THERMAL3R: PHYSICS-INDUCED FEED-FORWARD FRAMEWORK FOR THERMAL 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Thermal infrared (TIR) imaging enables reliable all-weather perception, yet generalizable 3D reconstruction from TIR remains underdeveloped. Existing feed-forward multi-view models designed for the visible spectrum fail in the thermal domain due to three physical hurdles: depth stretching from nonlinear emission, texture loss from heat conduction, and radiative distortion from atmospheric attenuation. To foster research in thermal 3D reconstruction, we introduce Largethermal3D, a comprehensive and fully-annotated benchmark for thermal 3D reconstruction, and Thermal3R, a physics-guided framework integrating radiation, conduction, and transmission priors. Specifically, we propose: (i) Physics-Aware Radiation Sensitivity Attention Mechanism (PAM) to handle nonlinear sensor response; (ii) Thermal Gradient Consistency (TGC) to sharpen geometric boundaries attenuated by heat conduction; and (iii) Attenuation-Aware Calibration (AAC) to correct atmospheric biases. Experiments show that Thermal3R achieves competitive performance, outperforming fine-tuned RGB baselines across all tasks. This underscores the superiority of physical priors over simple data-driven fine-tuning. The code is available at `https://anonymous.4open.science/r/Thermal3R-172F1re21`.

1 INTRODUCTION

3D reconstruction is undergoing a paradigm shift from traditional per-scene optimization (Huang et al., 2024; Barron et al., 2022) to generalizable feed-forward models like DUSt3R (Wang et al., 2024) and VGGT (Wang et al., 2025). While these models achieve remarkable zero-shot performance on visible (RGB) imagery, they are largely limited to the visible spectrum. Thermal infrared (TIR) imaging provides critical all-weather, illumination-invariant perception for nighttime operations and adverse environments (Liu et al., 2025; Zou et al., 2023; Liang et al., 2024). However, naively transferring RGB-based 3D pipelines to the thermal domain is non-trivial due to a significant modality gap governed by distinct physical laws.

The core challenge lies in three physical hurdles unique to TIR imaging that lead to severe geometric artifacts: (i) depth stretching (i.e., where the 3D scene is unnaturally elongated along the axial scale with an excessively large predicted depth span) arising from nonlinear radiance emission, which causes non-uniform sensor sensitivity; (ii) texture loss arising from heat conduction, which smooths local temperature variations and blurs boundaries; and (iii) radiative distortion arising from atmospheric attenuation, which introduces depth-dependent radiance decay (see Fig. 1). Existing thermal-specific works are either restricted to 2D tasks or rely on time-consuming per-scene optimization (Chen et al., 2024; 2025; Zhao et al., 2025), leaving generalizable, feed-forward thermal 3D reconstruction an underdeveloped field.

In this work, we address these challenges by incorporating thermal-imaging physics into geometric backbones. We first establish a standardized foundation by curating Largethermal3D, a comprehensive and fully-annotated benchmark for thermal 3D perception, providing unified camera calibration and dense geometry annotations for 90 diverse scenes. Building upon this, we propose Thermal3R, a physics-guided feed-forward framework that integrates radiation, conduction, and transmission priors. Specifically, Thermal3R incorporates three physics-induced modules: Physics-Aware Radiation Sensitivity Attention Mechanism (PAM) to handle nonlinear sensor response and rectify depth stretching, Thermal Gradient Consistency (TGC) to recover sharp boundaries from texture loss, and

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.

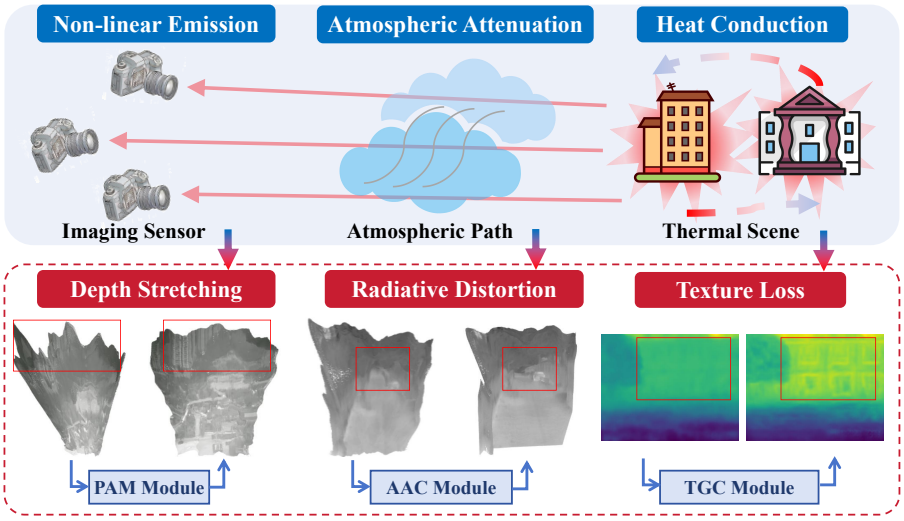


Figure 1: Physical hurdles in thermal infrared imaging and the corresponding geometric artifacts. We categorize physical hurdles into three origins: *Scene-induced Conduction*, *Path-induced Attenuation*, and *Sensor-induced Response*. These hurdles cause characteristic artifacts: texture loss, radiative distortion, and depth stretching, which are addressed by TGC, AAC, and PAM modules.

Attenuation-Aware Calibration (AAC) to correct atmospheric biases and radiative distortion. Extensive experiments demonstrate that Thermal3R achieves state-of-the-art performance, suggesting that physical priors can be more effective than straightforward modality transfer. Our main contributions are summarized as follows:

(1) We present Largethermal3D, the first comprehensive benchmark for thermal 3D reconstruction, featuring 90 high-quality scenes with unified calibration and dense geometry across indoor, outdoor, and UAV environments.

(2) We propose Thermal3R, a feed-forward framework that embeds thermal radiation, conduction, and transmission priors into a multi-view transformer to resolve TIR-specific geometric artifacts.

(3) We demonstrate that our approach consistently outperforms fine-tuned RGB baselines across monocular depth, point map, and camera pose estimation tasks, realizing a robust, physically-grounded understanding of thermal scenes.

2 Related Work

2.1 3D Reconstruction and Multi-View Transformers

Classical 3D reconstruction pipelines are built on structure-from-motion (SfM) (Schönberger & Frahm, 2016; Özyeşil et al., 2017; Agarwal et al., 2011; Moulon et al., 2013) and multi-view stereo (MVS) (Furukawa et al., 2015; Galliani et al., 2015; Gu et al., 2020), as exemplified by Bundler (Snavely et al., 2006), VisualSfM (Wu, 2013), and COLMAP (Schönberger & Frahm, 2016). While these methods recover camera poses and point clouds through image correspondences and geometric optimization, their dependence on hand-crafted features often limits robustness in complex scenes. Learning-based MVS approaches (Yao et al., 2018; Chen et al., 2019; Gu et al., 2020; Yan et al., 2020) address this by replacing hand-crafted priors with convolutional networks that infer depth across multiple views. However, they still suffer from long inference times. Recently, several large feed-forward 3D models (Shen et al., 2025; Deng et al., 2025; Jia et al., 2025) have emerged to achieve fast inference. VGGT (Wang et al., 2025) employs an alternating attention mechanism to enforce cross-view geometric constraints and utilizes a multi-head prediction module for parameter estimation. MapAnything (Keetha et al., 2026) proposes a unified transformer that regresses a factored multi-view scene representation to solve a broad range of 3D tasks in a single forward pass. But all these models operate on visible-spectrum RGB images and implicitly

assume reflectance-based image formation. They do not account for the nonlinear thermal emission, heat conduction, or atmospheric attenuation that govern thermal infrared (TIR) imaging. To our knowledge, our work is the first to adapt such a large feed-forward multi-view backbone to thermal imagery and to integrate thermal-domain physics directly into its training objective.

2.2 THERMAL 3D PERCEPTION

Motivated by recent advances in 3D reconstruction techniques such as NeRF (Mildenhall et al., 2021) and 3DGS (Kerbl et al., 2023), several works have adapted them to thermal scenes for achieving high-precision point cloud reconstruction: Thermal-NeRF (Ye et al., 2024) extends NeRF to operate solely on infrared images using thermal mapping and structural constraints for degraded scenes; ThermalNeRF (Lin et al., 2024) uses a multispectral radiance field calibrated between RGB and LWIR cameras; and on the Gaussian splatting side, ThermalGaussian (Lu et al., 2024) learns multimodal 3D Gaussians from calibrated RGB–thermal pairs with thermal-specific smoothing, Thermal3D-GS (Chen et al., 2024) introduces physics-induced 3D Gaussians by learning atmospheric transmission and thermal conduction fields. Although these approaches achieve strong results on thermal novel view synthesis, they rely on implicit or per-scene optimized scene representations. In addition, inference from thermal video sequences incurs a high computational cost. In contrast, we adapt a large multi-view transformer (VGGT) to thermal imagery and embed key parts of the thermal imaging physics directly into its architecture and loss. Additionally, we curate a comprehensive and fully-annotated benchmark for 3D reconstruction (Largethermal3D) from thermal infrared imagery, which comprises 90 distinct scenes.

3 METHOD

3.1 MOTIVATION & OVERALL FRAMEWORK

Thermal 3D reconstruction is primarily hindered by the modality gap between RGB reflectance and TIR physics. As illustrated in Fig. 1, we attribute the dominant TIR artifacts to three sources along the imaging chain: (i) **sensor-induced response nonlinearity** (yielding depth stretching); (ii) **scene-induced heat conduction** (yielding texture loss and blurred boundaries); (iii) **path-induced atmospheric attenuation** (yielding radiative distortion). Building on the VGGT backbone (Wang et al., 2025), we propose Thermal3R, a physics-induced feed-forward framework that injects these priors at different stages of the forward pass (Fig. 2). Concretely, Thermal3R integrates three modules with complementary roles: (i) **PAM** (see Fig. 2 top middle) intervenes at the representation level by compensating non-uniform radiometric sensitivity of thermal sensors, so that cold regions are not suppressed in attention and matching; (ii) **TGC** (see Fig. 2, middle right) intervenes at the geometric structure level by using thermal gradients as boundary cues to sharpen depth transitions that are softened by heat diffusion; (iii) **AAC** (see Fig. 2, bottom right) intervenes at the calibration level by explicitly modeling attenuation as a latent per-pixel field and enforcing a Beer–Lambert bottleneck to correct depth-dependent bias. This design moves beyond naive modality transfer and yields a physically consistent 3D reconstruction in a single forward pass.

3.2 PAM: PHYSICS-AWARE RADIATION SENSITIVITY ATTENTION MECHANISM

We first address **sensor-induced nonlinearity**, since it directly suppresses radiometric gradients and undermines cross-view matching; this failure then propagates to geometry as severe **depth stretching**. Uncooled microbolometer sensors exhibit a temperature-dependent sensitivity curve where the derivative of the output signal with respect to temperature decreases in low-temperature regions. As a result, for a fixed physical edge (constant ∇T), the observed gradient becomes $\nabla I = S(T) \cdot \nabla T$, which can be extremely small in background areas. Standard attention and matching mechanisms therefore under-utilize these regions, leading to ambiguous correspondence and stretched depth.

Thermal infrared (TIR) sensors measure band-integrated radiance $L_\lambda(T, \lambda)$ as a function of surface temperature T. The relationship is governed by Planck's law (Marr & Wilkin, 2012). For long-wave infrared (LWIR) at ambient temperatures, the spectral radiance can be approximated by Wien's law:

$$L_\lambda(T, \lambda) \approx C(\lambda) \exp\left(-\frac{a(\lambda)}{T}\right). \tag{1}$$

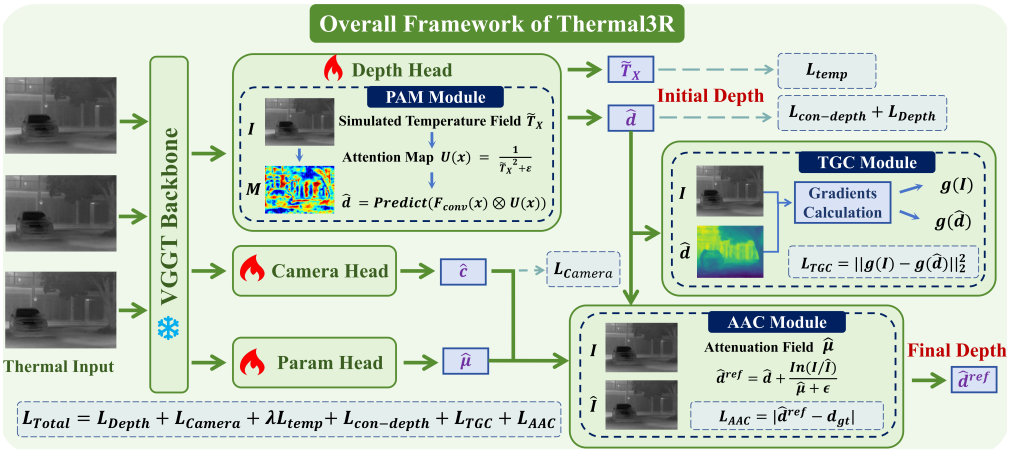


Figure 2: Overall framework of Thermal3R. Built upon the VGGT backbone (Wang et al., 2025), our framework employs three specialized heads to jointly estimate depth $\hat{d}$, camera parameters $\hat{c}$, and the attenuation field $\hat{\mu}$. By integrating the PAM, TGC, and AAC modules, the model ensures that the prediction results consistently adhere to infrared physical characteristics. Here, M represents the depth confidence heat map, the value of parameter λ is 0.3.

Where $C(\lambda)$ and $a(\lambda)$ are wavelength-dependent constants. The **absolute radiometric sensitivity**, defined as the change in radiance per unit temperature change ($S_{abs}(T) = \frac{\partial L}{\partial T}$), determines the strength of geometric cues (gradients) in the image. Differentiating Eq. equation 1 yields:

$$S_{abs}(T) = \frac{\partial L_\lambda(T,\lambda)}{\partial T} \propto \frac{1}{T^2} \exp\left(-\frac{a(\lambda)}{T}\right). \tag{2}$$

Crucially, due to the exponential term $\exp(-\frac{a(\lambda)}{T})$, $S_{abs}(T)$ **decreases rapidly** as temperature T drops. This leads to **radiometric gradient compression**: in cold/background regions, the observed image gradients ∇I become vanishingly small, causing standard attention mechanisms to overlook these areas. To compensate for this signal suppression, we need a weighting factor $W_{prior}(T)$ that is inversely proportional to the sensitivity. Considering the effective power-law approximation commonly used in thermal radiometry (Howell et al., 2010) around $T \approx 300$ K, the absolute radiometric sensitivity can be modeled as $S_{\text{abs}}(T) \propto T^{3.8}$. Inspired by this physical formulation, we explore the performance relationship between the inverse-sensitivity prior $W_{prior}(T)$ and different powers of T. It is found that $W_{\text{prior}}(T) = \frac{1}{T^2}$ achieves the best trade-off between numerical stability and performance (The detailed correlation analysis of the inverse-sensitivity prior with temperature T is provided in **Appendix A.7**). Meanwhile, considering the exponential term hinders effective modeling of the temperature T in the model, we take the logarithmic partial derivative with respect to T on both sides of Eq. equation 1, yielding:

$$\frac{\partial \ln L_\lambda(T,\lambda)}{\partial T} \approx \frac{a(\lambda)}{T^2}. \tag{3}$$

To instantiate this physical compensation, PAM estimates a pixel-wise sensitivity map and uses its inverse to re-weight features within the VGGT backbone. We employ a lightweight prediction head (Conv+ReLU) on intermediate features to estimate a simulated temperature map $\tilde{T}(x)$ at the input resolution. Based on $\tilde{T}(x)$, we construct the compensation weight map $U(x)$ as:

$$U(x) = \frac{1}{\tilde{T}(x)^2 + \epsilon}, \tag{4}$$

where ϵ is a small constant for numerical stability. Here, $U(x)$ serves as a proxy for the inverse radiometric sensitivity $W_{prior}(\tilde{T})$. Finally, to make the weight map compatible with deep features, we process $U(x)$ through a convolutional layer to obtain a modulation map $S(x)$. This map is injected into the depth head via residual multiplicative attention:

$$F_{\text{enhanced}}(x) = F_{\text{conv}}(x) \cdot (1 + \beta \cdot S(x)), \tag{5}$$

where $F_{\text{conv}}(x)$ is the input feature to the depth head, and β is a learnable scalar. This operation amplifies feature gradients in low-sensitivity (cold) regions, ensuring that the attention mechanism attends to background geometric structures as strongly as foreground ones. The visual analysis of the physical validity of the PAM module is provided in **Appendix A.4**.

To ensure the predicted temperature field $\tilde{T}(x)$ accurately reflects the physical properties required for correct sensitivity weighting, we introduce a consistency regularization loss. Since the exact sensor response curve and spatially varying emissivity are unknown, we do not enforce strict radiometric equality. Instead, we encourage the *structure* of the predicted temperature gradients to align with the observed image gradients, scaled by the physical sensitivity prior. Specifically, we minimize the discrepancy between the observed log-intensity gradients and the theory-driven temperature gradients modulated by the inverse squared temperature (a proxy for sensitivity drop-off):

$$\mathcal{L}_{\text{temp}} = \sum_{x} \rho \left\| \nabla \ln I_{\text{input}}(x) - \frac{\alpha}{\tilde{T}(x)^2 + \epsilon} \nabla \tilde{T}(x) \right\|_1, \tag{6}$$

where ∇ denotes the spatial gradient operator, α is a scaling factor, and $\rho(\cdot)$ is the Huber loss function robust to outliers. This loss ensures that the simulated temperature field $\tilde{T}$ respects the underlying non-linear physical laws governing thermal imaging. By enforcing this consistency, we prevent the PAM module from learning trivial or noisy attention maps that lack physical plausibility. Additionally, to maintain geometric accuracy, we retain the scale-invariant relative depth loss:

$$\mathcal{L}_{\text{con-depth}} = \mathbb{E}_x \left\| \frac{\hat{d}(x) - d_{\text{gt}}(x)}{d_{\text{gt}}(x) + \epsilon} \right\|_2, \tag{7}$$

where $\hat{d}(x)$ represents the predicted depth value and $d_{\text{gt}}(x)$ represents the ground-truth depth value. The total loss for the PAM module is:

$$L_{\text{PAM}} = L_{\text{con-depth}} + 0.3 \cdot L_{\text{temp}}. \tag{8}$$

3.3 TGC: Thermal Gradient Consistency

The second hurdle is the texture loss and boundary blurring caused by **scene-induced heat conduction**. In TIR imaging, heat diffusion smooths local temperature variations, attenuating sharp discontinuities at object interfaces (Bergman, 2011). While these boundaries appear "softer" in Fig. 1 than their visible counterparts, they remain intrinsically coupled with the scene's structural transitions. To leverage this coupling and restore sharp boundaries, we introduce **Thermal Gradient Consistency (TGC)**. This module (see Fig. 2, middle right) utilizes thermal gradients as a structural guide to regularize depth estimation, effectively counteracting the blurring effects of heat conduction.

From a physics perspective, thermal gradients in a scene arise from both geometric discontinuities and material property variations. While emissivity changes can induce thermal edges on flat surfaces, geometric boundaries typically generate the most significant and structurally consistent thermal gradients. To exploit this correlation while mitigating the risk of 'phantom edges' from pure material changes, TGC is designed to recover boundaries attenuated by heat conduction. Furthermore, since Thermal3R is built on a multi-view transformer, it inherently distinguishes between consistent geometric edges and view-dependent thermal textures, ensuring that the gradient consistency primarily reinforces true geometric structures. By maintaining gradient consistency, we use the thermal image as a structural guide to regularize depth estimation: (1) Edge Sharpening: Guiding the network to concentrate geometric transitions where thermal gradients are reliable, and (2) Artifact Suppression: Penalizing depth discontinuities in thermally smooth regions to eliminate phantom structures.

Let $\hat{d}(\mathbf{x})$ be the predicted depth map and $I(\mathbf{x})$ be the input thermal image. We compute the spatial gradients $\mathbf{g}_{\hat{d}(\mathbf{x})} = \nabla \hat{d}(\mathbf{x})$ and $\mathbf{g}_I(\mathbf{x}) = \nabla I(\mathbf{x})$ using finite difference operators. To focus on edge alignment rather than intensity scale, which may vary due to material emissivity, we normalize the gradients:

$$\tilde{\mathbf{g}}_{\hat{d}(\mathbf{x})} = \frac{\mathbf{g}_{\hat{d}(\mathbf{x})}}{\|\mathbf{g}_{\hat{d}(\mathbf{x})}\|_2 + \epsilon}, \qquad \tilde{\mathbf{g}}_I(\mathbf{x}) = \frac{\mathbf{g}_I(\mathbf{x})}{\|\mathbf{g}_I(\mathbf{x})\|_2 + \epsilon}, \tag{9}$$

The TGC loss is then defined as the mean squared error between the normalized gradient fields:

$$\mathcal{L}_{\text{TGC}} = \mathbb{E}_{\mathbf{x}}\Big[\big\|\tilde{\mathbf{g}}_{\hat{d}(\mathbf{x})} - \tilde{\mathbf{g}}_I(\mathbf{x})\big\|_2^2\Big]. \tag{10}$$

By focusing on normalized gradients, $\mathcal{L}_{\text{TGC}}$ encourages the depth map to inherit the topology of the thermal scene while remaining robust to global intensity fluctuations.

3.4 AAC: Attenuation-Aware Calibration

The third hurdle in thermal 3D reconstruction is the radiative distortion caused by **path-induced atmospheric attenuation**. As thermal radiance travels from the surface to the sensor, it is absorbed and scattered by atmospheric constituents, creating a depth-dependent bias in the observed intensity as visualized in Fig. 1. To address this, we introduce **Attenuation-Aware Calibration (AAC)**, a physics-based head integrated into our framework (see Fig. 2, bottom right). The AAC head predicts a per-pixel attenuation field $\hat{\mu}$ to explicitly model the radiative transfer process, thereby calibrating initial depth $\hat{d}$ into a physically consistent estimate $\hat{d}^{\text{ref}}$, which also serves as the final depth prediction output. The visual analysis of physical validity of $\hat{\mu}$ is provided in **Appendix A.6**.

The propagation of thermal radiance through a medium is commonly modeled by the Beer–Lambert law (Howell et al., 2010). For a ray with path length d_i, the transmitted radiance I_i relates to the initial surface emission I_{0i} as:

$$I_i = \tau_i\, I_{0i} = I_{0i}\, e^{-\mu_i d_i}, \tag{11}$$

where $\mu_i = \mu_{\text{abs},i} + \mu_{\text{sca},i}$ is the total attenuation coefficient (sum of absorption and scattering). By taking the logarithm, the true depth d_i can be expressed as:

$$d_i = \frac{\ln(I_i/I_{0i})}{-\mu_i}. \tag{12}$$

Equation equation 12 reveals a fundamental coupling: any discrepancy between the observed radiance and the ideal surface emission, if not accounted for by the attenuation field μ_i, will manifest as a bias in the estimated depth. In practice, the measured radiance I_i' often deviates from the ideal I_i due to atmospheric fluctuations or unmodeled sensor responses. If an initial depth $\hat{d}_i$ is predicted without accounting for these variations, the resulting depth bias can be derived by subtracting the logarithmically transformed radiance terms:

$$d_i - \hat{d}_i = \frac{\ln(I_i/I_{0i})}{-\mu_i} - \frac{\ln(I_i'/I_{0i})}{-\mu_i} = \frac{\ln(I_i/I_i')}{-\mu_i}. \tag{13}$$

This yields the physics-based correction formula:

$$d_i = \hat{d}_i + \frac{\ln(I_i'/I_i)}{\mu_i}. \tag{14}$$

Intuitively, if the observed intensity I_i' is weaker than the predicted ideal radiance I_i, the depth must be increased to account for the additional attenuation along the propagation path. A key challenge in modeling atmospheric effects is that the attenuation coefficient μ_i is spatially and temporally varying and, crucially, lacks ground-truth annotations for direct supervision. To address this, we design the Param head (see Fig. 2) to predict a per-pixel attenuation field $\hat{\mu}_i$ from the shared transformer features, treating it as a *latent atmospheric field* optimized through physical consistency. This per-pixel design choice allows the model to account for non-uniform atmospheric conditions, such as local fog patches or humidity gradients. Building on Eq. equation 14, the refined depth $\hat{d}_i^{\text{ref}}$ is computed by integrating the predicted field $\hat{\mu}_i$ with the initial depth $\hat{d}_i$:

$$\hat{d}_i^{\text{ref}} = \hat{d}_i + \frac{\ln(I_i'/\hat{I}_i)}{\hat{\mu}_i + \epsilon}, \tag{15}$$

where I_i' is the observed thermal intensity and $\hat{I}_i$ is the predicted radiance (It is obtained by fusing the point clouds of the same object from different camera viewpoints within a training batch and projecting them onto the current viewpoint.). To optimize this process, we supervise the refined depth against the ground truth d_{gt} via a regression loss:

$$\mathcal{L}_{\text{AAC}} = \mathbb{E}_i\big[\big\|\hat{d}_i^{\text{ref}} - d_{gt,i}\big\|_1\big]. \tag{16}$$

Although $\hat{\mu}_i$ is not explicitly supervised, Eq. equation 15 acts as a *physical bottleneck*. During training, gradients from $\mathcal{L}_{\text{AAC}}$ flow back through the radiative transfer equation, forcing $\hat{\mu}_i$ to converge to values that physically reconcile the discrepancy between observed radiance and estimated geometry.

3.5 TOTAL LOSS FUNCTION

Our final objective augments the original VGGT supervision (depth and camera) with three physics-induced terms, each targeting one dominant TIR artifact:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{PAM}} + \mathcal{L}_{\text{TGC}} + \mathcal{L}_{\text{AAC}}. \tag{17}$$

Where $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{camera}}$ denote depth prediction loss and camera pose prediction loss, as employed in VGGT. Specifically, $\mathcal{L}_{\text{PAM}}$ compensates sensor nonlinearity to mitigate depth stretching, $\mathcal{L}_{\text{TGC}}$ enforces edge-aware structural consistency against conduction-induced blur, and $\mathcal{L}_{\text{AAC}}$ corrects depth-dependent bias due to atmospheric attenuation, yielding a physically consistent reconstruction.

4 EXPERIMENTS

We introduce Largethermal3D datasets (Sec. 4.1) and evaluate on three tasks: monocular depth estimation (Sec. 4.2), point map estimation (Sec. 4.3) and camera pose estimation (Sec. 4.4). In Section 4.5, we conduct an ablation analysis.

4.1 DATASETS

Largethermal3D: Considering that current mainstream thermal 3D scene datasets (Chen et al., 2024; Lee et al., 2022; Dai et al., 2021; Li et al., 2024; Liu et al., 2025) suffer from limited scene categories and a small scale, which also lack critical camera parameters or ground-truth depth maps, hindering their direct utility for supervised model training. To address these limitations, we introduce the first comprehensive benchmark for thermal 3D reconstruction, Largethermal3D, which comprises 90 thermal 3D reconstruction scenes. Each scene in Largethermal3D contains thermal images, intrinsic camera parameters, extrinsic camera parameters and depth maps. Additionally, to facilitate subsequent research, we have provided pixel-level annotations for the objects of interest within several scenes. The detailed description of Largethermal3D dataset can be found in **Appendix A.8**. During training, we utilize 76 scenes for training and the remaining 14 scenes for testing. Furthermore, we compare Thermal3R with current mainstream models, including Dust3r (Wang et al., 2024), Mapanything (Keetha et al., 2026), Pi3, VGGT (Wang et al., 2025), and VGGT* (which denotes the fine-tuned version of VGGT on Largethermal3D dataset).

Table 1: Monocular Depth Estimation based on laser scanning scenes.

Method	Avg_MAE↓	Avg_Abs_Rel↓	Avg_Sq_Rel↓
Dust3r	0.7783	0.8102	0.7735
Mapanything	0.7469	0.7596	0.7098
Pi3	0.7359	0.7294	0.6834
VGGT	0.7382	0.7419	0.6895
VGGT*	<u>0.7217</u>	<u>0.6618</u>	<u>0.5852</u>
Thermal3R(Ours)	**0.7059**	**0.5324**	**0.4286**

Table 2: Monocular Depth Estimation based on computational photography scenes.

Method	Avg_MAE↓	Avg_Abs_Rel↓	Avg_Sq_Rel↓
Dust3r	0.4036	0.3279	0.3109
Mapanything	0.3528	0.2663	0.2711
Pi3	0.4423	0.2870	0.3306
VGGT	0.3707	0.2607	0.2292
VGGT*	<u>0.3035</u>	<u>0.2286</u>	<u>0.2184</u>
Thermal3R(Ours)	**0.2724**	**0.1924**	**0.1496**

4.2 MONOCULAR DEPTH ESTIMATION

We conduct experiments by dividing the test scenarios into two groups based on the type of depth ground truth acquisition: one using depth ground truth obtained via laser scanning, and the other using depth ground truth derived from computational photography. Following prior work (Keetha et al., 2026; Li et al., 2024), we report the MAE, Abs_Rel, and Sq_Rel metrics to evaluate the depth prediction accuracy. As reported in Table 1 and Table 2, our method achieves competitive performance results among all the 3D perception methods. Notably, our method achieves a substantial accuracy improvement when compared to the fine-tuned VGGT model (VGGT*). On the Avg_sq_rel metric, Thermal3R achieves improvements of 26.8% and 31.5% over the second-best method, respectively. Visualization of the depth estimation results from Fig. 3 shows that while models such as Dust3r and Mapanything can effectively distinguish objects in the foreground, distant objects are

often assigned similar depth values. In contrast, the Thermal3R model is capable of effectively differentiating depth variations among distant objects. Compared with the VGGT* results, Thermal3R not only reconstructs the detailed structures of distant objects more accurately but also maintains better overall consistency in depth prediction for objects at the same depth level.

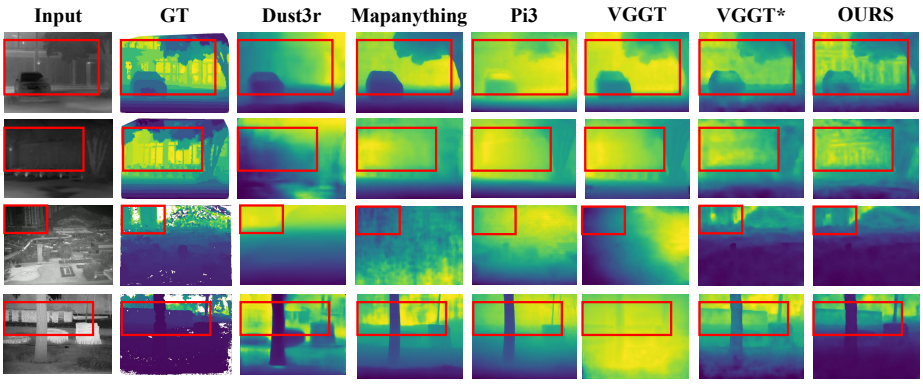


Figure 3: Comparative Visualization for Monocular Depth Estimation

4.3 Point Map Estimation

We also compare the accuracy of our predicted point cloud to other 3D perception methods on Largethermal3D dataset. For each scene, we uniformly sample 6-10 frames. Following VGGT, the predicted point cloud is aligned to the ground-truth point cloud using Umeyama algorithm (Umeyama, 1991). We use Accuracy, Completeness, and Overall for point map estimation. As shown in Table 3, our method achieves the highest predictive accuracy on metrics such as Accuracy and Overall, indicating that the point clouds predicted by Thermal3R are closer to the ground truth while also demonstrating superior completeness in point cloud reconstruction. As observed in the comparative point cloud prediction results in Fig. 4, compared to models like MapAnything and Pi3, Thermal3R demonstrates superior performance in holistic and coherent scene prediction across the first three relatively simple scenarios, with its results exhibiting minimal white void artifacts. While maintaining predictive coherence, Thermal3R also better preserves scene features and structures, avoiding issues such as collapsed building predictions in scene of the first row or distorted platform predictions in scene of the second row. In the fourth more complex scenario, models like Dust3R and Pi3 incorrectly predict the scene as a single flat plane with zero depth. Although VGGT* effectively distinguishes between foreground and background, it exhibits noticeable stretching, leading to distorted scene predictions. In contrast, Thermal3R not only distinguishes foreground and background effectively but also significantly mitigates the stretching effect, owing to PAM and TGC module.

Table 3: Point Map Estimation.

Method	Acc↓	Comp↓	Overall↓
Dust3r	2.1569	2.5086	2.3328
Mapanything	1.9564	<u>2.1212</u>	2.0388
Pi3	1.9175	2.7988	2.3581
VGGT	1.7135	2.7277	2.2206
VGGT*	<u>1.6113</u>	2.3816	<u>1.9964</u>
Thermal3R(Ours)	**1.4697**	**2.0459**	**1.7578**

Table 4: Camera Pose Estimation.

Method	RRA@3↑	RTA@3↑	RRA@5↑	RTA@5↑
Dust3r	13.33	8.67	21.67	17.00
Mapanything	16.67	10.65	26.30	19.63
Pi3	29.82	18.80	32.59	35.83
VGGT	46.67	21.11	68.06	38.24
VGGT*	<u>51.48</u>	<u>22.69</u>	<u>68.80</u>	<u>39.44</u>
Thermal3R(Ours)	**57.28**	**25.27**	**71.67**	**41.85**

4.4 Camera Pose Estimation

We evaluate Thermal3R on the Largethermal3D dataset for camera pose estimation, as shown in Table 4. Following (Wang et al., 2023; 2025), we randomly sampled three groups from each scene. Each group consists of 20 frames, and we calculate the translational and rotational errors for pose prediction using metrics such as RRA@3, RTA@3, RRA@5, and RTA@5. The experimental results demonstrate that the Thermal3R method achieves improvements of 5.8%, 2.6%, 2.9%, and 2.4%

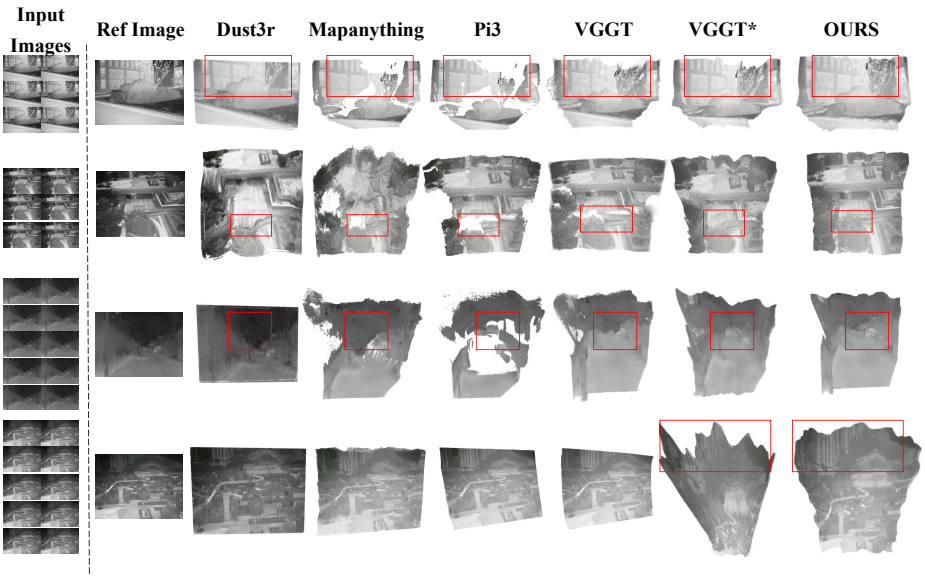


Figure 4: Comparative Visualization for Point Map Estimation

on the aforementioned four metrics, respectively, compared to the current state-of-the-art methods, indicating that our approach is capable of high-precision pose prediction.

4.5 Ablation Study

We conduct an ablation study on monocular depth estimation and point cloud prediction accuracy using the Largethermal3D dataset, the baseline model is VGGT*. The experimental result is shown in Table 5. which demonstrate that, with the addition of the three physically-driven components proposed in this paper, Thermal3R achieves the highest prediction accuracy across all metrics. The detailed ablation analysis is shown from **Appendix A.3** to **Appendix A.6**

Table 5: Effects of Different Components in Thermal3R.

PAM	TGC	AAC	Monocular Depth Estimation			Point Map Estimation		
			Avg_MAE↓	Avg_Abs_Rel↓	Avg_Sq_Rel↓	Acc↓	Comp↓	Overall↓
			0.7217	0.6618	0.5852	1.6113	2.3816	1.9964
✓			0.7113	0.5661	0.4542	1.4829	2.1591	<u>1.8170</u>
	✓		0.7168	0.5746	0.4620	1.5300	2.1950	1.8625
		✓	0.7157	0.6011	0.4959	1.5289	2.2117	1.8703
✓	✓		0.7093	<u>0.5473</u>	<u>0.4392</u>	<u>1.4761</u>	2.1667	1.8214
✓		✓	<u>0.7082</u>	0.5598	0.4536	1.4794	2.1724	1.8259
	✓	✓	0.7117	0.5626	0.4519	1.5184	<u>2.0523</u>	1.8754
✓	✓	✓	**0.7059**	**0.5324**	**0.4286**	**1.4697**	**2.0459**	**1.7578**

5 Conclusion

In this paper, we addressed the challenging problem of generalizable 3D reconstruction from thermal infrared imagery. By identifying the limitations of RGB-based geometric models in the thermal domain, we introduced Thermal3R, a feed-forward transformer that integrates core principles of thermal physics—emission, conduction, and attenuation—into its learning objective. We also contributed Largethermal3D, a comprehensive benchmark to bridge the data gap in thermal 3D vision. Our experimental results across multiple tasks and diverse scenarios demonstrate that incorporating physical priors allows the model to better resolve the ambiguities inherent in TIR imaging.

AI USE STATEMENT

In this work, we used generative AI tools solely for language refinement, proofreading, and improving the clarity and fluency of the manuscript. We did not use generative AI tools for designing the model framework, formulating the core contributions or technical innovations, designing experiments, analyzing results, drawing framework figures, producing code, or generating scientific claims. Other disclosure-related uses of generative AI are not applicable to this work. All AI-assisted edits were carefully reviewed by the authors to ensure that they preserved the intended technical meaning and did not introduce new claims, results, citations, or artifacts. We take full responsibility for the final content of this work, including all text, claims, figures, experiments, and conclusions.

ETHICS STATEMENT

This work focuses on thermal infrared 3D reconstruction and introduces Largethermal3D, a benchmark built from static thermal scenes with depth maps, camera parameters, and object-level annotations for research purposes. Thermal perception can benefit safety-critical applications such as search-and-rescue in low-light or smoke, all-weather autonomous navigation, and nighttime industrial monitoring. At the same time, thermal imaging may raise privacy and surveillance concerns because it can reveal the presence and spatial layout of people or objects even without visible-light appearance. To mitigate such risks, our dataset construction excludes dynamically moving objects and focuses on static indoor, outdoor, and UAV scenes. We encourage users of the dataset and code to follow the licenses of the original data sources, respect local privacy regulations, and avoid applications that enable unauthorized surveillance or harmful tracking.

REPRODUCIBILITY STATEMENT

We provide the code in an anonymous repository and will release the Largethermal3D benchmark to facilitate reproducibility and future research. The dataset split, evaluation protocols, and implementation details are described in Sec. 4 and Appendix A. In particular, Appendix A.1 reports the main training settings, including the trainable modules, optimizer, learning rate, weight decay, number of epochs, and GPU configuration. We also provide ablation studies and efficiency analyses to support independent verification of the proposed PAM, TGC, and AAC components.

REFERENCES

Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5460–5469. IEEE, 2022.

Theodore L Bergman. *Fundamentals of heat and mass transfer*. John Wiley & Sons, 2011.

Qian Chen, Shihao Shu, and Xiangzhi Bai. Thermal3d-gs: Physics-induced 3d gaussians for thermal infrared novel-view synthesis. In *European Conference on Computer Vision*, pp. 253–269. Springer, 2024.

Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1538–1547. IEEE, 2019.

Si Chen, Liuxiang Qiu, Da-Han Wang, Wentao Zhu, Yang Hua, and Yan Yan. Hierarchical token-aware cross-modality reconstruction for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 2025.

Weichen Dai, Yu Zhang, Shenzhou Chen, Donglei Sun, and Da Kong. A multi-spectral dataset for evaluating motion estimation systems. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5560–5566. IEEE, 2021.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025.

Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE international conference on computer vision*, pp. 873–881, 2015.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2492–2501. IEEE, 2020.

John R Howell, M Pinar Menguc, and Robert Siegel. *Thermal radiation heat transfer*. CRC press, 2010.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

Xiaosong Jia, Yanhao Liu, Yu Hong, Renqiu Xia, Junqi You, Bin Sun, Zhihui Hao, and Junchi Yan. Drivevggt: Calibration-constrained visual geometry transformers for multi-camera autonomous driving. *arXiv preprint arXiv:2511.22264*, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

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

Alex Junho Lee, Younggun Cho, Young-sik Shin, Ayoung Kim, and Hyun Myung. Vivid++: Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022.

Han Li, Yukai Ma, Yuehao Huang, Yaqing Gu, Weihua Xu, Yong Liu, and Xingxing Zuo. Riders: Radar-infrared depth estimation for robust sensing. *IEEE transactions on intelligent transportation systems*, 25(11):18764–18778, 2024.

Yiping Liang, Libing Bai, Lulu Tian, Xu Zhang, and Yong Gao. Thermal parameter reconstruction imaging for interlayer defect detection in ecpt. *IEEE Transactions on Industrial Informatics*, 21 (2):1931–1940, 2024.

Yvette Y Lin, Xin-Yi Pan, Sara Fridovich-Keil, and Gordon Wetzstein. Thermalnerf: Thermal radiance fields. In *2024 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–12. IEEE, 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. *arXiv preprint arXiv:2409.07200*, 2024.

Jonathan M Marr and Francis P Wilkin. A better presentation of planck's radiation law. *American Journal of Physics*, 80(5):399–405, 2012.

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.

Pierre Moulon, Pascal Monasse, and Renaud Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proceedings of the IEEE international conference on computer vision*, pp. 3248–3255, 2013.

Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 26:305–364, 2017.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113. IEEE, 2016.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. In *ACM SIGGRAPH*, pp. 835–846. ACM, 2006.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 1991.

Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9773–9783, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Changchang Wu. Towards linear-time incremental structure from motion. In *International Conference on 3D Vision (3DV)*, pp. 127–134. IEEE, 2013.

Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European conference on computer vision*, pp. 674–689. Springer, 2020.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pp. 767–783. Springer, 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.

Wenda Zhao, Hengshuai Cui, Haipeng Wang, You He, and Huchuan Lu. Freefusion: Infrared and visible image fusion via cross reconstruction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8040–8056, 2025.

Xiongxin Zou, Tangle Peng, and Yimin Zhou. Uav-based human detection with visible-thermal fused yolov5 network. *IEEE Transactions on Industrial Informatics*, 20(3):3814–3823, 2023.

A APPENDIX

A.1 TRAINING DETAILS

During training phase, we freeze the parameters of the DINO module, Global Attention, and Frame Attention components in the original VGGT model, which are used for image feature extraction, and only train the parameters of the Depth Head, Camera Head, and Param Head modules. The training process is specifically carried out on two GeForce RTX 3090 GPUs for 200 epochs. The optimizer selected is AdamW, with a learning rate set to $5e^{-5}$ and a weight decay coefficient set to 0.05.

A.2 EFFECTIVENESS ANALYSIS

To provide a more comprehensive evaluation of different models, we further conduct an effectiveness analysis on the LargeThermal3D dataset. Specifically, we measure the parameter count, inference time, and peak inference memory of the feed-forward methods. All efficiency measurements are performed on a single NVIDIA GeForce RTX 3090 GPU under identical input resolutions and view configurations to ensure a fair comparison. The experimental results Table 6 show that Thermal3R reconstructs 20 views in only 2.16 seconds without requiring test-time optimization, while achieving an inference speed comparable to VGGT. Across all evaluated view configurations, Thermal3R also consistently requires less peak inference memory than VGGT. In addition, it has a smaller overall parameter count, as it removes the unused point and tracking heads retained in VGGT.

Table 6: Model Effectiveness Analysis on the LargeThermal3D Dataset

Metric	DUSt3R	MapAnything	Pi3	VGGT	Thermal3R
Model parameters (M)	571.17	563.34	958.70	1256.67	1190.73
Inference time, 1 view (s)	0.21	0.06	0.08	0.12	0.12
Peak memory, 1 view (GB)	2.58	3.92	5.41	7.23	7.03
Inference time, 5 views (s)	1.93	0.31	0.60	0.61	0.64
Peak memory, 5 views (GB)	4.32	4.66	5.62	9.14	8.36
Inference time, 20 views (s)	34.58	0.92	2.09	1.98	2.16
Peak memory, 20 views (GB)	10.85	9.68	6.34	15.08	11.61

A.3 ABLATION STUDY

We conduct an ablation study on monocular depth estimation and point cloud prediction accuracy using the Largethermal3D dataset, the baseline model is VGGT*. The experimental result is shown in Table 7. The ablation experimental results demonstrate that, with the addition of the three physically-driven components proposed in this paper, Thermal3R achieves the highest prediction accuracy across all metrics. Observing the experimental results in the Monocular Depth Estimation section, it can be seen that incorporating both the PAM and TGC modules leads to a significant enhancement in depth prediction accuracy. This demonstrates their capability to effectively differentiate between objects at varying depth levels while maintaining the distinct physical architecture of each module. Observing the experimental results in the Point Map Estimation, following the integration of the PAM module, point cloud prediction accuracy is notably improved. The PAM module enhances the model's attention to distant objects and improves the prediction of their gradient variation rate. This approach effectively reduces the stretching artifacts in the reconstruction of distant structures such as buildings and forests, thereby improving the overall point cloud prediction accuracy of the model.

A.4 VISUAL ANALYSIS OF THE PAM MODULE

For the PAM module, we conduct a visual ablation study on both monocular depth estimation and point cloud prediction. Fig. 5 presents the results on monocular depth estimation. To highlight the ability of PAM to enhance the model's performance in distant, low-temperature regions, we visualize not only the predicted depth maps but also the corresponding depth-confidence heatmaps. The results show that, with PAM integrated into the model, the heatmap values in distant low-temperature areas are significantly increased, indicating that the model allocates greater attention to accurate

Table 7: Effects of Different Components in Thermal3R.

PAM	TGC	AAC	Monocular Depth Estimation			Point Map Estimation		
			Avg_MAE↓	Avg_Abs_Rel↓	Avg_Sq_Rel↓	Acc↓	Comp↓	Overall↓
			0.7217	0.6618	0.5852	1.6113	2.3816	1.9964
✓			0.7113	0.5661	0.4542	1.4829	2.1591	<u>1.8170</u>
	✓		0.7168	0.5746	0.4620	1.5300	2.1950	1.8625
		✓	0.7157	0.6011	0.4959	1.5289	2.2117	1.8703
✓	✓		0.7093	<u>0.5473</u>	<u>0.4392</u>	<u>1.4761</u>	2.1667	1.8214
✓		✓	<u>0.7082</u>	0.5598	0.4536	1.4794	2.1724	1.8259
	✓	✓	0.7117	0.5626	0.4519	1.5184	<u>2.0523</u>	1.8754
✓	✓	✓	**0.7059**	**0.5324**	**0.4286**	**1.4697**	**2.0459**	**1.7578**

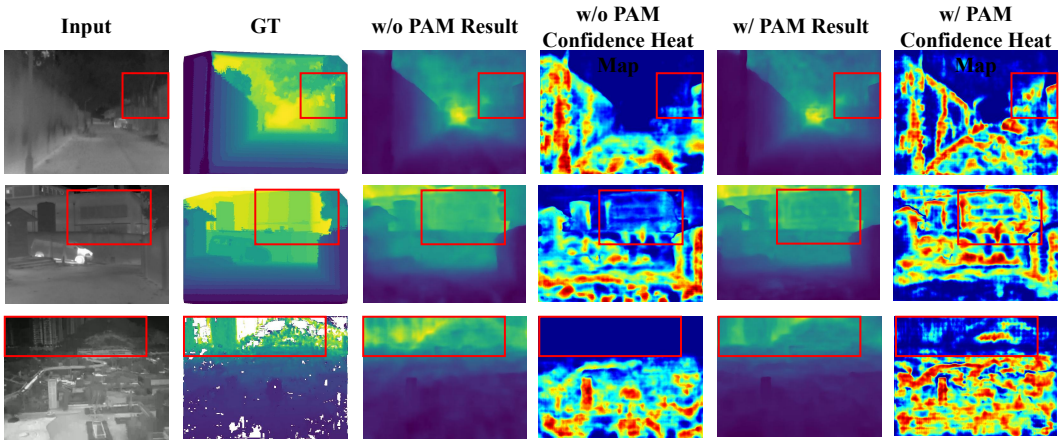


Figure 5: Ablation Visualization on Depth Estimation for PAM Module. For the Heat Map of depth confidence, colors tending toward deep red indicate higher confidence of the model in its depth prediction at that location, while colors tending toward deep blue indicate lower confidence.

predictions in these regions. Meanwhile, the monocular depth estimation results demonstrate that the model can better delineate the contours between different distant buildings after incorporating PAM. Fig. 6 shows the results on point cloud prediction. The first row comparison indicates that, with PAM, the depth stretching artifacts for distant regions are substantially mitigated. The second row further shows that PAM helps to more effectively separate near and far scenes, thereby reducing the "point cloud dragging" effect that previously occurred when predicting different objects at varying depths.

To further validate the physical plausibility of the simulated temperature map $\tilde{T}(x)$ in the PAM module, we select two representative scenes, **scene19** and **garden**. In both scenes, the top 15% of the image is dominated by the relatively low-temperature sky, whereas the bottom 15% mainly corresponds to the relatively high-temperature ground. We therefore compute five quantile statistics of $\tilde{T}(x)$ within the top and bottom 15% regions of each scene, respectively. The experimental results are reported in Table 8, which demonstrates that all sky quantiles are below the ground quantiles, matching the expected ordinal temperature relation.

Table 8: Physical Validity of the Simulated Temperature Map $\tilde{T}(x)$

Scene	Top 15% (sky)	Bottom 15% (ground)
scene19	27.45, 27.83, 28.04, 28.31, 28.97	31.95, 32.96, 33.27, 33.46, 33.66
garden	20.68, 21.55, 22.96, 24.23, 25.33	35.83, 36.80, 37.49, 38.71, 39.58

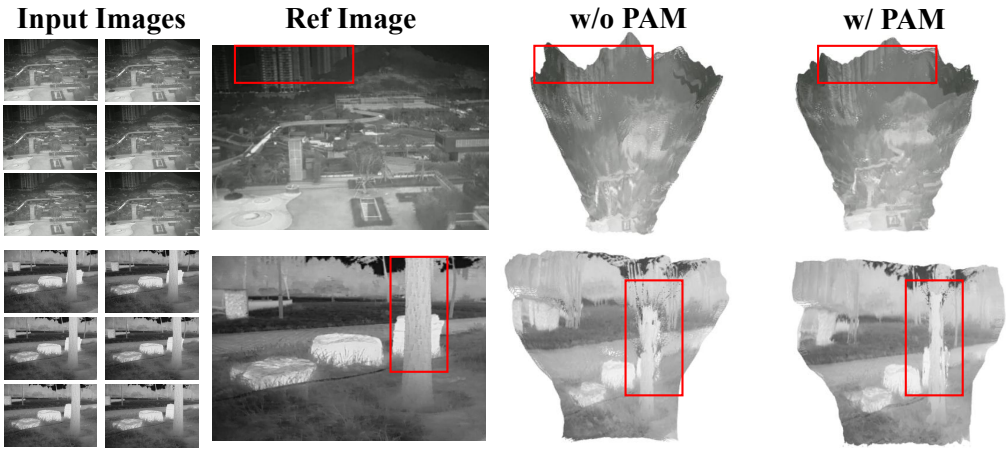


Figure 6: Ablation Visualization on Point Map for PAM Module.

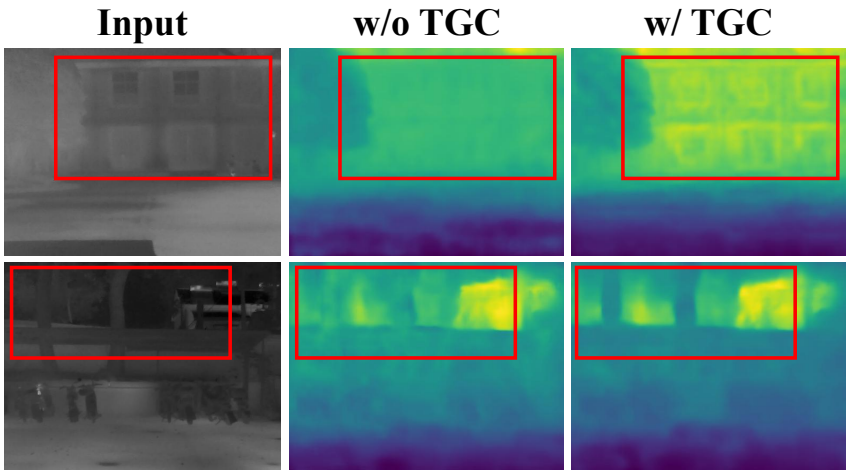


Figure 7: Ablation Visualization for TGC Module

A.5 Visual Analysis of the TGC Module

The ablation comparison results for the TGC module in Fig. 7 demonstrate that, incorporating the TGC module significantly enhances the reconstruction accuracy of detailed textures and object contours. For instance, in the first image row, the addition of the TGC module enables better recovery of the detailed architectural structures of the buildings in the background. In the second comparative image row, the integration of the TGC module yields clearer predictions of tree contours.

A.6 Visual Analysis of the AAC Module

For the AAC module, we conduct visual ablation analysis on both monocular depth estimation and point cloud prediction. Fig. 8 presents the experimental results on monocular depth estimation. The results show that incorporating the AAC module reduces depth prediction errors for objects such as trees and buildings in the scene. In the monocular depth estimation outputs, this improvement is reflected as more complete object boundaries and clearer contours. Meanwhile, we separately computed the error distribution curves for the Indoor Scene, Outdoor Scene, and UAV Scene, as shown in Fig. 9. Experimental results demonstrate that after incorporating the AAC module, the error distribution is significantly reduced in the Outdoor Scene and UAV Scene. This is because atmospheric effects are more pronounced in these scenes. The addition of the AAC module effectively mitigates the impact of atmospheric attenuation. This also indicates that the proposed AAC module

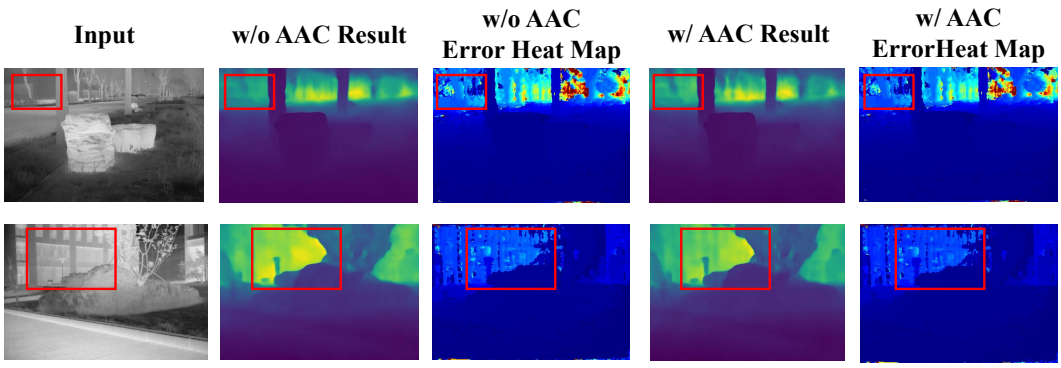


Figure 8: Ablation Visualization for AAC Module. For the Error Heat Map, colors tending toward deep blue indicate smaller errors, while colors tending toward deep red indicate larger errors.

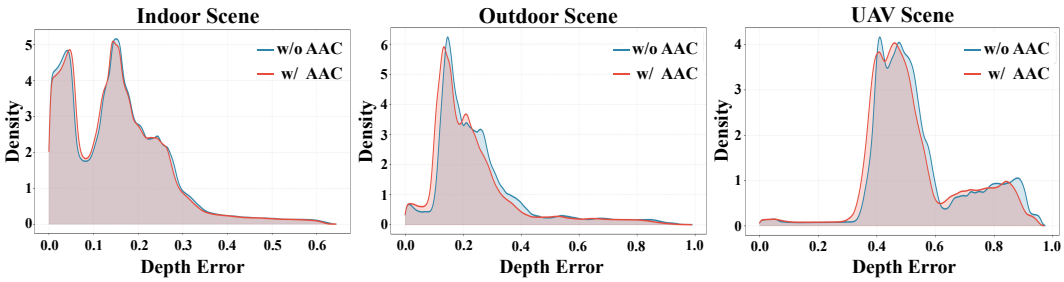


Figure 9: Error distribution curves for the ablation of the AAC module in Indoor Scene, Outdoor Scene, and UAV Scene.

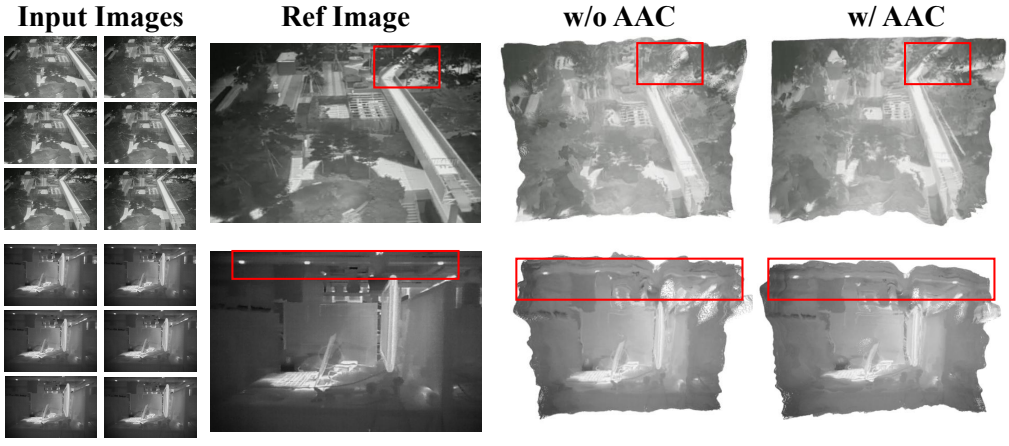


Figure 10: Ablation Visualization on Point Map for AAC Module

conforms to the physical characteristics of infrared radiation, enabling more significant accuracy improvements for the Outdoor Scene and UAV Scene.

Fig. 10 presents the point cloud ablation analysis for the AAC module. The experimental results demonstrate that incorporating the AAC module significantly improves the reconstruction accuracy of the point cloud. For instance, it effectively reconstructs and distinguishes between the road and the tree trunks above it. Additionally, for the beams in indoor scenes, it reduces the distortion issues present during the reconstruction process.

Furthermore, we analyze the physical validity of the parameter $\hat{\mu}$ in the AAC module. Because direct atmospheric attenuation measurements are unavailable, we performed a controlled Beer–Lambert

Table 9: Physical Validity of $\hat{\mu}$ in AAC module

Applied attenuation	scene1 (outdoor)	drone1 (UAV)	laptop (indoor)
0.00	0.2982	0.3931	0.2687
0.01	0.3061	0.3952	0.2777
0.02	0.3133	0.4049	0.2823
0.05	0.3221	0.4128	0.2868
0.08	0.3314	0.4223	0.2968

intervention:

$$I_\mu(x) = I_0(x) \exp\left[-\mu_{\text{gt}} d_{\text{gt}}(x)\right]. \tag{18}$$

Ground-truth depth was used only to synthesize test inputs and was not provided during inference. We evaluated representative outdoor (scene1), UAV (drone1), and indoor (laptop) scenes in Table 9

The predicted $\hat{\mu}$ increases monotonically with imposed attenuation in all three scene types, providing proxy evidence that it responds to radiative attenuation rather than merely absorbing arbitrary depth residuals. This does not establish $\hat{\mu}$ as an absolutely calibrated atmospheric coefficient. We describe it as an effective attenuation field learned under the Beer–Lambert constraint.

A.7 Temperature correlation analysis

We analyze the relationship between the inverse-sensitivity prior $W_{\text{prior}}(T)$ and the temperature T. Table 10 and Table 11 present the monocular depth estimation results in laser scanning scenarios and computational photography scenes, respectively, obtained under different orders of the selected temperature T.

Although in Section 3.2 we derive from physical principles that the mathematical relationship satisfies $W_{\text{prior}}(T) \propto 1/T^{3.8}$, in deep learning an excessively high exponent often leads to instability. Our experimental results also indicate that using $1/T^4$ does not yield high prediction accuracy. When using $1/T$, the model's attention to low-temperature regions can be enhanced to some extent, but the improvement in accuracy remains limited. In contrast, $1/T^3$ achieves an accuracy level comparable to $1/T^2$, yet it still suffers from instability issues associated with higher-order exponents, resulting in degraded performance on multiple metrics compared with $1/T^2$. Therefore, considering both stability and physical plausibility, we finally adopt $W_{\text{prior}}(T) = 1/T^2$.

Table 10: Analysis of Monocular Depth Estimation for T in Laser Scanning Scenarios.

Method	Avg_MAE↓	Avg_Abs_Rel↓	Avg_Sq_Rel↓
$1/T$	0.7115	0.5553	0.4483
$1/T^3$	0.7075	0.5451	**0.4228**
$1/T^4$	0.7157	0.5736	0.4631
$1/T^2$(Ours)	**0.7059**	**0.5324**	0.4286

Table 11: Analysis of Depth Estimation for Temperature T in computational photography scenes.

Method	Avg_MAE↓	Avg_Abs_Rel↓	Avg_Sq_Rel↓
$1/T$	0.2887	0.1998	0.1506
$1/T^3$	0.2823	**0.1912**	0.1541
$1/T^4$	0.2994	0.2033	0.1676
$1/T^2$(Ours)	**0.2724**	0.1924	**0.1496**

A.8 LargeThermal3D Dataset

Currently available open-source thermal imaging datasets are generally constrained by a limited number of sequences and a narrow range of captured scenarios, rendering them inadequate for large-scale scene training. For instance, the ZJU-Multispectral dataset employs a scanning-based methodology for street scene acquisition. However, due to the movement of vehicles and pedestrians during recording, it contains a significant proportion of dynamic scenes, which complicates its direct application for training models such as Thermal3R. In contrast, while the Thermal3DGS dataset encompasses three categories of scenarios—indoor, outdoor, and unmanned aerial vehicle (UAV) perspectives—the data volume for each category remains limited, thereby hindering the effective improvement of model robustness. Furthermore, all existing open-source thermal 3D scene

datasets share a critical shortcoming: the lack of comprehensive annotations for both depth maps and camera parameters. This deficiency substantially restricts their utility in training generalizable 3D perception models.

Therefore, we introduce the first comprehensive benchmark for thermal 3D reconstruction, Largethermal3D. The construction of Largethermal3D comprises two primary stages: scene collection, followed by scene modeling.

scene collection: To construct a large-scale 3D perception dataset, we collect and organize current mainstream datasets while supplementing them with additional scene captures to enhance scenario diversity. The specific criteria for scene selection are as follows: firstly, the video sequences must not contain dynamically moving objects; secondly, the captured sequences should exhibit rich viewpoint variations with clear image quality; and finally, we aim to balance the number of indoor, outdoor, and UAV scenarios as much as possible. The composition of scenarios per dataset in the final Largethermal3D dataset is summarized in Table 12.

Table 12: The composition of the Largethermal3D dataset

Dataset	Thermal3DGS	ThermalGaussian	ThermalGS	ThermalNerf	Multi-spectral	Vivid++	ZJU-Multispectrum	Ours	Total
Scene Number	19	10	10	1	5	1	16	28	90

scene modeling: After collecting the thermal scene sequences, it is necessary to acquire the corresponding depth maps, camera intrinsic parameters, camera extrinsic parameters, and mask images of the objects of interest for each thermal scene video sequence. Among these, the mask images of the interest objects are obtained by annotating each frame using the LabelMe software in several scenes. For the acquisition of depth maps and camera parameters, we refer to the production pipeline of the CO3D dataset and employ computational photography techniques.

For training Thermal3R, we partition the LargeThermal3D dataset into training and test sets. The detailed distribution of scene categories and the number of scenes in each split are presented in Table 13.

Table 13: Training and Test Split of the LargeThermal3D Dataset

Split	Indoor	Outdoor	UAV	Total scenes	Total images
Training	19	44	13	76	18670
Testing	3	9	2	14	3600
Total	22	53	15	90	22,270

Moreover, LargeThermal3D is, to the best of our knowledge, the largest infrared scene dataset currently available for 3D reconstruction. We therefore compare its number of scenes with those of existing mainstream infrared 3D reconstruction datasets, as shown in Table 14.

Table 14: Scene Count Comparison of Themal 3D Reconstruction Datasets

Dataset	RHD	TSDN	NTR	TI-NSD	ThermalGaussian	Vivid++	Largethermal3D
Number of scenes	10	5	4	20	10	5	90

Considering that depth maps obtained through computational photography may be subject to interference from noise, we conduct separate evaluations on depth values derived from laser scanning and those derived from computational photography within the depth estimation experimental module. The experimental results demonstrate that Thermal3R achieves higher accuracy improvement in the laser scanning-based scenario, thereby fully validating the effectiveness of the Thermal3R method.

A.9 Positive societal impacts and Limitation

This work presents Largethermal3D, the first large-scale benchmark, and Thermal3R, a physics-induced framework for generalizable thermal 3D reconstruction. By bridging the gap between in-

frared physics and deep learning, our research promotes the development of robust perception systems capable of operating in low-light and adverse weather conditions. These advancements have direct implications for safety-critical applications, including search-and-rescue operations in smoke or darkness, all-weather autonomous navigation, and nighttime industrial monitoring. Furthermore, the public release of our dataset and code lowers the barrier for future research in multi-spectral geometric vision, fostering more reliable and interpretable AI for real-world deployment.

However, our current work primarily focuses on 3D perception for thermal infrared static scenes. In future work, we will extend our research to feed-forward 3D perception models for thermal infrared dynamic scenes.