DarkVGGT: Seeing Through Darkness Using Thermal Geometry

Minseong Kweon[1] **Wenyuan Zhao**[2] **Nuo Chen**[2] **Lulin Liu**[1]
Huiwen Han[3] **Zihao Zhu**[2] **Srinivas Shakkottai**[2] **Chao Tian**[2] **Zhiwen Fan**[2]
[1]University of Minnesota [2]Texas A&M University [3]Stanford University

Project Page: `https://darkvggt.github.io`

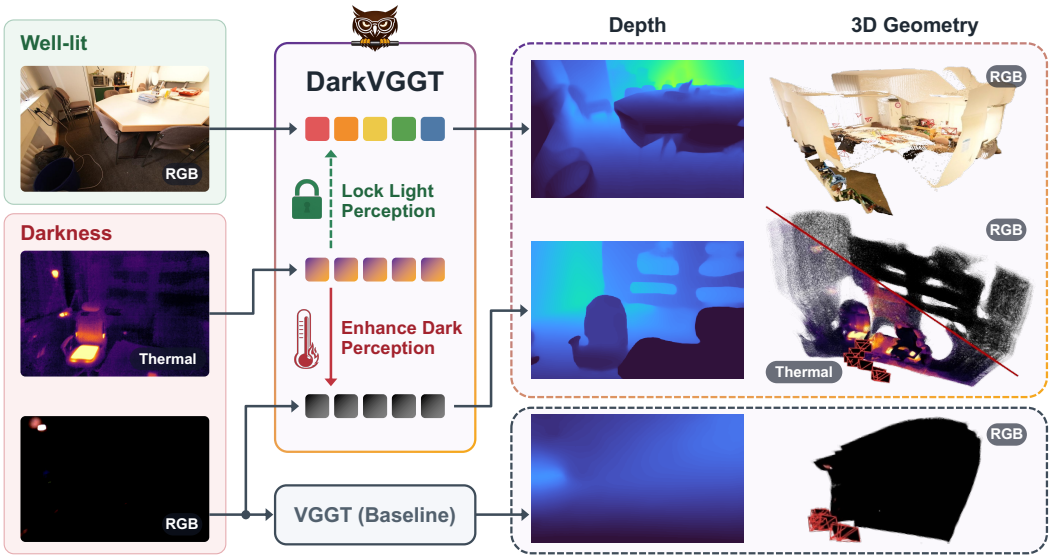

Figure 1: DarkVGGT achieves robust low-light 3D geometry estimation through effective thermal-to-RGB fusion. It performs reliably in dark environments where existing feed-forward foundation models often fail, while largely preserving well-lit performance.

Abstract

Recent feed-forward 3D reconstruction methods have demonstrated strong performance and flexibility in efficient end-to-end scene geometry estimation from image streams. However, their reliance on visible-light appearance makes them vulnerable in dark and low-visibility environments, where RGB cues are severely degraded and geometric evidence becomes ambiguous. To address this challenge, we propose DarkVGGT, a structured RGB-T adaptation of pretrained VGGT for robust 3D estimation in low-light scenes. DarkVGGT introduces two complementary modules. First, complementary thermal feature factorization processes thermal tokens through content-preserving and high-frequency residual pathways controlled by learned complementary allocation weights. Second, geometry-shared thermal routing aligns RGB and thermal representations in a shared latent space and selectively injects a gated feature correction into the RGB stream. Together, these components enable accurate thermal-informed geometry estimation under degraded RGB conditions while largely preserving performance in well-lit environments. Experiments on low-visibility RGB-T benchmarks demonstrate consistent improvements in both depth and camera pose estimation over existing feed-forward geometry baselines.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Feed-forward visual geometry models have emerged as a promising alternative to conventional SfM [31] and MVS [32] pipelines, directly predicting camera parameters and dense 3D structure from image streams with efficient inference. DUSt3R [45] demonstrates this paradigm through dense point-map regression, and subsequent work has extended it to matching [18, 6], dynamic scenes [53, 11, 7], and scalable multi-view aggregation [43, 3, 49]. VGGT [44] further unifies this direction by jointly predicting cameras, depth, point maps, and point tracks within a single feed-forward framework. These advances have broadened the applicability of feed-forward geometry models to robot navigation and manipulation [50], autonomous driving [21, 5], and real-time SLAM [34, 25, 23]. However, most existing models remain centered on well-lit RGB inputs, while reliable autonomous systems must also operate under challenging illumination, including nighttime roads and dark indoor scenes [1, 8]. Under low visibility, degraded RGB appearance and correspondence cues limit the reliability of feed-forward geometry models pretrained on well-lit images [13].

Recent methods have improved feed-forward geometry prediction in low-visibility scenes through clean-to-noisy distillation of image matching capabilities [9], event-camera-based feed-forward geometry modeling [29, 46], and thermal sensing with infrared cameras [38]. These methods show that low-light adaptation or additional sensing can recover geometric cues when RGB images become unreliable. Among these directions, thermal imaging is especially appealing for dark-scene geometry estimation because long-wave infrared radiation is largely insensitive to visible illumination and complements RGB cues when texture and color evidence degrade.

However, robust low-light performance alone is not sufficient for a reliable autonomous perception system. Such a system must improve 3D estimation under degraded illumination while preserving the geometric capability learned from standard well-lit RGB data. This requirement remains difficult because existing low-visibility adaptations may disturb the pretrained RGB geometry prior and reduce well-lit 3D perception, while multisensor inputs can introduce sparse or noisy modality-specific cues [54]. RGB-T fusion therefore needs to use thermal information as selective corrective evidence rather than as an unconstrained additional appearance stream. This is challenging for two reasons. First, foundation models such as VGGT acquire strong geometric priors from large-scale RGB pretraining, while paired RGB-T data with 3D supervision remains limited [24, 2]. Naive RGB-T mixing can therefore erode the pretrained well-lit representation and create a "daylight tax", where low-light gains come at the cost of degraded well-lit performance. Second, thermal appearance can vary independently of visible appearance and depth structure because of temperature, material properties, and background radiation [41]. Thus, RGB-T fusion should selectively modulate thermal features and route task-relevant corrections into the RGB geometry stream.

We therefore propose **DarkVGGT**, a feed-forward RGB-T geometry framework for robust 3D perception in dark and low-visibility scenes with low daylight performance degradation. DarkVGGT treats the thermal stream as a selective source of corrective evidence, rather than an unconstrained appearance channel. Within the proposed *Alternating Attention blocks*, it combines complementary thermal feature factorization with geometry-shared thermal routing: thermal features are processed through content-preserving and high-frequency residual pathways, while gated corrections in a shared latent space are routed to the RGB stream. This geometric adaptive fusion design uses thermal observations when RGB cues are unreliable, while preserving the pretrained RGB geometry prior and maintaining strong performance in well-lit scenes. Our contributions are summarized below:

- We formulate dark-scene RGB-T geometry estimation as a feed-forward multisensor fusion problem and propose **DarkVGGT**, a structured RGB-T adaptation of pretrained VGGT that retains the RGB geometry as the primary estimator while using thermal observations to provide gated corrections under low-light conditions.
- We introduce an RGB-T fusion strategy that combines complementary thermal feature factorization with geometry-shared thermal routing, providing complementary pathways of a content-preserving branch and a high-frequency residual branch, and selectively injecting corrections from a shared feature subspace to improve dark-scene geometry perception.
- DarkVGGT achieves strong depth and camera pose estimation performance across diverse low-light datasets, including ViViD++, STheReO, and Dark3R, while preserving geometric estimation performance on standard well-lit benchmarks such as ETH3D and ScanNet++, demonstrating the effectiveness of the proposed thermal-to-RGB fusion strategy.

2 Related Work

Feed-forward 3D reconstruction. DUSt3R [45] marked a significant advance in 3D reconstruction by directly regressing dense point maps from unconstrained pairwise images in an end-to-end framework. Subsequent works have evolved in several directions, including improving pairwise matching [18, 6], handling dynamic scenes [53], and accelerating inference by avoiding explicit global correction [43, 3]. VGGT [44] pushes this line further by jointly predicting cameras, depth, point maps, and 3D tracks in a unified Alternating Attention (AA) transformer, making it a natural foundation model for geometric reasoning. Beyond RGB-only reconstruction, recent works have begun to explore multimodal or degraded-visibility feed-forward geometry, including low-light structure-from-motion [9] and RGB-thermal adaptation of visual geometry transformers [38]. Most closely related to our setting, concurrent SEAR [38] shows that a pretrained VGGT can be adapted to RGB-T inputs using LoRA-based [14] fine-tuning, thermal camera tokens, and tailored batching strategies. DarkVGGT extends this baseline with complementary thermal feature factorization and geometry-shared thermal routing, selectively injecting thermal feature corrections into the RGB stream.

Low-light 3D vision with thermal images. Thermal imaging has long been recognized as complementary to RGB under low illumination, smoke, fog, and nighttime conditions. Prior work has explored thermal cues for low-light perception in tasks such as depth estimation, RGB-T sensor fusion, SLAM, and neural scene representation learning [36, 35, 4, 15, 22]. A parallel line studies RGB-T or multispectral depth fusion, including cross-spectrum depth estimation from dual-modality cameras [10], depth estimation from misaligned RGB-T image pairs [16], confidence-aware RGB-to-thermal distillation [56], and the recent align-and-fuse strategy that couples geometry-guided representation correction with an attachable fusion module [37]. Recent work on 3D scene modeling and novel-view synthesis jointly models RGB and thermal observations or introduces temperature-consistency and physics-aware priors [12, 20, 22]. Dual-band thermal videography [26] combines spectral and temporal cues to separate time-varying emission and reflection near ambient conditions. DarkVGGT focuses on learning geometry-oriented feature modulation from normalized single-band thermal inputs. For dense 3D point reconstruction, SEAR [38] proposes predicting 3D geometry by fine-tuning VGGT with LoRA on unpaired RGB-T streams. However, paired RGB-T images remain underexplored for dark-scene geometry estimation. DarkVGGT bridges this gap through an effective thermal-to-RGB fusion method.

3 Method

3.1 Overview of DarkVGGT

Feed-forward visual geometry models directly predict camera parameters, depth, point maps, and tracks from image observations, reducing reliance on conventional SfM/MVS pipelines. In this work, we leverage VGGT [44] as the base geometry predictor. Given a sequence $\{I_s\}_{s=1}^{S}$ of RGB image frames, where $I_s \in \mathbb{R}^{3\times H\times W}$, VGGT first patchifies each image and embeds it into a set of P tokens $\mathbf{x}_s \in \mathbb{R}^{P\times C}$ using DINOv2 [27]. Each frame is augmented with a camera token $\mathbf{c}_s \in \mathbb{R}^{1\times C}$ and four register tokens $\mathbf{r}_s \in \mathbb{R}^{4\times C}$. The resulting token sequence $\mathbf{t}_s := [\mathbf{c}_s; \mathbf{r}_s; \mathbf{x}_s]$ is processed by frame-wise and global Alternating Attention (AA) blocks, which model both per-view structure and cross-view correspondence. We denote the RGB patch tokens at the ℓ-th AA block as $\mathbf{x}^{\mathrm{rgb},(\ell)} \in \mathbb{R}^{BS\times P\times C}$, where B is the batch size and C is the token dimension. Task-specific prediction heads, including DPT-style dense prediction heads [28], then recover camera parameters and dense geometry from the shared representation.

Building on this, DarkVGGT introduces RGB-T thermal guidance for robust low-visibility geometry estimation. Given paired RGB and thermal frames, denoted by $\{(I_s^{\mathrm{rgb}}, I_s^{\mathrm{thr}})\}_{s=1}^{S}$, DarkVGGT predicts camera parameters (e.g., the predicted pose $\hat{P}$) and depth maps $\hat{D}$ on the RGB stream while leveraging thermal observations as an auxiliary modality in dark scenes. The framework builds on a pretrained VGGT backbone [44] with a frozen DINOv2 encoder [27]. At the ℓ-th AA block, the patched RGB and thermal tokens are denoted by $\mathbf{x}^{\mathrm{rgb},(\ell)}, \mathbf{x}^{\mathrm{thr},(\ell)} \in \mathbb{R}^{BS\times P\times C}$, respectively. For RGB-T adaptation, we use shared LoRA adapters [14] on the AA blocks with modality-specific camera tokens. Details are provided in Appendix A.1.

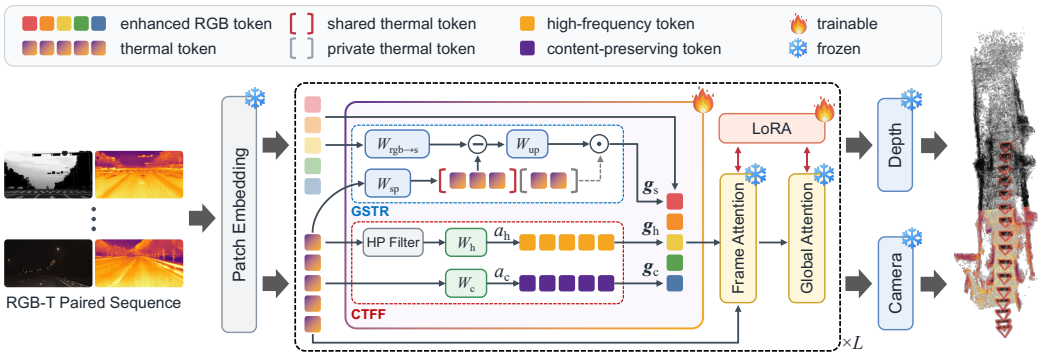

Figure 2: Overview of the DarkVGGT framework. DarkVGGT modulates thermal embeddings through content-preserving and high-frequency residual pathways using a high-pass (HP) filter and projection branches. Geometry-shared thermal routing (GSTR) learns shared/private thermal representations and injects gated corrections from the shared latent space into the RGB stream.

Figure 2 provides an overview of **DarkVGGT**. On top of the VGGT-based backbone, DarkVGGT introduces two additional modules: 1) Complementary Thermal Feature Factorization (CTFF) and 2) Geometry-Shared Thermal Routing (GSTR).

3.2 Complementary Thermal Feature Factorization

Figure 3 illustrates Complementary Thermal Feature Factorization (CTFF), which processes thermal tokens through a content-preserving (c) pathway and a high-frequency residual (h) pathway. Thermal appearance exhibits spatial variations arising from temperature, material properties, and environmental conditions [41, 42, 26]. We process the unfiltered thermal features and their local high-pass residuals through separate learned projections to accommodate these variations. Complementary allocation weights couple the two pathways, while separate gates regulate their respective contributions to the RGB stream. The single-band thermal inputs undergo normalization and CLAHE before feature extraction.

Complementary feature modulation. At each AA block, we learn a joint RGB-T cross-modal feature $\mathbf{h}^{(\ell)}$ from the joint RGB and thermal patch tokens and use it to predict complementary patch-wise allocation weights:

$$\mathbf{h}^{(\ell)} = \mathrm{FFN}\left([\mathbf{x}^{\mathrm{rgb},(\ell)}; \mathbf{x}^{\mathrm{thr},(\ell)}]\right), \qquad a_{\mathrm{c}}^{(\ell)} = \sigma\left(W_a \mathbf{h}^{(\ell)}\right), \quad a_{\mathrm{h}}^{(\ell)} = 1 - a_{\mathrm{c}}^{(\ell)}, \tag{1}$$

where $\mathrm{FFN}(\cdot)$ denotes the lightweight feed-forward network used to extract joint RGB-T features. This parameterization determines both allocation weights from a single scalar per patch. The content-preserving pathway receives unfiltered thermal tokens, while the high-frequency residual pathway emphasizes local spatial variations through a high-pass filter. The modulated thermal features are

$$\mathbf{z}_{\mathrm{c}}^{(\ell)} = a_{\mathrm{c}}^{(\ell)} \odot W_{\mathrm{c}} \mathbf{x}^{\mathrm{thr},(\ell)}, \qquad \mathbf{z}_{\mathrm{h}}^{(\ell)} = a_{\mathrm{h}}^{(\ell)} \odot W_{\mathrm{h}}\, \mathrm{HP}(\mathbf{x}^{\mathrm{thr},(\ell)}), \tag{2}$$

where W_{c} and W_{h} are learnable projections for the content-preserving and high-frequency residual pathways, respectively. $\mathrm{HP}(\cdot)$ denotes the high-pass filter on the patch grid, implemented as $x - \mathrm{AvgPool}_{3\times3}(x)$. The RGB token is adapted by

$$\tilde{\mathbf{x}}^{\mathrm{rgb},(\ell)} = \mathbf{x}^{\mathrm{rgb},(\ell)} + \boldsymbol{g}_{\mathrm{c}}^{(\ell)} \odot \mathbf{z}_{\mathrm{c}}^{(\ell)} + \boldsymbol{g}_{\mathrm{h}}^{(\ell)} \odot \mathbf{z}_{\mathrm{h}}^{(\ell)}, \tag{3}$$

where $\boldsymbol{g}_k^{(\ell)} = \sigma(W_{g,k}\mathbf{h}^{(\ell)})$ for $k \in \{\mathrm{c}, \mathrm{h}\}$ denotes the gate. The content-preserving gate is additionally biased by a learned per-patch offset parameterized as a log-variance difference $\Delta u^{(\ell)} = \log \sigma^2_{\mathrm{rgb}} - \log \sigma^2_{\mathrm{thr}}$ predicted from $\mathbf{h}^{(\ell)}$. As shown in Figure 3, complementary allocation weights and pathway-specific gates jointly modulate the thermal correction added to the RGB tokens. To encourage complementary specialization, we introduce three auxiliary regularizers.

The sparsity bias $\mathcal{L}_{\text{sparse}}$ penalizes the averaged high-frequency allocation $a_{\text{h}}^{(\ell)}$, the edge-disagreement loss $\mathcal{L}_{\text{edge}}$ supervises this allocation with a soft target derived from thermal–depth edge disagreement, and the orthogonality penalty $\mathcal{L}_{\text{ortho}}$ discourages redundancy between the two factorized embeddings:

$$\mathcal{L}_{\text{sparse}} = \mathbb{E}\left[a_{\text{h}}^{(\ell)}\right], \tag{4}$$

$$\mathcal{L}_{\text{edge}} = \text{BCE}\left(a_{\text{h}}^{(\ell)}, a_{\text{h,tar}}^{(\ell)}\right), \tag{5}$$

$$\mathcal{L}_{\text{ortho}} = \mathbb{E}\left[\text{Sim}^2\left(\mathbf{z}_{\text{c}}^{(\ell)}, \mathbf{z}_{\text{h}}^{(\ell)}\right)\right], \tag{6}$$

where $\mathbb{E}[\cdot]$ denotes expectation over batch, frame, and patch locations, $\text{BCE}(\cdot,\cdot)$ denotes binary cross-entropy, and $\text{Sim}(\cdot,\cdot)$ denotes cosine similarity. The soft target $a_{\text{h,tar}}^{(\ell)} = \sigma(\beta(\bar{E}_{\text{thr}}^{(\ell)} - \bar{E}_{D}^{(\ell)} - \tau))$ is derived from thermal–depth edge disagreement, with $\bar{E}_{\text{thr}}^{(\ell)}$ and $\bar{E}_{D}^{(\ell)}$ denoting patch-pooled thermal and ground-truth depth edge magnitudes, β controlling target sharpness, and τ setting the disagreement margin.

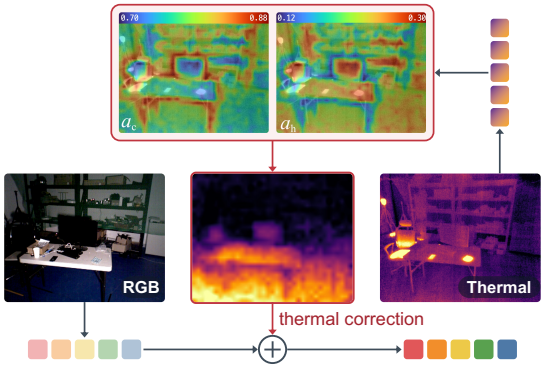

Figure 3: Visualization of complementary thermal feature factorization. Per-patch allocation weights control the content-preserving and high-frequency residual pathways before gated RGB fusion.

3.3 Geometry-Shared Thermal Routing

While thermal cues can compensate for degraded RGB in dark scenes, they contain both geometry-relevant and modality-specific components [41]. We introduce a Geometry-Shared Thermal Routing module that encourages cross-modal alignment in a shared latent space via stop-gradient distillation and uses a private branch for thermal reconstruction.

Geometry-shared thermal decomposition. Given RGB and thermal patch tokens $\mathbf{x}^{\text{rgb},(\ell)}$ and $\mathbf{x}^{\text{thr},(\ell)}$ at an ℓ-th AA block, we decompose the thermal tokens into geometry-shared and thermal-private components [47] and project RGB into the geometry-shared space:

$$[\mathbf{u}^{\text{s},(\ell)};\, \mathbf{u}^{\text{p},(\ell)}] = W_{\text{sp}}\, \mathbf{x}^{\text{thr},(\ell)}, \qquad \mathbf{v}^{\text{s},(\ell)} = W_{\text{rgb}\to\text{s}}\, \mathbf{x}^{\text{rgb},(\ell)}. \tag{7}$$

Here, $\mathbf{u}^{\text{s},(\ell)}$ and $\mathbf{u}^{\text{p},(\ell)}$ denote thermal features trained for cross-modal alignment and thermal reconstruction, respectively, and $\mathbf{v}^{\text{s},(\ell)}$ denotes the RGB feature projected into the geometry-shared space. We use $\bar{\mathbf{u}}^{\text{s},(\ell)}$, $\bar{\mathbf{u}}^{\text{p},(\ell)}$, and $\bar{\mathbf{v}}^{\text{s},(\ell)}$ to denote their layer-normalized realizations.

Thermal routing supervision. With the geometry-shared thermal branch as a stop-gradient teacher, modality-specific projection heads followed by L_2 normalization map $\bar{\mathbf{v}}^{\text{s},(\ell)}, \bar{\mathbf{u}}^{\text{s},(\ell)}$ into a common embedding space, yielding $\check{\mathbf{v}}^{\text{s},(\ell)}, \check{\mathbf{u}}^{\text{s},(\ell)}$:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}\left[1 - \left\langle \check{\mathbf{v}}^{\text{s},(\ell)}, \text{sg}\left[\check{\mathbf{u}}^{\text{s},(\ell)}\right]\right\rangle\right]. \tag{8}$$

where $\text{sg}[\cdot]$ blocks gradients on the teacher branch (see Appendix A.1 for projection-head dimensions). We then encourage the private branch to retain thermal input information by reconstructing the thermal tokens with a variance-normalized regression loss:

$$\mathcal{L}_{\text{recon}} = \frac{\text{MSE}\left(W_{\text{rec}}\mathbf{u}^{\text{p},(\ell)},\, \text{sg}[\mathbf{x}^{\text{thr},(\ell)}]\right)}{\text{Var}\left(\text{sg}[\mathbf{x}^{\text{thr},(\ell)}]\right) + \epsilon}. \tag{9}$$

where W_{rec} maps the thermal-private features back to the thermal token space, and ϵ is a small constant for numerical stability. Finally, to discourage redundancy between the geometry-shared and thermal-private subspaces, we add a small cross-branch decorrelation loss $\mathcal{L}_{\text{decorr}} = \|(\mathbf{u}^{\text{s},(\ell)})^{\top}\mathbf{u}^{\text{p},(\ell)}\|_F^2$.

Gated shared-feature correction. The RGB stream is updated using the difference between thermal and RGB features in the shared latent space, while the thermal-private branch supports

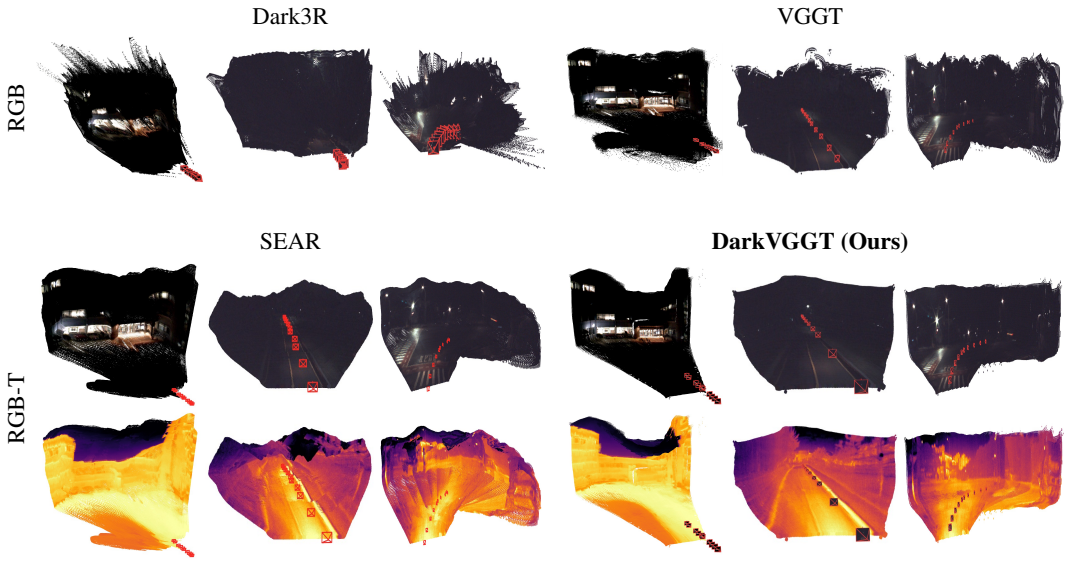

Figure 4: Qualitative comparison of nighttime 3D geometry estimation across Dark3R, VGGT, SEAR, and DarkVGGT. Red cameras indicate predicted poses. For RGB-T methods, we show RGB-textured and thermal-intensity-colored point-cloud reconstructions.

thermal reconstruction and gate conditioning. We compute a per-token feature-conditioned gate $\boldsymbol{g}_{\mathrm{s}}^{(\ell)}$ by comparing the RGB geometry-shared and thermal-private representations,

$$\Delta u_{\mathrm{sp}}^{(\ell)} = W_{\sigma}^{\mathrm{rgb}}\,\bar{\mathbf{v}}^{\mathrm{s},(\ell)} - W_{\sigma}^{\mathrm{thr}}\,\bar{\mathbf{u}}^{\mathrm{p},(\ell)}, \qquad \boldsymbol{g}_{\mathrm{s}}^{(\ell)} = \sigma\Big(W_{\mathrm{g}}[\bar{\mathbf{v}}^{\mathrm{s},(\ell)};\bar{\mathbf{u}}^{\mathrm{s},(\ell)}] + \Delta u_{\mathrm{sp}}^{(\ell)}\Big), \tag{10}$$

where $\Delta u_{\mathrm{sp}}^{(\ell)}$ is a learned offset that modulates the gate using geometry-shared and thermal-private features. We then route the gated geometry-shared thermal residual to the RGB tokens:

$$\tilde{\mathbf{x}}^{\mathrm{rgb},(\ell)} = \mathbf{x}^{\mathrm{rgb},(\ell)} + \boldsymbol{g}_{\mathrm{s}}^{(\ell)} \odot W_{\mathrm{up}}\Big(\mathrm{sg}[\bar{\mathbf{u}}^{\mathrm{s},(\ell)}] - \bar{\mathbf{v}}^{\mathrm{s},(\ell)}\Big), \tag{11}$$

where W_{up} projects the geometry-shared latent space back to the token space, and $\tilde{\mathbf{x}}^{\mathrm{rgb},(\ell)}$ denotes the enhanced RGB tokens. In practice, this module is instantiated only in the final k AA blocks.

3.4 Training

Thermal dropout and RGB-prior preservation. To reduce over-reliance on thermal cues and preserve the pretrained RGB prior, we apply thermal-branch dropout during training on batches whose clips all come from RGB-dominant scenes (see Appendix C.3). For such batches, the thermal inputs $\{I_s^{\mathrm{thr}}\}_{s=1}^{S}$ are dropped with a fixed probability, so the model performs an RGB-only forward pass. On these dropout steps, we additionally regularize the adapted model toward a frozen base-VGGT reference evaluated on the same RGB frames: $\mathcal{L}_{\mathrm{drop}} = \|\hat{D}_{\mathrm{rgb}} - \hat{D}^0_{\mathrm{rgb}}\|_2^2 + \|\hat{P}_{\mathrm{rgb}} - \hat{P}^0_{\mathrm{rgb}}\|_2^2$, where $(\hat{D}_{\mathrm{rgb}}, \hat{P}_{\mathrm{rgb}})$ and $(\hat{D}^0_{\mathrm{rgb}}, \hat{P}^0_{\mathrm{rgb}})$ denote RGB-only depth and pose predictions from the adapted student and the frozen base-VGGT reference, respectively. This term directly counteracts the daylight tax by anchoring the RGB-only forward of the adapted model to its pretrained behavior.

Training losses. DarkVGGT optimizes

$$\mathcal{L} = \mathcal{L}_{\mathrm{VGGT}} + \mathcal{L}_{\mathrm{ctff}} + \mathcal{L}_{\mathrm{gstr}} + \mathcal{I}_{\mathrm{drop}}\,\lambda_{\mathrm{drop}}\,\mathcal{L}_{\mathrm{drop}}, \tag{12}$$

where $\mathcal{I}_{\mathrm{drop}}$ equals 1 on thermal-dropout steps and 0 otherwise, and $\mathcal{L}_{\mathrm{VGGT}}$ follows VGGT [44]. The terms $\mathcal{L}_{\mathrm{ctff}}$ and $\mathcal{L}_{\mathrm{gstr}}$ are weighted sums of the auxiliary losses from the complementary thermal feature factorization module described in Section 3.2 and the geometry-shared thermal routing module described in Section 3.3, respectively.

Table 1: Depth estimation across low-visibility scenes. DarkVGGT is compared with feed-forward 3D reconstruction baselines, reporting quantitative depth metrics to evaluate robustness under severe illumination degradation. † denotes a model fine-tuned under our training setup. 1st 2nd 3rd.

Method	ViViD++			STheReO			Dark3R		
	AbsRel↓	$\delta<1.25\uparrow$	$\mathrm{RMSE}_{\log}\downarrow$	AbsRel↓	$\delta<1.25\uparrow$	$\mathrm{RMSE}_{\log}\downarrow$	AbsRel↓	$\delta<1.25\uparrow$	$\mathrm{RMSE}_{\log}\downarrow$
(A) RGB-only									
DUSt3R [45]	0.3765	0.4883	0.4008	0.4759	0.3146	0.6437	0.2118	0.6391	0.2562
MASt3R [18]	0.2685	0.6121	0.3132	0.5093	0.3418	0.6619	0.2004	0.6933	0.2496
Dark3R [9]	0.3036	0.5739	0.3392	0.4843	0.3404	0.6437	0.1510	0.7914	0.1960
DepthAnything3 [19]	0.2552	0.6239	0.2990	0.3145	0.4746	0.5246	0.2054	0.6528	0.2489
VGGT [44]	0.2746	0.5874	0.3234	0.2950	0.5156	0.4567	0.2433	0.6095	0.2861
(B) RGB-T									
SEAR [38]	0.1956	0.7502	0.2421	0.2370	0.6143	0.4147	0.1703	0.7058	0.2151
SEAR† [38]	0.1806	0.7670	0.2378	0.1780	0.7323	0.3399	0.1305	0.8274	0.1696
DarkVGGT (ours)	0.1274	0.8692	0.1970	0.1600	0.7651	0.3077	0.1246	0.8387	0.1651

4 Experiments

Implementation details. We initialize from the publicly released VGGT-1B weights [44] and keep the backbone frozen. Trainable parameters comprise LoRA adapters ($r = 64$, $\alpha = 128$) on the AA blocks, a thermal camera token and RoPE offset, the CTFF allocation weights and fusion gates, and the shared–private decomposition with ranks $(d_s, d_p, d_{\mathrm{mi}}) = (96, 192, 128)$ active in the last $k = 8$ AA blocks. We train DarkVGGT on four NVIDIA A100 GPUs at a resolution of 518×518 using bfloat16 AMP for 10 epochs with 200 optimization steps per epoch and thermal dropout probability $p = 0.5$. We optimize the model with AdamW using a learning rate of 5×10^{-5}, a weight decay of 10^{-2}, and cosine decay, with the remaining hyperparameters provided in Appendix C.3.

Additional controlled input and fusion comparisons and the scope of the learned supervision prior are provided in Appendices B.3 and C.2.

The RGB-only baselines are evaluated zero-shot. We report both off-the-shelf SEAR and SEAR†, fine-tuned with the same paired splits, depth supervision, input resolution, VGGT-1B initialization, optimization schedule, and training budget as DarkVGGT. Comparisons use the stronger SEAR variant for each dataset and metric, since fine-tuning does not uniformly improve pose estimation.

Datasets. We evaluate DarkVGGT on two real multi-sensor RGB-T datasets, ViViD++[17] and STheReO[52], and report Dark3R [9] as a complementary pseudo-thermal low-light evaluation. Since Dark3R does not provide measured thermal images, we synthesize pseudo-thermal inputs with ThermalGen [48] and treat it separately from real RGB-T benchmarks. For ViViD++ and STheReO, we supervise dense depth prediction using LiDAR-derived dense depth maps refined with Lingbot-Depth [40], while for Dark3R, we use the released canonical dense depth maps. For evaluation, we use `indoor_aggresive_dark`, `indoor_unstable_dark`, and `outdoor_robust_night2` from ViViD++, `kaist_evening` and `valley_evening` from STheReO, and `dark_chapel` and `dark_kitchen` from Dark3R as test sequences. To assess RGB-prior preservation under standard light conditions, we evaluate RGB-only performance on ETH3D [33] and ScanNet++ [51]. Each test sequence is split into non-overlapping 50 frame chunks with 10 frames uniformly sampled per chunk following VGGT [44], except for Dark3R dataset, where we use a fixed stride of 20 across the full scene. Detailed data preprocessing and train-test sequence splits are provided in Appendix C.1. Dark3R frame-sampling variance is analyzed in Appendix B.4.

Metrics. We evaluate depth and pose estimation performance following the protocol used in prior work [44, 55]. For depth estimation, we report AbsRel, $\delta < 1.25$, and $\mathrm{RMSE}_{\log}$, which measure relative depth error, threshold-based accuracy, and logarithmic depth deviation, respectively. We use raw LiDAR measurements as ground-truth depth for ViViD++ [17] and STheReO [52], and the released canonical depth maps for Dark3R [9]. For pose estimation, we report RRA, RTA, and AUC, corresponding to relative rotation accuracy, relative translation accuracy, and the area under the cumulative pose-accuracy curve, respectively.

Depth estimation performance comparison. Table 1 compares the depth estimation performance of DarkVGGT with RGB-only and RGB-T feed-forward 3D reconstruction models across three

Table 2: Camera pose estimation on low-visibility RGB-T scenes. We compare DarkVGGT with feed-forward 3D reconstruction baselines and report pose metrics under weak geometric correspondences.

Method	ViViD++			STheReO			Dark3R		
	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑
(A) RGB-only									
DUSt3R [45]	0.72	0.22	6.4	0.88	0.43	27.0	0.98	0.20	10.0
MASt3R [18]	0.76	0.46	27.1	0.98	0.80	71.3	1.00	0.89	62.5
Dark3R [9]	0.74	0.41	22.7	0.99	0.79	69.8	1.00	0.94	68.0
DepthAnything3 [19]	0.73	0.28	12.0	0.99	0.85	76.6	0.55	0.18	5.1
VGGT [44]	0.81	0.48	25.9	1.00	0.99	96.3	0.64	0.47	15.3
(B) RGB-T									
SEAR [38]	0.91	0.59	36.1	1.00	0.97	88.5	0.66	0.52	18.6
SEAR† [38]	0.84	0.47	28.1	1.00	0.98	95.2	0.84	0.71	32.5
DarkVGGT (ours)	0.98	0.74	46.0	1.00	0.98	95.4	1.00	0.95	57.4

Table 3: Depth and pose estimation on well-lit ETH3D and ScanNet++ scenes. We evaluate RGB-only performance to examine whether DarkVGGT preserves the pretrained VGGT geometry prior when thermal guidance is unavailable.

Method	**ETH3D**				**ScanNet++**			
	$\delta < 1.25$ ↑	$\mathrm{RMSE}_{\log}$ ↓	RRA@30 ↑	RTA@30 ↑	$\delta < 1.25$ ↑	$\mathrm{RMSE}_{\log}$ ↓	RRA@30 ↑	RTA@30 ↑
MASt3R [18]	0.9413	0.1496	0.67	0.68	0.8762	0.1890	0.89	0.90
Dark3R [9]	0.8992	0.1739	0.63	0.65	0.8889	0.1802	0.88	0.89
VGGT [44]	0.9729	0.0595	1.00	0.91	0.9479	0.1106	0.98	0.93
DarkVGGT (ours)	0.9584	0.0803	1.00	0.88	0.9608	0.1016	0.98	0.92

low-visibility datasets. DarkVGGT achieves the lowest numerical AbsRel and $\mathrm{RMSE}_{\log}$, as well as the highest $\delta < 1.25$ across all three datasets. Compared with fine-tuned SEAR [38], which achieves better depth scores than its off-the-shelf variant on all three benchmarks, DarkVGGT reduces average AbsRel from 0.1630 to 0.1373, increases $\delta < 1.25$ from 0.7756 to 0.8243, and lowers $\mathrm{RMSE}_{\log}$ from 0.2491 to 0.2232. DarkVGGT and fine-tuned SEAR achieve comparable depth performance on Dark3R dataset. Against the RGB-only model Dark3R [9], DarkVGGT yields lower AbsRel on all three benchmarks, with the largest gains on ViViD++ and STheReO, while also reducing $\mathrm{RMSE}_{\log}$ across the board. On the ViViD++ and STheReO datasets, DepthAnything3 [19] and VGGT [44] obtain the best numerical scores among RGB-only models, respectively, yet both remain substantially behind RGB-T methods. This suggests that thermal cues provide an effective inductive bias when adapting feed-forward 3D reconstruction models to dark scenes. In particular, compared with our baseline model, VGGT, DarkVGGT reduces the average AbsRel and $\mathrm{RMSE}_{\log}$ over the three datasets by 0.1337 and 0.1322, respectively, while improving the average $\delta < 1.25$ by 0.2535.

These results support complementary thermal feature factorization and gated shared-subspace correction for multimodal adaptation and geometric prediction in low-visibility scenes. As shown in Figure 4, DarkVGGT improves depth estimation in nighttime scenes and reconstructs more accurate dense 3D point clouds than existing RGB-only models, including Dark3R and VGGT, as well as the RGB-T model SEAR.

Pose estimation performance comparison. Table 2 shows pose estimation performance comparing DarkVGGT with existing methods using $\mathrm{RRA@30}$, $\mathrm{RTA@30}$, and $\mathrm{AUC@30}$, where @30 denotes evaluation up to a 30° angular threshold for relative rotation, translation-direction, and joint pose errors. DarkVGGT achieves the strongest average pose performance across the three low-visibility datasets. Compared with the average of RGB-only baselines, DarkVGGT improves $\mathrm{RRA@30}$, $\mathrm{RTA@30}$, and $\mathrm{AUC@30}$ by 16.7%, 59.1%, and 66.8%, respectively. Using the stronger SEAR [38] variant for each dataset and metric, DarkVGGT raises the average $\mathrm{RRA@30}$, $\mathrm{RTA@30}$, and $\mathrm{AUC@30}$ from 0.917 to 0.993, from 0.760 to 0.890, and from 54.6 to 66.3, respectively. On STheReO, the two methods match in $\mathrm{RRA@30}$ and $\mathrm{RTA@30}$, with a small $\mathrm{AUC@30}$ difference of 95.2 versus 95.4. Relative to Dark3R [9], DarkVGGT also shows consistent average gains, with the largest improvement in $\mathrm{RTA@30}$. The overall pose scores on the STheReO dataset are higher than on the other two datasets, likely because its forward-driving scenes induce smaller relative pose errors.

Figure 4 provides a qualitative view of these pose differences. Dark3R produces trajectories that collapse toward one side, leading to distorted dense point-cloud reconstructions. SEAR, as an

Table 4: Effectiveness of each proposed module and computational cost for RGB-T multimodal nighttime 3D geometry estimation.

	Depth			Camera Pose			Cost	
Method	AbsRel ↓	$\delta < 1.25$ ↑	RMSE_{log} ↓	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	#Param. (M)	FPS
AA	0.1672	0.7512	0.2569	0.95	0.77	61.5	604.73	9.0
LoRA	0.1658	0.7590	0.2542	0.96	0.79	62.2	50.33	8.9
LoRA + CTFF	0.1422	0.8088	0.2298	1.00	0.86	68.9	107.21	7.5
LoRA + GSTR	0.1538	0.7879	0.2386	0.97	0.79	62.4	70.81	7.5
LoRA + CTFF + GSTR (Full)	0.1373	0.8243	0.2232	0.99	0.89	66.3	124.34	7.4

Table 5: Ablation study on the effectiveness of each loss function for both depth and pose estimation.

	Depth			Camera Pose		
Configuration	AbsRel ↓	$\delta < 1.25$ ↑	$\text{RMSE}_{\log}$ ↓	RRA@5 ↑	RTA@5 ↑	AUC@30 ↑
Full	0.1373	0.8243	0.2232	0.68	0.45	66.3
w/o $\mathcal{L}_{\text{sparse}}$	0.1381	0.8188	0.2249	0.68	0.45	69.2
w/o $\mathcal{L}_{\text{recon}}$	0.1404	0.8151	0.2274	0.68	0.44	65.9
w/o $\mathcal{L}_{\text{edge}}$	0.1418	0.8145	0.2287	0.67	0.43	65.3

RGB-T baseline, improves over VGGT [44] in 3D reconstruction quality but still exhibits locally warped camera trajectories. In contrast, DarkVGGT estimates more stable camera poses, supporting higher-quality 3D geometry prediction under nighttime conditions.

RGB-only performance in well-lit scenes. To assess whether DarkVGGT preserves the pretrained VGGT prior when thermal guidance is unavailable, we evaluate RGB-only performance on the well-lit ETH3D [33] and ScanNet++ [51] datasets. As shown in Table 3, DarkVGGT remains close to VGGT [44] overall and even improves some metrics. On ETH3D dataset, it achieves 0.9584 $\delta < 1.25$ and 0.0803 $\text{RMSE}_{\log}$, compared with 0.9729 and 0.0595 for VGGT. On ScanNet++ dataset, it improves both metrics, reaching 0.9608 and 0.1016 versus 0.9479 and 0.1106 for VGGT.

Averaged over the two datasets, DarkVGGT differs from VGGT by only 0.0008 in $\delta < 1.25$ and 0.0059 in $\text{RMSE}_{\log}$, exactly matches the average RRA@30, and shows only a 0.02 decrease in RTA@30 from 0.92 to 0.90. This deviation is notably smaller than the teacher gap observed for Dark3R [9], which shows a 0.0147 decrease in average $\delta < 1.25$ and a 0.0078 increase in $\text{RMSE}_{\log}$ relative to MASt3R [18]. The degradation is more pronounced on ETH3D dataset, where $\delta < 1.25$ decreases from 0.9413 to 0.8992 and $\text{RMSE}_{\log}$ increases from 0.1496 to 0.1739. For pose estimation, Dark3R likewise underperforms MASt3R by 0.025 in RRA@30 and 0.02 in RTA@30, while DarkVGGT achieves higher pose accuracy than Dark3R in our experiments on well-lit sequences. Please refer to Appendix B.5 for the results obtained when using RGB-T inputs on daylight sequences.

Ablation study. Table 4 demonstrates the effectiveness of each proposed module for RGB-T nighttime 3D geometry estimation. AA denotes full fine-tuning of the alternating attention blocks without LoRA, while CTFF and GSTR denote the complementary thermal feature factorization module and geometry-shared thermal routing, respectively. Adding LoRA to the AA baseline yields only modest gains, improving AbsRel from 0.1672 to 0.1658, $\delta < 1.25$ from 0.7512 to 0.7590, and AUC@30 from 61.5 to 62.2. In contrast, CTFF provides the largest standalone improvement, reducing AbsRel to 0.1422 and $\text{RMSE}_{\log}$ to 0.2298, while also increasing RRA@30, RTA@30, and AUC@30 to 1.00, 0.86, and 68.9, respectively. GSTR provides a smaller standalone gain, but complements CTFF in the full model, reaching 0.1373 AbsRel, 0.8243 $\delta < 1.25$, and 0.89 RTA@30 (Table 4). Relative to CTFF alone, these depth and translation gains are accompanied by decreases in RRA@30 from 1.00 to 0.99 and aggregate AUC@30 from 68.9 to 66.3. Selected sequence-level comparisons between LoRA+CTFF and the full model are provided in Appendix B.2. Compared with AA, the full model improves all metrics, including a 17.9% reduction in AbsRel and a 7.8% increase in AUC@30. Regarding computational efficiency, our full model achieves superior performance with the proposed modules while requiring 79.4% fewer learnable parameters than full AA fine-tuning, reducing the parameter count from 604.73M to 124.34M and achieving an inference speed of 7.4 FPS. These results highlight the benefit of combining thermal feature modulation with shared-subspace routing for nighttime geometry estimation.

Table 5 ablates the three regularizers by removing each term individually. Since pose differences are relatively small, we report $\mathrm{RRA@5}$ and $\mathrm{RTA@5}$ for a more fine-grained comparison. Dropping $\mathcal{L}_{\mathrm{edge}}$ causes the largest AbsRel increase, from 0.1373 to 0.1418, as it supervises residual allocation using thermal–depth edge disagreement. Without $\mathcal{L}_{\mathrm{recon}}$, AbsRel degrades to 0.1404, as the thermal reconstruction constraint on the private subspace is removed. $\mathcal{L}_{\mathrm{sparse}}$ has the smallest AbsRel impact, but its removal lowers $\delta < 1.25$, suggesting that limiting residual allocation benefits depth estimation. Although this variant improves $\mathrm{AUC@30}$ from 66.3 to 69.2, it does so at the cost of degraded depth estimation. These findings support the complementary roles of allocation regularization and private-subspace reconstruction.

5 Conclusion

We presented DarkVGGT, a structured RGB-T adaptation of pretrained VGGT that leverages thermal observations for robust 3D geometry estimation in low-visibility scenes. DarkVGGT modulates content-preserving and high-frequency residual thermal pathways through learned complementary allocation weights. Additionally, geometry-shared thermal routing injects gated corrections derived from thermal–RGB feature differences in a shared latent space. Across low-visibility RGB-T benchmarks, DarkVGGT outperforms existing RGB-only and RGB-T feed-forward models in depth estimation and in camera pose estimation averaged across the three datasets. Experiments on well-lit datasets including ETH3D and ScanNet++ show that DarkVGGT largely preserves the original capability of VGGT, thereby reducing the daylight tax of low-light adaptation. Ablation studies further indicate that complementary thermal feature factorization drives the main gains, while geometry-shared thermal routing provides complementary refinement with little runtime overhead. Overall, DarkVGGT demonstrates that thermal guidance can improve dark-scene 3D perception while largely preserving RGB robustness in well-lit settings.

Limitation. Our experiments assume moderately aligned paired RGB-T inputs, which may limit applicability to weakly aligned or unpaired settings. Addressing such cases would require additional cross-modal correspondence or correction mechanisms. Cross-modal appearance differences can also weaken correspondence when ordinary glass is transparent in visible light but opaque in long-wave infrared, or when internal heat sources induce surface thermal patterns without corresponding visible texture. Our modules also introduce a few hyperparameters for residual scaling, sparsity, and fusion budgeting, which may require validation-based tuning or sensor-specific calibration. These limitations are orthogonal to the VGGT backbone.

Future work. Future work includes adaptive RGB-T fusion, weaker supervision settings such as partially aligned or unpaired RGB-T data, and extensions to other multimodal geometry cues such as event cameras or LiDAR. While this direction may support safer robotic and autonomous perception in low-visibility environments, RGB-T systems can also raise privacy concerns in public or sensitive spaces due to thermal sensing. Future deployment should therefore follow appropriate privacy, safety, and data-governance safeguards.

6 Acknowledgements

This work was supported in part by GPU hardware and cloud credits provided to the Texas A&M University PHAI Lab through the NVIDIA Academic Grant Program. We also acknowledge computational resources provided by the Texas A&M VISION GPU Cluster. Z. Fan was supported in part by research gifts from Meta and the US AI Glasses Impact Grants program. The work of W. Zhao and C. Tian was supported in part by the National Science Foundation (NSF) via grants DMS-2312173 and ECCS-2433631.

References

[1] Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019.

[2] Martin Brenner, Napoleon H Reyes, Teo Susnjak, and Andre LC Barczak. RGB-D and thermal sensor fusion: A systematic literature review. *IEEE Access*, 11:82410–82442, 2023.

[3] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view network for stereo 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1050–1060, 2025.

[4] 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*, 2024.

[5] Hoonhee Cho, Jae-Young Kang, Giwon Lee, Hyemin Yang, Heejun Park, Seokwoo Jung, and Kuk-Jin Yoon. Vr-drive: Viewpoint-robust end-to-end driving with feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2510.23205*, 2025.

[6] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3R-SfM: A fully-integrated solution for unconstrained structure-from-motion. In *2025 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2025.

[7] Juntong Fang, Zequn Chen, Weiqi Zhang, Donglin Di, Xuancheng Zhang, Chengmin Yang, and Yu-Shen Liu. More: Motion-aware feed-forward 4d reconstruction transformer. *arXiv preprint arXiv:2603.05078*, 2026.

[8] Bahareh Ghari, Ali Tourani, Asadollah Shahbahrami, and Georgi Gaydadjiev. Pedestrian detection in low-light conditions: A comprehensive survey. *Image and Vision Computing*, 148:105106, 2024.

[9] Andrew Y Guo, Anagh Malik, SaiKiran Tedla, Yutong Dai, Yiqian Qin, Zach Salehe, Benjamin Attal, Sotiris Nousias, Kyros Kutulakos, and David B Lindell. Dark3R: Learning structure from motion in the dark. *arXiv preprint arXiv:2603.05330*, 2026.

[10] Yubin Guo, Haobo Jiang, Xinlei Qi, Jin Xie, Cheng-Zhong Xu, and Hui Kong. Unsupervised visible-light images guided cross-spectrum depth estimation from dual-modality cameras. *arXiv preprint arXiv:2205.00257*, 2022.

[11] Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. Dˆ 2ust3r: Enhancing 3d reconstruction with 4d pointmaps for dynamic scenes. *arXiv e-prints*, pages arXiv–2504, 2025.

[12] Mariam Hassan, Florent Forest, Olga Fink, and Malcolm Mielle. ThermoNeRF: Joint RGB and thermal novel view synthesis for building facades using multimodal neural radiance fields. *arXiv preprint arXiv:2403.12154*, 2024.

[13] Yuze He, Yubin Hu, Wang Zhao, Jisheng Li, Yong-Jin Liu, Yuxing Han, and Jiangtao Wen. DarkFeat: Noise-robust feature detector and descriptor for extremely low-light RAW images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 826–834, 2023.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[15] Minseong Kweon, Janghyun Kim, Ukcheol Shin, and Jinsun Park. MrGS: Multi-modal radiance fields with 3D gaussian splatting for RGB-Thermal novel view synthesis. *arXiv preprint arXiv:2511.22997*, 2025.

[16] Byeongjun Kwon and Munchurl Kim. Multi-modal depth estimation from misaligned thermal and RGB images. In *Proceedings of the Korean Institute of Broadcast and Media Engineers Summer Conference*, pages 912–915, 2024.

[17] 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.

[18] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European conference on computer vision*, pages 71–91. Springer, 2024.

[19] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[20] 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)*, pages 1–12. IEEE, 2024.

[21] Hao Lu, Tianshuo Xu, Wenzhao Zheng, Yunpeng Zhang, Wei Zhan, Dalong Du, Masayoshi Tomizuka, Kurt Keutzer, and Yingcong Chen. Drivingrecon: Large 4d gaussian reconstruction model for autonomous driving. *arXiv preprint arXiv:2412.09043*, 2024.

[22] 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.

[23] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025.

[24] Parv Maheshwari, Jay Karhade, Yogesh Chawla, Isaiah Adu, Florian Heisen, Andrew Porco, Andrew Jong, Yifei Liu, Santosh Pitla, Sebastian Scherer, et al. AnyThermal: Towards learning universal representations for thermal perception. *arXiv preprint arXiv:2602.06203*, 2026.

[25] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025.

[26] Sriram Narayanan, Mani Ramanagopal, and Srinivasa Narasimhan. Dual Band Thermal Videography: Separating Time-Varying Reflection and Emission Near Ambient Conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 199–208, 2026.

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[28] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[29] Yinrui Ren, Jinjing Zhu, Kanghao Chen, Zhuoxiao Li, Jing Ou, Zidong Cao, Tongyan Hua, Peilun Shi, Yingchun Fu, Wufan Zhao, et al. EventVGGT: Exploring cross-modal distillation for consistent event-based depth estimation. *arXiv preprint arXiv:2603.09385*, 2026.

[30] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38(1): 35–44, 2004.

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

[32] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016.

[33] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017.

[34] Dharmendra Selvaratnam and Dena Bazazian. 3D reconstruction in robotics: A comprehensive review. *Computers & Graphics*, 130:104256, 2025.

[35] Ukcheol Shin, Kyunghyun Lee, Seokju Lee, and In So Kweon. Self-supervised depth and ego-motion estimation for monocular thermal video using multi-spectral consistency loss. *IEEE Robotics and Automation Letters*, 7(2):1103–1110, 2021.

[36] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[37] Ukcheol Shin, Kyunghyun Lee, and Jean Oh. Bridging spectral-wise and multi-spectral depth estimation via geometry-guided contrastive learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6299–6305. IEEE, 2025.

[38] Vsevolod Skorokhodov, Chenghao Xu, Shuo Sun, Olga Fink, and Malcolm Mielle. SEAR: Simple and efficient adaptation of visual geometric transformers for RGB+Thermal 3D reconstruction. *arXiv preprint arXiv:2603.18774*, 2026.

[39] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[40] Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, et al. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026.

[41] Rubén Usamentiaga, Pablo Venegas, Jon Guerediaga, Laura Vega, Julio Molleda, and Francisco G Bulnes. Infrared thermography for temperature measurement and non-destructive testing. *Sensors*, 14(7):12305–12348, 2014.

[42] Michael Vollmer and Klaus-Peter Möllmann. *Infrared Thermal Imaging: Fundamentals, Research and Applications*. Wiley, August 2010. ISBN 9783527630868. doi: 10.1002/9783527630868. URL `http://dx.doi.org/10.1002/9783527630868`.

[43] Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025.

[44] 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*, pages 5294–5306, 2025.

[45] 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*, pages 20697–20709, 2024.

[46] Xiaoshan Wu, Yifei Yu, Xiaoyang Lyu, Yihua Huang, Bo Wang, Baoheng Zhang, Zhongrui Wang, and Xiaojuan Qi. EAG3R: Event-augmented 3D geometry estimation for dynamic and extreme-lighting scenes. *arXiv preprint arXiv:2512.00771*, 2025.

[47] Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 4730–4738, 2021.

[48] Jiuhong Xiao, Roshan Nayak, Ning Zhang, Daniel Tortei, and Giuseppe Loianno. ThermalGen: Style-disentangled flow-based generative models for RGB-to-Thermal image translation. *arXiv preprint arXiv:2509.24878*, 2025.

[49] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025.

[50] Sizhe Yang, Linning Xu, Hao Li, Juncheng Mu, Jia Zeng, Dahua Lin, and Jiangmiao Pang. Robo3r: Enhancing robotic manipulation with accurate feed-forward 3d reconstruction. *arXiv preprint arXiv:2602.10101*, 2026.

[51] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[52] Seungsang Yun, Minwoo Jung, Jeongyun Kim, Sangwoo Jung, Younghun Cho, Myung-Hwan Jeon, Giseop Kim, and Ayoung Kim. STheReO: Stereo thermal dataset for research in odometry and mapping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3857–3864. IEEE, 2022.

[53] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

[54] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Qinghua Hu, Cheng Deng, Cai Xu, Jie Wen, Di Hu, et al. Multimodal fusion on low-quality data: A comprehensive survey. *Information Fusion*, page 104437, 2026.

[55] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. FLARE: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21936–21947, 2025.

[56] Xingxing Zuo, Nikhil Ranganathan, Connor Lee, Georgia Gkioxari, and Soon-Jo Chung. MonoTher-Depth: Enhancing thermal depth estimation via confidence-aware distillation. *IEEE Robotics and Automation Letters*, 10(3):2830–2837, 2025.

A Methodological details

A.1 Architecture and adaptation

LoRA and camera tokens. After loading the pretrained VGGT [44] checkpoint, we initialize the thermal camera token by duplicating the pretrained camera token, following the same token-initialization strategy as SEAR [38]. This allows both modalities to start from the same geometric token prior. We then insert rank-64 LoRA [14] adapters, shared across the RGB and thermal streams, into the `qkv`, `proj`, `fc1`, and `fc2` projections of both the frame-attention and global-attention blocks, while keeping the original backbone weights frozen. The parameter counts in Table 4 refer to learnable parameters. The VGGT aggregator contains 24 alternating-attention rounds, each comprising one frame-attention and one global-attention block, for 48 blocks in total. The input-plus-output dimensions of the adapted projections sum to 16,384 per block, yielding $48 \times 64 \times 16{,}384 = 50{,}331{,}648 \approx 50.33$M learnable LoRA parameters.

Late geometry-shared thermal routing. The geometry-shared decomposition uses ranks $(d_s, d_p, d_{\mathrm{mi}}) = (96, 192, 128)$, with the distillation projection heads in Eq. 8 implemented as single linear maps $\mathbb{R}^{d_s} \to \mathbb{R}^{d_{\mathrm{mi}}}$. The pathway is instantiated in the last $k = 8$ AA blocks, with the residual scale initialized to $\alpha_{\mathrm{gstr}} = \sigma(-2.2) \approx 0.1$ and the corrective injection and gating heads zero-initialized, so the branch is effectively inactive at initialization and is activated only as training learns a consistent corrective signal. The reconstruction and alignment losses $\mathcal{L}_{\mathrm{recon}}$ and $\mathcal{L}_{\mathrm{distill}}$ follow a cosine warmup from 0.01 to 0.10 over the first 1500 of 4000 gradient-accumulation steps. Without this warmup, applying the target weight from step zero pulls the private subspace prematurely toward thermal reconstruction and degrades early depth/pose accuracy.

Thermal positional adaptation. To account for residual spatial offsets between RGB and thermal views, we introduce a learnable 2D thermal RoPE [39] offset. Thermal tokens share the rotary positional grid of RGB tokens, with an additive offset $(\Delta y, \Delta x)$ applied before rotary encoding. The offset is initialized to zero, allowing the multimodal model to start from the aligned RGB prior and learn cross-sensor positional corrections only when supported by the training signal. This lightweight mechanism preserves the RGB branch while adapting only the thermal positional coordinates.

A.2 Geometry-Shared Thermal Routing visualization

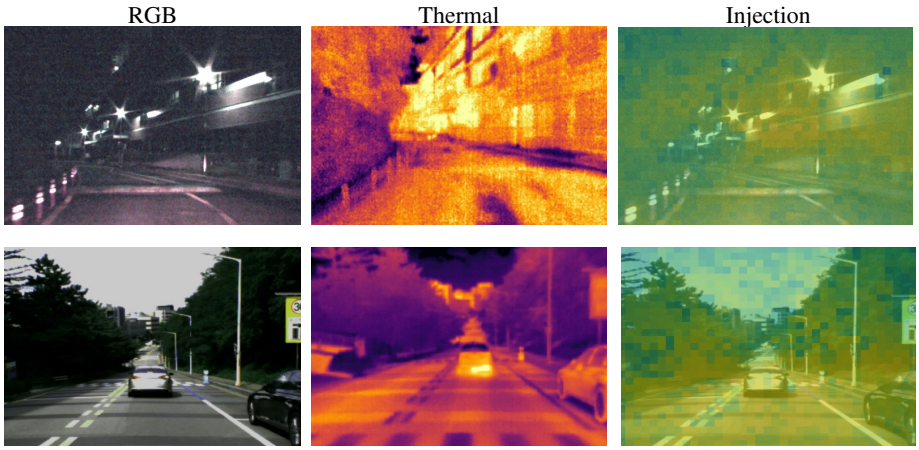

Figure 5: Reliability-gated injection samples during training in dark and light scenes.

Figure 5 visualizes the per-token ℓ_2 norm of the gated geometry-shared token injection $\boldsymbol{g}_{\mathrm{s}}^{(\ell)} \odot W_{\mathrm{up}}\big(\mathrm{sg}[\bar{\mathbf{u}}^{\mathrm{s},(\ell)}] - \bar{\mathbf{v}}^{\mathrm{s},(\ell)}\big)$ from Eq. 11, evaluated on paired RGB-T inputs sampled from day and night training scenes. In the dark night scene shown in the top row, the RGB image is severely underexposed and dominated by saturated headlights, while the corresponding thermal frame depicts the road geometry, surrounding buildings, and structural edges clearly. The geometry-shared routing magnitude in this case concentrates on the unlit road surface and the building facades on either side,

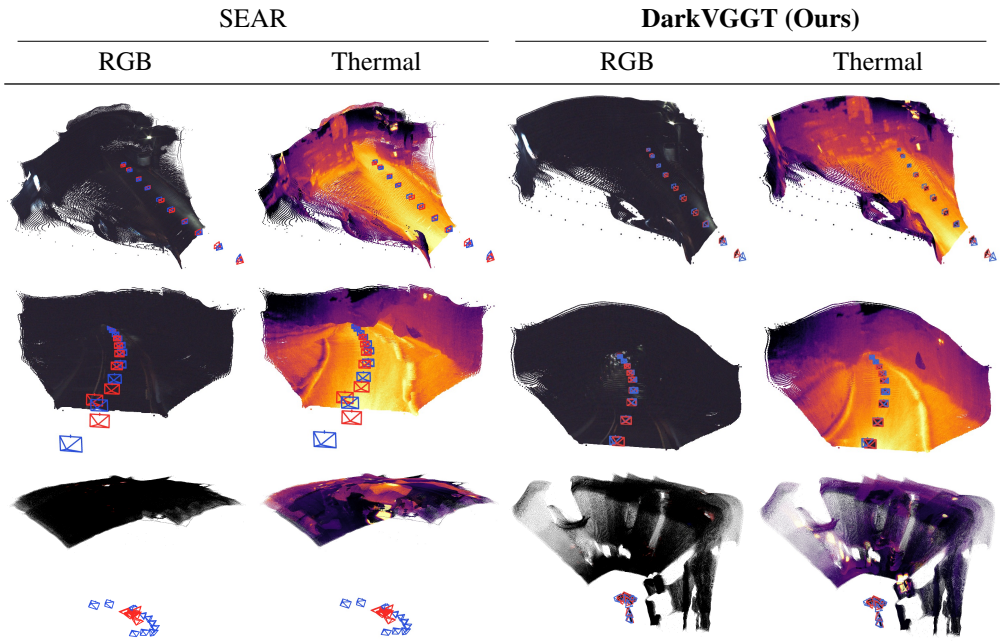

Figure 6: Qualitative comparison between SEAR and our method. Blue and red cameras represent ground-truth and predicted camera poses, respectively.

which are precisely the regions where the RGB pathway carries little usable information and the thermal stream is informative. In the well-lit daytime scene shown in the bottom row, where the RGB image is properly exposed and rich in texture, the injection magnitude remains uniformly low across the entire field of view, with no regions of pronounced activation. At the deepest active AA block, the per-token feature-conditioned gate $\boldsymbol{g}_{\mathrm{s}}^{(\ell)} \in [0,1]^P$ modulates the shared-thermal residual before RGB-token injection. This contrast suggests that the gate suppresses geometry-shared residual injection when RGB cues are sufficient and selectively routes thermal-recovered structure where RGB fails.

A.3 Edge supervision and target sensitivity

The depth edge map used by $\mathcal{L}_{\mathrm{edge}}$ is computed from ground-truth depth using Sobel filters. Thermal and depth edge magnitudes are pooled over each patch to construct the soft target $a_{\mathrm{h,tar}}^{(\ell)}$. This supervision is used only during training.

Sensitivity of the edge target. The edge-disagreement target uses sharpness $\beta = 8$ and margin $\tau = 0.15$. For positive β, changing sharpness preserves the ordering of the disagreement scores and the target's 0.5 decision boundary, while changing target confidence. In the sensitivity evaluation, the measured Spearman rank correlation is at least 0.999. Halving sharpness to $\beta = 4$ increases target entropy by approximately $1.8\times$. Doubling the margin to $\tau = 0.30$ changes the fraction of flagged patches by at most 8.5 percentage points. Retraining with these changes individually alters AbsRel by at most 0.0015 on each benchmark, with aggregate AbsRel changes of -0.0005 and -0.0008, respectively. These small changes support robustness to the tested sharpness and margin settings.

B Additional experiments and analysis

B.1 DarkVGGT vs. SEAR

Figure 6 compares additional dense 3D point-cloud reconstruction results between SEAR [38], an RGB-T feed-forward foundation model, and our proposed DarkVGGT on outdoor and indoor scenes.

Table 6: Controlled input and fusion comparison on a separate ViViD++ split. Fine-tuned methods share the split, schedule, and supervision. Dark and light columns report results on the three dark and two daylight test sequences, respectively. The top three values across all rows in each metric column are marked as 1st, 2nd, and 3rd.

		Dark		Light	
Input / method	Setup	AbsRel↓	AUC@30↑	AbsRel↓	AUC@30↑
RGB-only (raw)	Zero-shot	0.2746	25.9	0.0977	71.6
RGB-only + Enh	Zero-shot	0.2841	24.9	0.0959	72.1
RGB-only + Enh	Fine-tuned	0.1736	28.8	0.0729	73.0
Thermal-only	Zero-shot	0.2254	57.4	0.2413	53.3
Thermal-only	Fine-tuned	0.0940	64.9	0.0914	73.7
Naive multimodal	Fine-tuned	0.1722	29.1	0.0718	73.6
DarkVGGT (full)	Fine-tuned	0.1229	48.1	0.0732	76.9

In the nighttime outdoor sequences shown in the top two rows, both methods produce comparable point-map quality, likely because weak illumination cues such as streetlights remain available, allowing LoRA-based RGB-T adaptation to exploit residual RGB structure. However, the camera-pose results show that DarkVGGT estimates trajectories closer to the ground truth than SEAR.

The gap becomes more pronounced in dark indoor scenes, where light and texture cues are nearly absent and effective fusion of thermal information into degraded RGB embeddings becomes crucial. Under these conditions, SEAR struggles to recover reliable geometry. In contrast, with the proposed thermal-to-RGB fusion modules, DarkVGGT robustly captures object-level depth variations and reconstructs plausible 3D geometry even in extremely dark indoor environments.

B.2 Selected CTFF-only and full comparisons

We examine selected sequence-level results to contextualize the aggregate comparison between LoRA+CTFF and the full model in Table 4. Adding GSTR reduces average AbsRel from 0.1422 to 0.1373 and increases RTA@30 from 0.86 to 0.89, while RRA@30 decreases from 1.00 to 0.99 and AUC@30 decreases from 68.9 to 66.3. Thus, the full model improves depth and translation accuracy at the reported threshold but does not uniformly improve pose performance.

On `indoor_aggresive_dark`, adding GSTR increases RTA@30 from 0.3733 to 0.4222 (13.1% relative improvement), with the full model reaching AUC@30 of 19.3. On `indoor_unstable_dark`, AbsRel decreases from 0.1251 to 0.1228 and RTA@30 increases from 0.8259 to 0.8481 (2.7% relative improvement). On STheReO, LoRA+CTFF and the full model achieve similar dataset-average AbsRel values of 0.1613 and 0.1600, respectively.

These selected results demonstrate local improvements alongside a lower aggregate AUC@30. Since AUC@30 integrates joint pose accuracy over thresholds, improvements in RTA@30 alone do not imply an improvement in AUC@30. This analysis is limited to the selected sequences and dataset average above.

B.3 Controlled input and fusion baselines

Table 6 compares input modalities and fusion strategies on a separate controlled ViViD++ split. The test scenes are `indoor_aggresive_dark`, `indoor_unstable_dark`, `outdoor_robust_night2`, `indoor_unstable_global`, and `outdoor_robust_day2`. The last two scenes are moved from the main experiment's training split to evaluation. All fine-tuned rows use the same split, training schedule, and ground-truth supervision. DarkVGGT is retrained for this comparison. These results are not evaluations of the main experiment's checkpoint. "Enh" denotes CLAHE enhancement of RGB inputs. The naive multimodal baseline fine-tunes the AA blocks.

Zero-shot RGB enhancement degrades performance on the dark split, increasing aggregate AbsRel from 0.2746 to 0.2841 and reducing AUC@30 from 25.9 to 24.9. Under matched fine-tuning, naive multimodal adaptation remains close to enhanced RGB-only (0.1722 versus 0.1736 AbsRel), whereas

Table 7: Per-sequence performance on Dark3R dataset over 10 frame-sampling trials (mean $\pm$ standard deviation).

Scene	AUC@30↑	RRA@30↑	RTA@30↑	AbsRel↓	$\delta < 1.25 \uparrow$	$\mathrm{RMSE}_{\log} \downarrow$
`chapel`	64.9 ± 8.2	1.000 ± 0.000	0.908 ± 0.055	0.114 ± 0.013	0.824 ± 0.046	0.163 ± 0.016
`kitchen`	46.7 ± 3.5	0.998 ± 0.004	0.959 ± 0.022	0.125 ± 0.006	0.868 ± 0.020	0.157 ± 0.006

Table 8: Quantitative comparison of depth estimation on well-lit RGB-T ViViD++ sequences. This shows that our thermal-to-RGB fusion method remains effective even in well-lit scenes.

	`city_day1`			`city_day2`			`campus_day2`		
Method	AbsRel↓	$\delta < 1.25 \uparrow$	$\mathrm{RMSE}_{\log} \downarrow$	AbsRel↓	$\delta < 1.25 \uparrow$	$\mathrm{RMSE}_{\log} \downarrow$	AbsRel↓	$\delta < 1.25 \uparrow$	$\mathrm{RMSE}_{\log} \downarrow$
MASt3R [18]	0.3133	0.5769	0.4079	0.2785	0.4284	0.3720	0.1613	0.8154	0.2548
Dark3R [9]	0.3479	0.4617	0.4703	0.4187	0.3587	0.5548	0.2268	0.6811	0.3157
VGGT [44]	0.1739	0.7395	0.2661	0.1882	0.7281	0.3371	0.1365	0.8639	0.2474
DarkVGGT (ours)	0.1786	0.7496	0.2550	0.1942	0.7139	0.3235	0.1502	0.8153	0.2328

DarkVGGT reaches 0.1229 AbsRel and 48.1 AUC@30. Relative to naive multimodal adaptation, this is a 28.6% reduction in dark AbsRel.

Thermal-only achieves better performance on the dark split. Its fine-tuned variant achieves 0.0940 AbsRel and 64.9 AUC@30, and its zero-shot variant also exceeds DarkVGGT in dark AUC@30 (57.4 versus 48.1). Thermal-only requires thermal input and cannot directly operate on RGB-only benchmarks such as ETH3D and ScanNet++. Where thermal input is available in daylight, fine-tuned thermal-only has approximately 25% higher AbsRel than DarkVGGT (0.0914 versus 0.0732). DarkVGGT recovers approximately 64% of the dark AbsRel gap between fine-tuned enhanced RGB-only and thermal-only, while remaining close to enhanced RGB-only in daylight AbsRel (0.0732 versus 0.0729) and achieving the highest daylight AUC@30 in this comparison (76.9). Naive multimodal adaptation achieves lower daylight AbsRel than DarkVGGT (0.0718 versus 0.0732). These results characterize a trade-off across illumination and sensor availability.

B.4 Dark3R frame-sampling variance

Dark3R provides a complementary evaluation using synthesized thermal inputs. The `chapel` and `kitchen` scenes are held out from training. Table 7 reports per-sequence means and standard deviations over ten frame-sampling trials, supplementing the deterministic evaluation used in the main benchmark tables. These standard deviations characterize sensitivity to frame selection, not variability across independently trained models.

AUC@30 exhibits different sampling variability across the two scenes (standard deviations 8.2 and 3.5), whereas the reported AbsRel and RRA@30 standard deviations are at most 0.013 and 0.004, respectively. Pairwise pose evaluation aggregates $\binom{N}{2}$ frame pairs for N sampled frames, so a difficult segment can affect multiple pairs simultaneously. This offers a possible explanation for sensitivity to frame selection, but the sampling experiment alone does not isolate its cause or establish training stability. These estimates are based on two held-out Dark3R scenes.

B.5 Daylight scene performance on RGB-T datasets

To verify that DarkVGGT's RGB-T fusion strategy remains effective under daylight conditions, we further evaluate depth estimation performance on three held-out well-lit sequences in Table 8. We use the `city_day1`, `city_day2`, and `campus_day2` sequences from ViViD++ [17] dataset. Since the extrinsics of these sequences are released as identity matrices, we additionally evaluate only depth estimation in this ablation study.

DarkVGGT achieves the lowest $\mathrm{RMSE}_{\log}$ on all three held-out sequences. Averaged over the three sequences, DarkVGGT reaches 0.1743 AbsRel, 0.7596 $\delta < 1.25$, and 0.2704 $\mathrm{RMSE}_{\log}$, compared with VGGT's 0.1662, 0.7772, and 0.2835. DarkVGGT improves $\mathrm{RMSE}_{\log}$ at the cost of slightly worse average AbsRel and $\delta < 1.25$, while largely preserving VGGT's RGB geometry performance.

Table 9: Dataset split statistics used in the main experiments.

Dataset	Train Sequences	Train Frames	Eval Sequences	Eval Chunks	Eval Sampled Frames
ViViD++ [17]	11	6,691	3	31	310
STheReO [52]	4	9,837	2	60	600
Dark3R [9]	6	2,260	2	2	50
Total	21	18,788	7	93	960

C Experimental setup

C.1 RGB-T dataset processing

Table 9 summarizes the number of samples used for our overall training and evaluation.

ViViD++. We train on 11 ViViD++ scenes and evaluate on three disjoint dark scenes. The training split includes `indoor_robust_dark`, `outdoor_robust_night1`, `indoor_aggresive_global`, `indoor_robust_global`, `indoor_unstable_global`, `outdoor_robust_day1`, `outdoor_robust_day2`, `indoor_robust_varying`, `indoor_aggresive_local`, `indoor_robust_local`, and `indoor_unstable_local`, while the held-out set contains `indoor_aggresive_dark`, `indoor_unstable_dark`, and `outdoor_robust_night2`. Sequence-mode training uses all frames from the selected scenes. We use refined depth maps `Depth_RGB_refined` and `Depth_T_refined` for training. Thermal images are percentile-clipped, min-max normalized, enhanced with CLAHE [30], and colorized with the Inferno colormap before being passed to the frozen DINOv2 encoder, while the grayscale thermal image is retained for edge-disagreement supervision. Both modalities are resized to a common resolution under the same crop/resize transform.

STheReO. We train on four STheReO scenes, `snu_evening`, `kaist_afternoon`, `snu_afternoon`, and `valley_afternoon`, comprising 44,612 raw frames in total, and evaluate on `kaist_evening` and `valley_evening` (14,586 raw frames). We use 9,837 training frames and 13,128 evaluation frames after trimming 5% from each evaluation sequence end. Sequence-mode evaluation uses strict non-overlapping K=50 windows capped at 30 per scene, yielding 60 windows and 600 sampled frames. Raw STheReO sequences are converted into the canonical ViViD++ layout. For each RGB timestamp, we match the nearest thermal frame and LiDAR scan, interpolate the LiDAR-body pose to obtain camera-to-world poses, and undistort both modalities. Sparse depth maps are obtained by projecting the matched LiDAR scan into each camera, with optional dense refinements stored under `Depth_RGB_refined` and `Depth_T_refined`. During training, we disable static load-time subsampling and instead randomize the temporal spacing between nearby sampled frames with a stride drawn from $[2, 4]$.

Dark3R. We train on the clean and dark variants of `university_college`, `quad`, and `fireplace` (six sequences) and evaluate on `dark_chapel` and `dark_kitchen` under the public Dark3R protocol. Figure 7 shows our RGB-T conversion. As Dark3R lacks a calibrated dual-sensor rig, each exposure bracket yields a clean–dark pair with shared poses and released dense metric ground-truth depth. Clean RGB uses the cleanest sRGB image from the depth-annotation ISO folder, while Dark RGB uses the noisiest exposure ($\sim 40\times$ darker). ThermalGen [48] generates paired pseudo-thermal supervision, and camera poses are converted from NeRF to OpenCV.

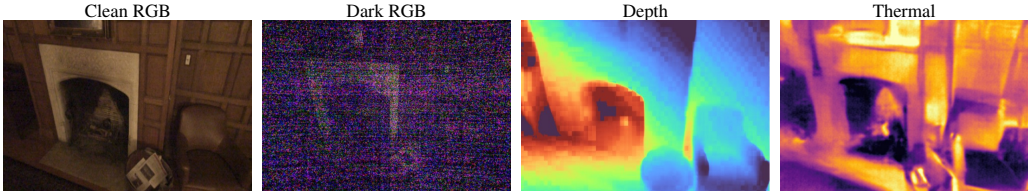

Figure 7: Preprocessed Dark3R dataset training samples.

C.2 Lingbot-Depth supervision and its scope

For ViViD++ and STheReO, Lingbot-Depth [40] is used as an off-the-shelf completion tool to densify sparse LiDAR measurements into pixel-dense training targets and reduce optimization difficulties associated with sparse supervision. DarkVGGT and the matched fine-tuned SEAR baseline use the same refined supervision, whereas evaluation uses raw LiDAR ground truth rather than the refined targets. This parity applies to the matched training comparison, not to the off-the-shelf baselines.

Lingbot-Depth can introduce learned priors into the training targets. The contribution of depth refinement relative to training on raw sparse LiDAR, including any architecture-dependent effects, remains unquantified.

C.3 Hyperparameters

Table 10 summarizes the key training hyperparameters used in DarkVGGT. We initialize the model from the public VGGT-1B checkpoint [44] and keep the backbone frozen throughout training. Multimodal adaptation is handled by the lightweight modules listed in Table 10, while the GSTR residual is initialized conservatively to avoid perturbing the pretrained RGB prior.

The auxiliary losses are kept small relative to the main VGGT supervision. The CTFF regularizers remain weak, the GSTR reconstruction and distillation terms use cosine warmup, and the decorrelation term is delayed and ramped gradually. In the final configuration, we omit the explicit patch-level correction loss and apply RGB-prior preservation only on thermal-dropout steps. Thermal dropout is restricted to batches whose clips all come from a predefined set of daylight training sequences, identified by scene-name matching, yielding an effective dropout rate of about 25%. At inference, no scene-name rules are used. The learned feature-conditioned gates modulate thermal injection.

Table 10: Key training hyperparameters for DarkVGGT.

Architecture / data	
Backbone initialization	VGGT-1B, pretrained checkpoint
Image size / patch size	518 / 14
Clip lengths	$\{2, 3, 4, 6, 12\}$
Max images per GPU	24
Optimization	
Optimizer	AdamW
Learning rate	5×10^{-5}
Weight decay	10^{-2}
Schedule	10% linear warmup, 90% cosine decay
Mixed precision	bfloat16
Gradient accumulation	2
Epochs / steps	10 / 2000
Effective steps (accum. $\times$ steps)	$4000 = 2 \times 2000$
Adaptation modules	
LoRA rank / alpha	64 / 128
Geometry-shared ranks	$(d_s, d_p, d_{mi}) = (96, 192, 128)$
GSTR active blocks	$k = \mathbf{8}$
GSTR residual init α_{gstr}	$\sigma(-2.2) \approx 0.10$
Loss weights	
Camera $\mathcal{L}_{\mathrm{cam}}$	5.0
Depth $\mathcal{L}_{\mathrm{depth}}$ (log-norm)	1.0
Residual allocation $\mathcal{L}_{\mathrm{sparse}}$	0.03
Edge disagreement $\mathcal{L}_{\mathrm{edge}}$	0.01
Orthogonality $\mathcal{L}_{\mathrm{ortho}}$	0.003
RGB-prior preservation $\mathcal{L}_{\mathrm{drop}}$	0.5
GSTR recon $\mathcal{L}_{\mathrm{recon}}$	$0.01 \rightarrow 0.10$ (cosine, step 0–1500 of 4000)
GSTR distill $\mathcal{L}_{\mathrm{distill}}$	$0.01 \rightarrow 0.10$ (cosine, step 0–1500 of 4000)
GSTR decorrelation $\mathcal{L}_{\mathrm{decorr}}$	$0 \rightarrow 2 \times 10^{-5}$ (cosine, step 1000–3000 of 4000)
Multimodal training	
Thermal dropout	$p = 0.5$, RGB-dominant batches only

C.4 Evaluation protocol

All main benchmark results are reported in sequence mode on held-out scenes only. For each held-out scene, we form strict non-overlapping $K{=}50$-frame windows and uniformly sample $N{=}10$ frames per window, capped at 30 windows per scene. ViViD++ scenes are short and never reach this cap (up to 20 windows per scene), whereas STheReO sequences are long driving recordings for which the cap is active, yielding 30 windows per scene. For the Dark3R dataset, we use a single deterministic chunk per scene, sampling the full scene with a stride of 20 to obtain approximately 25 frames. Depth metrics are averaged over the sampled evaluation frames, while pose metrics are pooled over all frame pairs within each sampled window.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the main contributions, including the proposed RGB-T geometry framework, the two fusion modules, and the evaluation settings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Section 5, including the assumption of moderately aligned paired RGB-T inputs.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper is about application in computer vision. We do not have explicit theorems.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have detailed description in Appendix and provide a zip file for code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets used in this work are publicly available, and anonymized code is attached as a zip file. Detailed preprocessing procedures are described in Appendix C.1.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: See Appendix A.1, C.1, C.3 for detailed recipe of experimental setting.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: 3D feed-forward models are geometry foundation models. Usually does not require error bar to run it for multiple times. We include diverse scenes to show generalizibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments were conducted on NVIDIA four A100 GPUs (80GB). Training for 10 epochs takes approximately 7 hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe that the work is consistent with its guidelines. Our study is limited to RGB-T 3D geometry estimation and does not involve human-subject experiments, identifiable information, or applications designed to cause harm.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The future work in the Section 5 discusses benefits for low-visibility robotic perception and privacy concerns from thermal sensing.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk data or models requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the sources for all existing models and datasets used in the paper, and include usage information in the appendix where publicly available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release the code and model assets with documentation covering usage and training details.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is only used for polishing English and code debugging.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.