, Lihan Jiang, Yuanbo Xiangli, and Bo Dai. Gsdf: 3dgs meets sdf for improved neural rendering and reconstruction. *Advances in Neural Information Processing Systems*, 37:129507–129530, 2024.

Zheng Zhang, Wenbo Hu, Yixing Lao, Tong He, and Hengshuang Zhao. Pixel-gs: Density control with pixel-aware gradient for 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 326–342. Springer, 2024.