

Supercharging Thermal Gaussian Splatting with Depth Estimation

Manoj Biswanath^{1*}, Chenxin Cai^{2*}, Hannah Schieber³, Daniel Roth³, Benjamin Busam¹

¹ Photogrammetry and Remote Sensing, Munich Center for Machine Learning (MCML),
Technical University of Munich, Munich, Germany - (manoj.biswanath, b.busam)@tum.de

² Technical University of Munich, Munich, Germany - chenxin.cai@tum.de

³ Human-Centered Computing and Extended Reality Lab, TUM University Hospital,
Clinic for Orthopedics and Sports Orthopedics, Munich Institute of Robotics and Machine Intelligence (MIRMI),
Technical University of Munich, Munich, Germany – (hannah.schieber, daniel.roth)@tum.de

Keywords: Thermal imaging, Gaussian Splatting, Depth Estimation, Radiance fields, Novel view synthesis.

Abstract

Efficient and robust 3D scene representation is crucial in autonomous driving, robotics, and related fields. While RGB images provide valuable content for 3D reconstruction, other modalities like thermal or depth can enable additional information on the environment. Lately, novel view synthesis methods like 3D Gaussian Splatting have started using multiple modalities to further boost their performance. But fusing or combining multimodal data can make the process slower and can bring in additional challenges. Therefore, our project aims to use single modality based on thermal infrared domain, by removing the reliance on visible light as much as possible. This single modality can be expected to be faster as it does not rely on multimodal data. We propose a method, **Thermal-to-Depth Gaussian Splatting (TDg)**, that uses only thermal images and depth estimation in its architecture to derive the radiance fields. Our TDg method outperforms the MSMG (Multiple Single-Modal Gaussians) baseline in most cases on our test datasets, RGBT-Scenes and ThermalMix. On average, the rendering quality metrics such as learned perceptual image patch similarity (LPIPS), structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR) of TDg are 1.12%, 0.034%, and 0.01% better than the baseline MSMG values. It also reduces the training time significantly, by 12 mins 47 secs (55% improvement). Overall, our method is successful in deriving these thermal radiance fields, which can ultimately have several applications, such as identifying heat sources critical in surveillance, search or rescue operations, and industrial inspections where temperature is widely used to monitor machines.

1. Introduction

Robotics (Ji et al., 2024; Li and Pathak, 2024), autonomous driving (Yan et al., 2024), virtual reality (Schieber et al., 2025; Kleinbeck et al., 2026), construction industry (Rüter et al., 2024) and similar fields demand highly efficient and robust 3D scene representation. In recent years, explicit rendering techniques, particularly novel view synthesis (NVS) methods such as Gaussian Splatting (GS) (Kerbl et al., 2023), received significant attention due to their exceptional rendering speed and visual quality (Li et al., 2025). Mainstream methods heavily rely on RGB images, whose performance deteriorates significantly in visually degraded environments, such as low-light, fog, smoke, or extreme weather conditions. This can make the direct application of RGB-designed GS frameworks highly challenging. Introducing other modalities alongside RGB, such as thermal, depth, or LiDAR data, can provide additional information about the environment, or can especially help in cases where drones and ground robots are often equipped with additional sensors, or only contain thermal and depth cameras, rather than RGB.

Infrared imaging cameras, which detect the inherent thermal radiation of objects, can provide a robust perception capability, independent of visible light. It is suitable in challenging environments, regardless of lighting and weather conditions. Thermal imaging converts temperature information into interpretable images. One common application of thermography is energy inspection of objects. Classically, this has been done

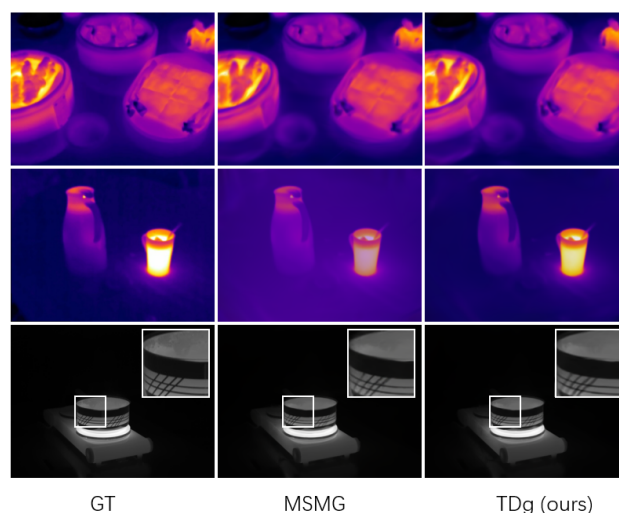


Figure 1. Thermal images rendered from the GS model of our TDg method, compared against Ground Truth (GT) and baseline (MSMG).

by visually evaluating the images. But a direct 3D spatial reference is not established for the measured values. This deficiency becomes obvious for complex structures, when their images taken from different angles need to be fused, or the measured thermal radiations need to be stored in an object-related manner, for further processing. In such cases, creation of a thermal radiance field (as a 3D representation) can be help-

* These authors contributed equally to this work.

ful. It can provide us a 3D spatial reference of the temperature measurements. For example, one specific application could be energy inspection in buildings, where knowledge of both the inside and outside wall temperatures would be necessary to provide insights about the geometry and material properties of the walls. This knowledge could be acquired from the 3D representations. Further, these continuous scene representations can provide correct view-dependent effects such as reflections, translucency, etc., and can serve as either digital twins or an enrichment of already available digital twins.

Novel views are required to get the correct heat signatures of objects, avoiding parallax errors. Thermal patterns of objects can be very complex. For example, heat leakages from building walls mostly look like blobs, elongated structures, or irregular shapes (Krawczyk et al., 2015). To obtain a more accurate picture of these complex patterns, many different viewpoints would be necessary to have the ability to rotate, view, and inspect them. Novel views can provide realistic images from viewpoints that were never actually captured, because taking photos from all angles is not feasible. Also, continuous viewpoint changes can be useful to study space/time-related heat transfer behavior. And predict heat radiation from new viewpoints.

Recent studies have attempted to integrate thermal imaging into the GS framework by using multiple modalities. This can further boost their performance by providing accurate information, enabling an exploration of the 3D environment, and providing photorealistic results. But fusing or combining these multimodal data can make the process slower and can bring in additional challenges. Leveraging a single modality eliminates this dependency, enables faster convergence, and simplifies the processing pipeline. Thermal and RGB cameras typically require precise co-registration or alignment. Thermal images are inherently characterized by low contrast, sparse texture, and non-uniform brightness distribution. Current approaches still fundamentally rely on paired RGB images for supervision or joint optimization, failing to establish a truly independent and purely thermal-based Gaussian representation system. Since RGB images can enhance the quality of thermal models because of their rich geometry and semantic information, removing the reliance on RGB while still maintaining good quality is challenging and demands an intelligent architecture.

Therefore, to address the challenges of thermal-based GS fundamentally, we propose an enhanced and self-contained Thermal GS framework. We want to build a GS model to synthesize thermal images. Through core innovation, our framework achieves high-quality 3D reconstruction. In summary, we contribute:

1. A thermal-guided depth estimation module integrated within GS
2. A novel purely thermal GS approach, denoted **Thermal-to-Depth Gaussian Splatting (TDg)**¹
3. A strong evaluation on two state-of-the-art benchmark datasets, RGBT-Scenes and ThermalMix.

Figure 1 shows a teaser of our method's results. The thermal images rendered from our resulting GS model, are compared against the baseline method (MSMG) and Ground Truth (GT). As shown in the comparison, our method shows better resolution.

¹ Code available: <https://hannahhaensen.github.io/TDg/>

2. Related Work

Since our approach integrates depth estimation with thermal NVS, it lies at the intersection of geometric scene understanding and multi-modal 3D reconstruction. Consequently, we review prior work on depth estimation methods as well as approaches that extend NVS to multi-modal or cross-spectral settings. Particular emphasis is placed on techniques that leverage complementary sensing modalities to improve geometric and photometric consistency.

2.1 Novel View Synthesis

Early approaches such as GS (Kerbl et al., 2023) and neural radiance fields (NeRF) (Mildenhall et al., 2021) rely on RGB images and camera poses typically provided by COLMAP (Schönberger and Frahm, 2016). While later works improve joint camera optimization (Schischka et al., 2024; Schieber et al., 2024; Bian et al., 2023), COLMAP-based initialization remains common. Recent research extends radiance fields to the thermal domain (Chen et al., 2024; Lu et al., 2024; Carmichael et al., 2025; Hassan et al., 2025), enabling 3D reconstruction beyond the visible spectrum.

The most frequent multi-modal approaches add depth data instead of thermal images. For example, (Chung et al., 2024) leverages dense depth maps aligned with initial COLMAP data to optimize for a more precise geometry of the scene. Boosting GS performance with depth is further investigated by (Xu et al., 2025). They combine depth estimation and GS in a shared architecture. Both show advantages for NVS.

2.2 Thermal Novel View Synthesis

Similar to recent developments in general NVS, both GS (Kerbl et al., 2023) and NeRF (Mildenhall et al., 2021) have also demonstrated strong applicability to thermal NVS. Thermal-NeRF (Ye et al., 2024) leverages infrared imagery to enhance rendering in low-light environments: their method converts 16-bit infrared inputs into 8-bit thermal representations, defines a fixed-size sampling volume, and employs an MLP to regress thermal intensities. While Thermal-NeRF focuses solely on thermal inputs, ThermoNeRF (Hassan et al., 2025) integrates RGB and thermal modalities through two coupled MLPs. In a similar direction, (Özer et al., 2024) proposes a joint training framework for RGB and thermal images, systematically exploring different integration strategies and evaluating reconstruction quality across both modalities. (Carmichael et al., 2025) introduces TRNeRF, a framework designed to restore blurry, rolling-shutter, and noisy thermal imagery; their approach explicitly considers the various types of thermal cameras and the distinct challenges they introduce, such as the motion blur commonly observed in uncooled microbolometers, addressing these issues by using models of rolling-shutter timing, noise having fixed-patterns, and motion blur directly into NeRF.

(Lu et al., 2024) extends GS to the multimodal NVS setting by jointly training on RGB and thermal images simultaneously, rendering Gaussians separately for each modality while optimizing them through a combined loss. In this way, they allow each Gaussian to specialize in rendering its respective modality. They use a shared, multimodal-initialized point cloud and also introduce multimodal regularization based on the number of Gaussians associated with each modality. They introduce

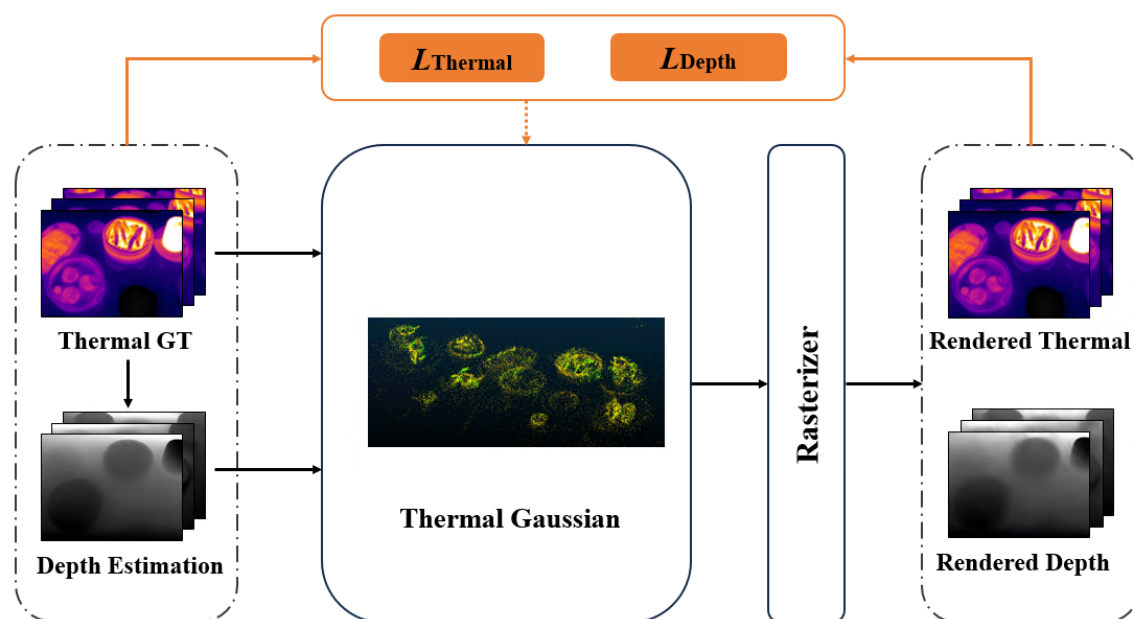


Figure 2. TDg architecture. A unified 3D Gaussian representation (center) is optimized via dual rasterization. By comparing the rendered thermal and depth maps (right) against the input thermal GT and estimated depth priors (left), a joint loss (top) is computed to backpropagate and guide the geometric and photometric training of the Gaussians.

three different training strategies. One of them is MSMG (Multiple Single Modal Gaussians), which we select as the baseline method for comparison of our work. The reason for selecting this method as our baseline is because it closely resembles our proposed method, in the sense that our method replaces the RGB modality with purely thermal images advanced with depth estimation. Additionally, this method demonstrates fairly good results in thermal-GS domain, which is ideal for baseline comparison. (Nam et al., 2025) presents Veta-GS, a thermal GS approach that is aware of scene dynamics. By integrating a Thermal Feature Extractor and a deformation field which is view-dependent, the method captures thermal variations while maintaining robustness, and additionally incorporates the MonoS-SIM loss, which emphasizes frequency, appearance, and edge information to preserve stability. Beyond RGB–thermal fusion, (Chen et al., 2024) proposes a physics-inspired thermal GS framework. Their method explicitly models thermal conduction effects, atmospheric transmission, and incorporates a temperature consistency constraint tailored to infrared signal behavior.

2.3 Depth Estimation

A wide range of approaches has been proposed for estimating depth from RGB images (Aramatzakis et al., 2023; Ke et al., 2024). Recently, depth estimation via diffusion-based techniques has shown particularly strong performance (Ke et al., 2024). They adapt Stable Diffusion to the monocular depth estimation task by fine-tuning its U-Net backbone for depth prediction. (Chung et al., 2024) leverages dense depth maps aligned with initial COLMAP data to optimize for a more precise geometry of the scene.

Thermal Depth Estimation In addition to RGB-based methods, thermal depth estimation provides an alternative line of research. Marigold (Ke et al., 2024) uses thermal images to estimate depth maps. Enhancing thermal images by maximizing self-supervision has been attempted by (Shin et al., 2022).

They demonstrate the importance of temporally consistent normalization and adopt contrast limited adaptive histogram equalization (CLAHE) (Zuiderveld, 1994) to intensify image details locally, by abstaining from noise amplification. (Shin et al., 2023) and (Zuo et al., 2025) investigate depth estimation directly from thermal imagery, demonstrating its potential, especially under difficult lighting conditions or in adverse environments.

Depth Estimation and Novel View Synthesis Depth estimation can be incorporated into NVS pipelines to further enhance reconstruction quality and overall robustness, e.g. (Xu et al., 2025). For improving NVS, depth can be a valuable factor. Traditional GS (Kerbl et al., 2023) allows integration of depth within further code refinement to supervise GS. While often depth-supervised methods focus on the center of Gaussians and ignore the unique Gaussian shape during training, SAD-GS (Kung et al., 2024) directly addresses this problem explicitly, by enhancing GS with depth supervision. Another work specifically optimizing depth and GS is, for example, introduced in DepthSplat (Xu et al., 2025). DepthSplat connects GS with single and multi-view depth estimation, resulting in superior-quality GS. It uses pre-trained monocular depth features. (Roessle et al., 2022) use dense depth priors for deriving the neural radiance fields. They reduce the required number of images significantly by an order of magnitude, to synthesize the novel views.

2.4 Research Gap

We have seen how research has evolved from initial RGB-based NVS to accommodating multiple modalities or cross-spectral domains. Different attempts are seen to integrate thermal into NVS using various fusion techniques. Eventually, some thermal-based GS and NeRF are developed. Additionally, people tried to use depth to supervise NVS. They explored different depth estimation techniques, both from RGB and

thermal modalities. While all these methods tried to integrate multi-modal data, but they rather treated these additional modalities as complementary, mostly for improvement purposes. Eventually, none of them developed a purely single modality (thermal infrared) based GS framework. So, we have attempted to bridge this gap through our work. We also use the depth factor for supervision and the depth is estimated solely from thermal images. Therefore, our attempt is to retain our framework within a single modality.

3. Method

We introduce our enhanced thermal GS framework TDg. We replace the RGB modality, and purely rely on thermal images advanced with depth estimation, purely queried by thermal images, see Figure 2. This method significantly improves the geometric accuracy and radiance consistency of the thermal scene representation.

In our TDg framework, rather than maintaining separate models for different modalities, we utilize a single, unified set of 3D Gaussians to inherently align the thermal radiance and geometry of the scene. Specifically, a set of N Gaussians $\mathcal{G}$ is used to represent the scene:

$$\mathcal{G} = \{\mu_i, \Sigma_i, \alpha_i, c_i\}_{i=1}^N \quad (1)$$

where each Gaussian i is parameterised by its covariance matrix Σ_i , a thermal radiance feature c_i , center position μ_i , and opacity α_i . During optimization, this unified model is rendered into both thermal images (using the radiance feature c_i) and depth maps (using the camera-space Z -coordinate of μ_i). This design allows the pre-estimated depth prior to directly constrain the 3D geometric distribution of the thermal radiance field, avoiding the complexity and feature conflicts of multi-branch architectures.

3.1 Thermal-to-Depth Gaussian Estimation

Building upon the GS rasterization strategy (Kerbl et al., 2023), we remove the reliance on RGB and integrate depth instead. Our approach purely leverages estimated depth from thermal images. This explicitly enforces geometric constraints and addresses the instability of pure thermal supervision in low-texture or low-light conditions.

3.1.1 Thermal-to-Depth Estimation Given a thermal image I_{Th} , a thermal-guided depth estimator f_θ predicts a dense depth map:

$$D = f_\theta(I_{Th}). \quad (2)$$

3.1.2 Depth Gaussian Loss Similar to the color rendering of thermal images, the rendered depth map $\hat{D}$ is obtained via α -blending integration along the ray direction. For a given pixel i , its expected depth $\hat{D}_i$ is accumulated based on the Z -axis depth z_k of the Gaussians in the camera coordinate system:

$$\hat{D}_i = \sum_{k \in \mathcal{N}_i} z_k \alpha_k \prod_{j=1}^{k-1} (1 - \alpha_j), \quad (3)$$

$\mathcal{N}_i$ represents the arranged set of overlapping Gaussians along the ray corresponding to pixel i .

Since the depth predicted by monocular depth estimation networks typically contains scale and shift ambiguities, it cannot be directly compared with the absolute rendered depth. Therefore, we apply Min-Max normalization to both rendered depth $\hat{D}$ and predicted depth D , mapping them into a unified scale $[0, 1]$. Let $\tilde{D}$ and $\tilde{D}$ represent normalized estimated and rendered depth maps, respectively. The depth loss is formulated as a combination of SSIM (Structural Similarity) and L_1 distance:

$$\mathcal{L}_{\text{depth}} = \|\tilde{D} - \tilde{D}\|_1 + (1 - \text{SSIM}(\tilde{D}, \tilde{D})). \quad (4)$$

3.1.3 Thermal Gaussian Loss Following (Lu et al., 2024), the thermal supervision combines pixel fidelity, structural similarity, and a smoothness term penalizing local discontinuities:

$$\begin{aligned} \mathcal{L}_{\text{thermal}} = & (1 - \lambda_{\text{SSIM}}) \mathcal{L}_1(I_{Th}, \hat{I}_{Th}) \\ & + \lambda_{\text{SSIM}} (1 - \text{SSIM}(I_{Th}, \hat{I}_{Th})) + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}, \end{aligned} \quad (5)$$

where $\mathcal{L}_1$ denotes the mean absolute error between the ground-truth thermal image I_{Th} and the rendered thermal image $\hat{I}_{Th}$. It is used to assess the geometric accuracy.

$\mathcal{L}_{\text{smooth}}$ is a smoothness loss term introduced for regularization:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{4M} \sum_{i,j} (|T_{i\pm 1,j} - T_{i,j}| + |T_{i,j\pm 1} - T_{i,j}|), \quad (6)$$

where M is the number of rendering pixels in total and $T_{i,j}$ is the temperature at pixel (i, j) . Because objects above absolute zero continuously radiate heat and reach thermal equilibrium with their surroundings, most regions in thermal images exhibit gradual temperature changes. Hence, introducing the smoothness term effectively penalizes unnatural abrupt changes and noise.

3.1.4 Progressive Joint Optimization Instead of maintaining separate models for different modalities, we optimize a single, unified 3D Gaussian model using a combined loss function. To address the training instability caused by the sparse texture and low contrast of thermal images, we propose a progressive joint optimization strategy. In the early stages of training, the 3D geometry is highly unconstrained. Therefore, we incorporate the depth rendering loss to enforce structural accuracy based on the estimated depth prior. As the training progresses and the overall geometry stabilizes, we gradually decay the weight of the depth loss, allowing the model to focus mainly on the photometric fidelity of the thermal rendering.

The progressive weight factor $w_{\text{decay}}(t)$ at iteration t is formulated as a linear decay:

$$w_{\text{decay}}(t) = \max\left(0, 1 - \frac{t - t_{\text{start}}}{t_{\text{end}} - t_{\text{start}}}\right), \quad (7)$$

where t_{start} and t_{end} define the boundaries of the decay phase.

The final joint loss function is then formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{thermal}} + w_{\text{decay}}(t) \cdot \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{depth}}, \quad (8)$$

where λ_{depth} is a constant scaling hyperparameter for depth supervision. This progressive formulation ensures a smooth transition from geometry-guided structuring to fine-grained thermal

Datasets	MSMG (Lu et al., 2024)				Ours (TDg)			
	PSNR ↑	SSIM ↑	LPIPS ↓	Time [h] ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Time [h] ↓
RGBT-Scenes								
Day light conditions								
Dimsum	26.80	0.893	0.127	0.407	26.90	0.895	0.121	0.376
DailyStuff	21.04	0.830	0.203	0.402	21.17	0.836	0.199	0.106
Ebike	23.20	0.872	0.206	0.390	24.99	0.865	0.237	0.289
Road Block	26.51	0.921	0.221	0.388	26.73	0.921	0.212	0.159
Truck	25.90	0.875	0.145	0.384	26.28	0.878	0.133	0.227
Rotary Kiln	26.80	0.928	0.145	0.391	26.85	0.929	0.134	0.105
Parterre	24.06	0.877	0.219	0.385	24.42	0.879	0.205	0.118
Glass Cup	25.12	0.861	0.129	0.387	25.66	0.873	0.124	0.102
Dark light conditions								
Dark Scenes	19.40	0.811	0.322	0.359	19.84	0.815	0.307	0.104
Average	24.31	0.874	0.191	0.388	24.76	0.877	0.186	0.176
ThermalMix								
360-degree scenes								
Laptop	30.03	0.912	0.190	0.388	29.88	0.904	0.202	0.102
Pan	32.15	0.897	0.068	0.407	32.63	0.887	0.066	0.284
Lion	38.78	0.992	0.045	0.388	38.72	0.992	0.041	0.288
Forward-facing scenes								
Hand	26.84	0.834	0.246	0.388	26.72	0.832	0.253	0.104
Face	33.91	0.951	0.215	0.384	33.13	0.953	0.222	0.102
Average	32.34	0.917	0.153	0.391	32.22	0.914	0.157	0.176
Average (Overall)	27.18	0.8896	0.177	0.389	27.42	0.8899	0.175	0.176

Table 1. Quantitative evaluation of our TDg method on datasets RGBT-Scenes and ThermalMix, compared to MSMG baseline.

optimization, effectively avoiding the optimization shocks associated with a hard supervision switch.

4. Evaluation

In the following, we first report our implementation details, followed by the evaluation strategy and our experiments.

4.1 Implementation Details

The code is written in Python using PyTorch, with CUDA/C++ components for efficient GS rasterization. All experiments are conducted on a laptop equipped with an NVIDIA GeForce RTX 3060 GPU (6 GB VRAM), 16 GB RAM, and an Intel 12th Gen Core i7-12700H CPU (2.30 GHz). The implementation of the proposed framework, along with the experimental scripts and codes, are made available within the repository **TDg**².

For data preprocessing, depth maps are estimated from thermal images using the method of Marigold (Ke et al., 2024). The images are posed and the initial sparse point clouds are obtained from COLMAP (Schönberger and Frahm, 2016) by using the Structure-from-Motion (SfM) algorithm. Since thermal images cannot provide the necessary geometric information to construct these point clouds, we leveraged cross-modal prior knowledge from RGB images. But our core TDg method still uses only thermal images. Only for acquiring the sparse point clouds, we used RGB data.

We adopted the training settings from the official GS implementation, keeping hyperparameters unchanged while adjusting the number of iterations to between 10k and 30k, as our method converges faster due to its simplified design. As

a baseline, we select MSMG (Lu et al., 2024). For tuning the progressive weight factor (7), in our implementation, t_{start} is the first iteration, and t_{end} is typically set to 50% of the total iterations to ensure that the depth prior is fully faded out in the later stages.

4.2 Datasets

We conducted comprehensive evaluations on the standard datasets used for thermal imaging and NVS.

As datasets, we use RGBT-Scenes (Lu et al., 2024) and ThermalMix (Özer et al., 2024), following a standard protocol where 80% of the images in each scene are used for training and 20% for testing. We evaluated on 14 different scenes in total.

RGBT-Scenes: This dataset provides multi-view RGB images paired with pseudo-color thermal images, with a resolution of 640×480 (Lu et al., 2024). We evaluated on 9 scenes, including one low-light environment. The dataset is originally collected using a handheld thermal-infrared camera with accurate calibration, making it suitable for multimodal 3D reconstruction tasks. In addition, the pseudo-color thermal images can be further exploited for depth estimation during preprocessing.

ThermalMix: This dataset contains paired RGB and gray-scale thermal images of multiple indoor objects, also with a resolution of 640×480 (Özer et al., 2024). We use 5 representative scenes from this dataset for evaluation. The precise RGB–Thermal alignment facilitates reliable 3D reconstruction, and the grayscale thermal images can also be directly employed for depth estimation in our pipeline.

² Code available: <https://hannahaensen.github.io/TDg/>

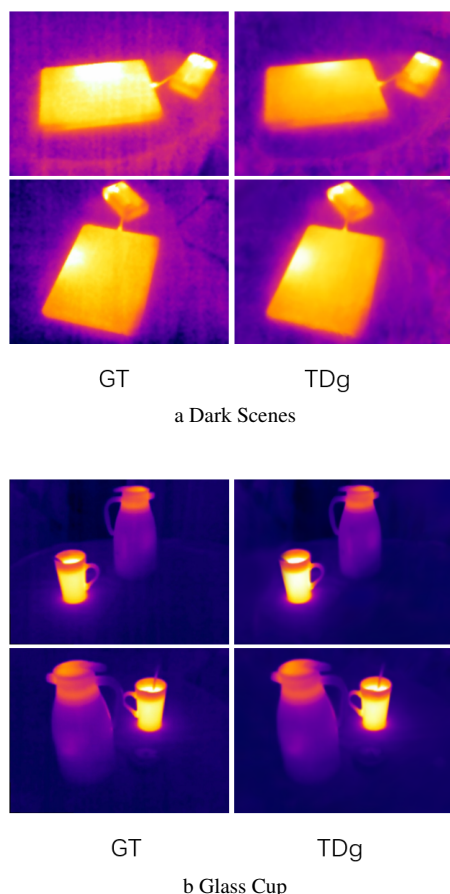


Figure 3. Rendering image results on the dataset scenes: (a) Dark Scenes (top) and (b) Glass Cup (bottom).

4.3 Metrics

For evaluation, rendering quality is measured using PSNR, SSIM, and LPIPS, which together capture pixel-level fidelity, structural similarity, and perceptual consistency. Computational efficiency is reported as the average training time per scene in hours. Lower LPIPS value, together with higher SSIM and PSNR values, signify better reconstruction fidelity. LPIPS also indicates how visually appealing the results are to humans. SSIM and LPIPS are especially important in low-contrast thermal scenarios.

4.4 Quantitative and Qualitative Results

We evaluate TDg quantitatively and compare with state-of-the-art methods. Table 1 shows a higher performance of TDg compared to MSMG overall in mean PSNR, SSIM, LPIPS, and runtime. For the runtime, overall we outperform the baseline, while for NVS metrics we outperform MSMG on all scenes from the RGBT-Scenes dataset, except for one SSIM result. For the ThermalMix dataset, we also show an overall lower runtime across all scenes. However, for PSNR, SSIM, and LPIPS, results are more mixed. Nevertheless, for PSNR we are within ± 1.0 .

For qualitative results, we can visualize Figure 3, which shows rendering images on the datasets, Glass Cup and Dark Scenes. The thermal images rendered from our resulting GS model are compared against Ground Truth (GT), thereby showing the results of correct reconstruction.

Overall, experimental results show that the depth-estimation-based approach improves reconstruction quality while maintaining accuracy, with a reduction in training time. It outperforms the baseline across various rendering quality metrics (PSNR, LPIPS, and SSIM) as well as in terms of training efficiency. E.g., on average, the accuracy metric PSNR of our method is 27.42, which is better than the baseline value of 27.18. Another notable advantage of our TDg method is the significant reduction in training time by 55% (12 mins 47 secs), while still outperforming the baseline in NVS quality. The runtime improvement is apparent in every scene, see Table 1.

4.5 Failure Cases

Random Point Cloud Initialization: We experimented with initializing the 3D Gaussians from a random sparse point cloud, bypassing the SfM requirement for point cloud generation. While this approach successfully minimized the photometric loss on the training set, it exhibited severe overfitting and failed to generalize to novel views.

Partly Occluded Areas: Additionally, our framework encounters difficulties when reconstructing large-scale targets, such as complete buildings. For such expansive structures, capturing a continuous and exhaustive 360-degree thermal multi-view trajectory is often logistically infeasible, leaving certain areas, such as the rear facades, sparsely observed. Figure 4 shows such an example of a building, where the front view was correctly reconstructed, but the back view failed due to insufficient perspective. While our depth-guided TDg approach effectively regularizes the geometry in observed regions, it cannot properly densify Gaussians in rear-seen spaces.

5. Discussion

In summary, we contribute a thermal-only GS approach, outperforming the baseline in training time and NVS quality.

A fundamental challenge in jointly optimizing thermal and RGB modalities is the inherent discrepancy in their physical properties. RGB data is highly sensitive to ambient lighting variations and high-frequency texture noise, which can introduce conflicting gradient signals during the optimization of thermal Gaussians. During the training of thermal Gaussians, it can sometimes confuse the model rather than helping it. Furthermore, forcing the model to accommodate redundant, low-quality, or misaligned RGB features can distract the optimization process, ultimately degrading the representation of critical thermal signatures. It can reduce the focus on key thermal features. This justifies our thermal-centric pipeline, which avoids these cross-modal conflicts during the core rendering phase.

For runtime efficiency, we outperform the multi-modal baseline due to an efficiency gain. This efficiency gain is directly attributed to the depth-guided supervision. In the baseline MSMG, the lack of geometric constraints causes the 3D Gaussians to perform excessive and inefficient densification in empty spaces, to minimize the photometric loss, resulting in "floaters". By incorporating the depth rendering loss ($\mathcal{L}_{depth}$), TDg provides an explicit spatial gradient that quickly guides the Gaussians to the correct geometry planes, drastically reducing unnecessary splitting and pruning operations, thereby accelerating the overall convergence.

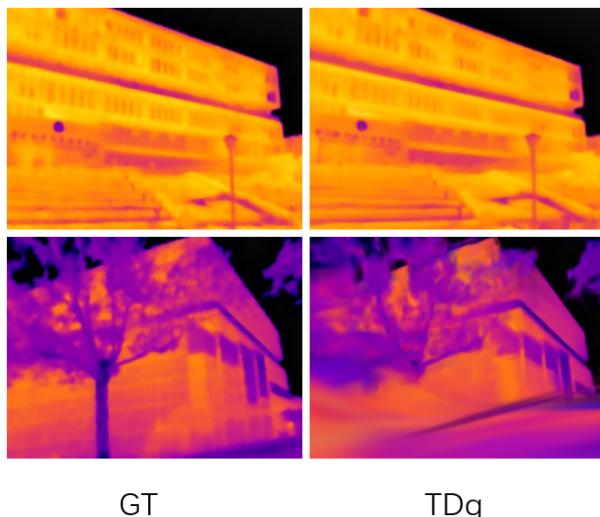


Figure 4. Building reconstruction: The front view was correctly reconstructed, but the back view failed due to insufficient perspective.

5.1 Limitations

Currently, our model still requires a sparse point cloud for initialization. Our initialization step relies on an RGB-based SfM producing a COLMAP point cloud. Because the low resolution of thermal images poses challenges for the SfM algorithm to carry out tasks such as feature extraction, matching descriptors, etc.

However, GS also allows random sparse point cloud initialization, instead of using points already in a proper geometric format. While random point clouds can be used, the NVS quality degrades. Notably, also for the baseline. This failure case confirms that without reliable geometric priors, purely thermal-based optimization easily falls into local minima, thereby justifying our design choice of utilizing RGB-assisted SfM for the initial geometric scaffolding.

As shown in Table 2, our approach, as well as the baseline, relies on a proper geometric initialization. An interesting observation was that jointly training thermal and RGB splats with random point cloud initialization also deteriorates the quality of rendered RGB images.

5.2 Future Work

Future work can focus on completely removing the reliance on RGB data, by exploring the possibility of using random sparse point clouds as initialization. This will also reduce the reliance on algorithms such as SfM. Or alternatively, perform temporally consistent normalization and local detail enhancement on thermal images to get better features for point cloud initialization. In this way, we can move towards adapting existing feature extraction and matching approaches for RGB images to thermal images and do an end-to-end reconstruction, using only thermal images. And it will also remove the constraint of performing RGB-thermal registration or alignment. Another scope of future study is to further improve the reliability and accuracy of depth estimation from thermal images.

	Random	COLMAP	PSNR / SSIM
MSMG (Lu et al., 2024)	✓	×	9.64 / 0.3471
MSMG (Lu et al., 2024)	×	✓	27.18 / 0.8896
Ours (TDg)	✓	×	9.69 / 0.3472
Ours (TDg)	×	✓	27.42 / 0.8899

Table 2. Ablation study on RGBT-Scenes and ThermalMix dataset, we report mean overall PSNR and SSIM. We compare random initialization and COLMAP-based initialization of the point cloud used to start the GS training.

6. Conclusion

In this paper, we have presented a thermal-guided depth estimation module, Thermal-to-Depth Gaussian Splatting (TDg), that uses thermal images and depth estimation in its architecture to derive the radiance fields, to accurately represent scenes. Experimental results show that our method performs better than the MSMG baseline across various rendering quality metrics and in terms of training efficiency, when tested on standard datasets. The accuracy metric PSNR of our method is 0.01% better than the baseline value, and the training time reduces by 12 mins 47 secs (55% reduction). From the training part, we can summarize that thermal images alone are powerful. Even without RGB, they provide strong structural and semantic signals. Estimated depth can be useful. Despite its noise, it helps when guided by well-crafted loss terms.

Therefore, overall, our method is successful in deriving the 3D thermal models in the form of radiance fields, which can ultimately have several applications, such as identifying heat sources (e.g., people, machinery) critical in surveillance, search and rescue operations, or industrial inspection. They can be suitable for robust robot vision regardless of lighting and weather conditions, and can be used in factories where temperature is widely used to monitor machines. Furthermore, an additional advantage is that there is no need for multi-modal data and hardware synchronization.

References

- Arampatzakis, V., Pavlidis, G., Mitianoudis, N., Papamarkos, N., 2023. Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4), 2396–2414.
- Bian, W., Wang, Z., Li, K., Bian, J.-W., Prisacariu, V. A., 2023. Nope-nerf: Optimising neural radiance field with no pose prior. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4160–4169.
- Carmichael, S., Bhat, M., Ramanagopal, M., Buchan, A., Vasudevan, R., Skinner, K. A., 2025. Trnerf: Restoring blurry, rolling shutter, and noisy thermal images with neural radiance fields. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 7980–7990.
- Chen, Q., Shu, S., Bai, X., 2024. Thermal3d-gs: Physics-induced 3d gaussians for thermal infrared novel-view synthesis. *European Conference on Computer Vision*, Springer, 253–269.
- Chung, J., Oh, J., Lee, K. M., 2024. Depth-regularized optimization for 3d gaussian splatting in few-shot images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 811–820.

- Hassan, M., Forest, F., Fink, O., Mielle, M., 2025. Ther-moNeRF: A multimodal Neural Radiance Field for joint RGB-thermal novel view synthesis of building facades. *Advanced Engineering Informatics*, 65, 103345.
- Ji, M., Qiu, R.-Z., Zou, X., Wang, X., 2024. Grasp-splats: Efficient manipulation with 3d feature splatting.
- Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C., Schindler, K., 2024. Repurposing diffusion-based image generators for monocular depth estimation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9492–9502.
- Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 1–14.
- Kleinbeck, C., Theelke, L., Schieber, H., Eck, U., von Eisenhart-Rothe, R., Roth, D., 2026. Hybrid foveated path tracing with peripheral gaussians for immersive anatomy.
- Krawczyk, J., Mazur, A. M., Sasin, T., Stokłosa, A., 2015. Infrared building inspection with unmanned aerial vehicles. *Prace Instytutu Lotnictwa*, 32–48.
- Kung, P.-C., Isaacson, S., Vasudevan, R., Skinner, K. A., 2024. Sad-gs: Shape-aligned depth-supervised gaussian splatting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2842–2851.
- Li, K., Masuda, M., Schmidt, S., Mori, S., 2025. Radiance Fields in XR: A Survey on How Radiance Fields are Envisioned and Addressed for XR Research. *IEEE Transactions on Visualization and Computer Graphics*.
- Li, Y., Pathak, D., 2024. Object-aware gaussian splatting for robotic manipulation. *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*.
- Lu, R., Chen, H., Zhu, Z., Qin, Y., Lu, M., Zhang, L., Yan, C., Xue, A., 2024. Thermalgaussian: Thermal 3d gaussian splatting. *arXiv preprint arXiv:2409.07200*.
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. 65(1), 99–106. Publisher: ACM New York, NY, USA.
- Nam, M., Park, W., Kim, M., Hur, H., Lee, S., 2025. Veta-gs: View-dependent deformable 3d gaussian splatting for thermal infrared novel-view synthesis. *2025 IEEE International Conference on Image Processing (ICIP)*, IEEE, 965–970.
- Özer, M., Weiherer, M., Hundhausen, M., Egger, B., 2024. Exploring multi-modal neural scene representations with applications on thermal imaging. *European Conference on Computer Vision*, Springer, 82–98.
- Roessle, B., Barron, J. T., Mildenhall, B., Srinivasan, P. P., Nießner, M., 2022. Dense depth priors for neural radiance fields from sparse input views. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12892–12901.
- Rüter, S., Waldow, K., Bartels, N., Fuhrmann, A., 2024. 3d gaussian splatting for construction sites. *International Conference on Construction Applications of Virtual Reality*, Springer, 37–52.
- Schieber, H., Deuser, F., Egger, B., Oswald, N., Roth, D., 2024. Nerftrinsics four: An end-to-end trainable nerf jointly optimizing diverse intrinsic and extrinsic camera parameters. *Computer Vision and Image Understanding*, 249, 104206.
- Schieber, H., Young, J., Langlotz, T., Zollmann, S., Roth, D., 2025. Semantics-controlled gaussian splatting for outdoor scene reconstruction and rendering in virtual reality. *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, IEEE, 318–328.
- Schischka, N., Schieber, H., Karaoglu, M. A., Gorgulu, M., Grötzner, F., Ladikos, A., Navab, N., Roth, D., Busam, B., 2024. Dynamon: Motion-aware fast and robust camera localization for dynamic neural radiance fields. *IEEE Robotics and Automation Letters*.
- Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Shin, U., Lee, K., Lee, B.-U., Kweon, I. S., 2022. Maximizing Self-Supervision From Thermal Image for Effective Self-Supervised Learning of Depth and Ego-Motion. *IEEE Robotics and Automation Letters*, 7(3), 7771–7778.
- Shin, U., Park, J., Kweon, I. S., 2023. Deep depth estimation from thermal image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1043–1053.
- Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M., 2025. Depthsplat: Connecting gaussian splatting and depth. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16453–16463.
- Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., Lang, X., Zhou, X., Peng, S., 2024. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. *European Conference on Computer Vision*, Springer, 156–173.
- Ye, T., Wu, Q., Deng, J., Liu, G., Liu, L., Xia, S., Pang, L., Yu, W., Pei, L., 2024. Thermal-nerf: Neural radiance fields from an infrared camera. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 1046–1053.
- Zuiderveld, K., 1994. *Contrast limited adaptive histogram equalization*. Academic Press Professional, Inc., USA, 474–485.
- Zuo, X., Ranganathan, N., Lee, C., Gkioxari, G., Chung, S.-J., 2025. MonoTher-Depth: Enhancing Thermal Depth Estimation via Confidence-Aware Distillation. *IEEE Robotics and Automation Letters*.