# SEAR: Simple and Efficient Adaptation of Visual Geometric Transformers for Unpaired RGB-Thermal 3D Reconstruction

Vsevolod Skorokhodov[1], Chenghao Xu[2], Shuo Sun[3], Olga Fink[2], and Malcolm Mielle[1]

[1] Schindler EPFL Lab, Lausanne, Switzerland
[2] EPFL, Lausanne, Switzerland
[3] Örebro University, Örebro, Sweden

**Abstract.** Foundational feed-forward visual geometry models enable accurate and efficient camera pose estimation and scene reconstruction by learning strong scene priors from massive RGB datasets. However, their effectiveness drops when applied to mixed sensing modalities, such as RGB-thermal (RGB-T) images. We observe that while a visual geometry grounded transformer pretrained on RGB data generalizes well to thermal-only reconstruction, it struggles to align RGB and thermal modalities when processed jointly. To address this, we propose SEAR, a simple yet efficient fine-tuning strategy that adapts a pretrained geometry transformer to multimodal RGB-T inputs. Despite being trained on a relatively small RGB-T dataset, our approach significantly outperforms state-of-the-art methods for 3D reconstruction and camera pose estimation, achieving significant improvements over all metrics and delivering higher detail and consistency between modalities with negligible overhead in inference time compared to the original pretrained model. Notably, SEAR enables reliable multimodal pose estimation and reconstruction even under challenging conditions, such as low lighting and dense smoke. We validate our architecture through extensive ablation studies and demonstrate how the model aligns both modalities. Additionally, we introduce a new dataset featuring RGB and thermal sequences captured at different times, viewpoints, and illumination conditions, providing a robust benchmark for future work in multimodal 3D scene reconstruction. Code and models are publicly available at https://doi.org/10.5281/ZENODO.21077295.

**Keywords:** Multimodal Pose Estimation · Thermal imagery · Feed-Forward Neural Network

## 1 Introduction

Recent work introduced large-scale 3D geometric foundation models as adaptable solutions for in-the-wild 3D vision tasks. Unlike traditional photogrammetry pipelines—which depend on feature matching and iterative optimization—these models use transformers to predict camera poses and scene structure from sparse

RGB inputs in a single feed-forward pass [41]. Pretrained on diverse datasets, they achieve strong cross-scene generalization, addressing classical limitations like computational inefficiency and sensitivity to image quality and environmental conditions. Despite these advances, current geometric foundation models rely solely on RGB inputs, limiting their robustness in real-world scenarios where complementary modalities are needed. E.g., thermal cameras capture long-wave infrared radiation, enabling applications such as low-light reconstruction [46], thermal simulation [2], and infrastructure inspection [9].

When applying a pretrained feed-forward model (e.g., VGGT [41]) to mixed RGB-thermal inputs, we observe independent reconstructions that fail to align into a coherent 3D representation (see Fig. 3). This indicates that pretrained models encode strong geometric priors, but lack explicit cross-modal consistency for pose estimation and reconstruction. Most prior works address this issue via either synchronous data collection [9, 2, 23, 20], or post-hoc alignment [38]. In simultaneous capture, using paired images from both modalities, a "strong" modality (e.g., RGB) infers poses for a "weaker" one (e.g., thermal) using classical pipelines like COLMAP [33]. In practice, synchronous captures necessitate complex or expensive sensor setups, and real-world data is often asynchronous (e.g., thermal imagery at sunrise for steady-state facades [44] vs. RGB at daytime for visual clarity [9]). Post-hoc alignment methods, on the other hand, use cross-modal feature matching to align poses/images using specialized descriptors; those methods are thus sensitive to image, feature, and match quality, leading to lower robustness. Thus, estimating consistent multimodal camera poses and 3D scene reconstruction in realistic scenarios is still a challenge.

In this work, we address camera pose estimation and scene reconstruction from unpaired RGB and thermal (RGB-T) images. Our key insight is that, since VGGT performs well on each modality independently when processed jointly but lacks geometric consistency, large-scale multimodal retraining should be unnecessary. Instead, a pretrained model can be adapted with minimal parameter updates to bridge the modality gap. Our contributions are threefold:

- We propose SEAR, a lightweight fine-tuning strategy for cross-modal pose estimation and 3D reconstruction, based on LoRA-based adapters and a specific batching strategy. Our method requires fewer than 5% of the original model's parameters and preserves the inference speed of the base model.
- We show that SEAR only requires a modest multimodal dataset for tuning (∼15,000 pairs of RGB and thermal images), making it practical for real-world applications where collecting large multimodal datasets is challenging.
- We present a new dataset comprising 9 scenes (1,890 images) with distinct RGB/thermal trajectories captured under varying illumination and viewpoints. This dataset provides a new benchmark for evaluating RGB-T cross-spatial multimodal reconstruction in challenging conditions.

We demonstrate large improvements against eight state-of-the-art baselines in pose estimation and point cloud reconstruction metrics. Complementary ablation studies further support our key design choices.

## 2 Related Works

RGB-T imaging enables low-light scene mapping and non-invasive infrastructure inspection (e.g., detecting thermal defects via finite element analysis [2]). Though NeRF [9, 19] and Gaussian Splatting [23, 22] have been extended to RGB-T data, they depend on known camera poses—typically estimated via SfM [33] or single-modality feed-forward networks, which fail for cross-modal datasets (see Fig. 3). Thus, most prior works either assume paired multimodal datasets, using RGB to obtain pose estimates for other modalities [9, 2, 23] or are limited by the complexity of performing feature matching between the two modalities [3].

To estimate camera poses and scene reconstruction, traditional 3D reconstruction relies on iterative optimization (e.g., SfM [33] or SLAM [36]), on the detection of salient features (e.g., ORB [27]), and on computationally expensive matching algorithms (e.g., PnP [45] or RANSAC [7]). While these methods can struggle with robustness and generalization, recent feed-forward models use transformers to predict camera poses and 3D geometry in a single pass [42, 17, 41], achieving strong zero-shot generalization. However, these methods are trained on unimodal inputs (e.g., RGB), and their use for applications where multimodal data (e.g., thermal, depth) is necessary for robustness remains limited.

Retraining large models for multimodal tasks is computationally prohibitive for most machine learning practitioners. Furthermore, RGB and thermal datasets for scene reconstruction are scarce (e.g., ThermoNeRF's 20 scenes [9]), with larger RGB+thermal datasets focusing on segmentation/navigation. Parameter-efficient fine-tuning (PEFT) methods address this by reducing the number of trainable parameters while preserving performance.

Key approaches include Adapter Layers [11] (lightweight modules inserted between transformer layers), Prompt Tuning [18] (optimizes continuous prompts in embedding space), and Low-Rank Adaptation (LoRA) [12] (decomposes weight updates into low-rank matrices). While PEFT has been applied in NLP [12] and visual segmentation [24], existing work focuses on task specialization—not modality expansion. No prior work explores PEFT to adapt pretrained 3D geometric models to multimodal reconstruction. By bridging this gap, our work lays the foundation for resource-efficient, generalizable multimodal reconstruction.

## 3 Preliminaries: VGGT

VGGT [41] is a transformer-based model able to estimate, from a sequence of $N$ RGB images observing a 3D scene, the intrinsic and extrinsic camera parameters, a depth map, a point map, and a grid of $C$-dimensional features for point tracking. DINOv2 [28] performs feature extraction under the hood, producing a sequence of tokens for each input frame. After that, a learnable *camera token* is concatenated to each set of tokens of each frame. The resulting token sequence is processed by 24 self-attention blocks, with alternating (frame-wise and global) attention (AA) blocks. The tokens from the last layer are passed to the camera parameters prediction head, while intermediate tokens from layers 4, 11, 17, and 23 are

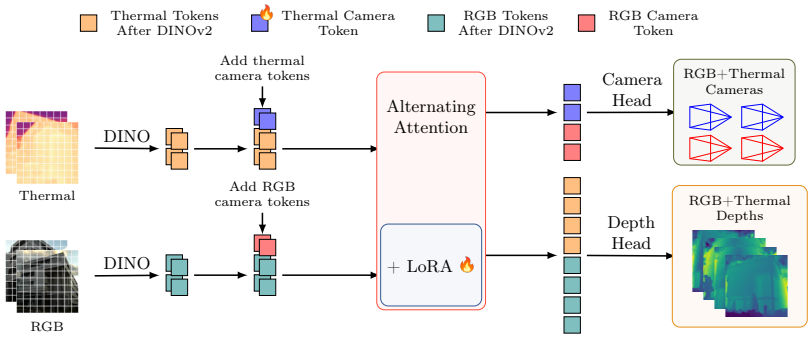



**Fig. 1:** SEAR architecture. RGB and thermal images are first tokenized using DINOv2. For each modality, camera-specific tokens are concatenated with the corresponding DINO tokens. The combined tokens are then processed by an Alternating-Attention (AA) module with LoRA adapters. Finally, the refined tokens are passed to separate prediction heads for camera parameter estimation and depth estimation. Trainable parameters are highlighted with a flame symbol.

concatenated and passed to depth, point map, and tracking prediction heads. For more details, refer to the original paper [41].

## 4 Methodology

While VGGT can be used to estimate RGB or thermal 3D scenes, estimated RGB-T scenes often result in two disjoint reconstructions (one per modality), misaligned in pose and scale (see Fig. 3). We hypothesize that this knowledge gap can be closed by fine-tuning the pretrained model, drastically reducing computational cost compared to retraining.

We introduce a framework (see Fig. 1) to adapt VGGT [41] for joint estimation of RGB and thermal intrinsic/extrinsic camera parameters and depth maps. We introduce two innovations: 1) Lightweight LoRA adapters in the AA module to bridge the domain gap between RGB and thermal inputs (Sec. 4.1) 2) A batching strategy designed to prevent the model from relying on RGB-T image correspondences (Sec. 4.2). This simple, albeit effective, methodology, combined with a small amount of training data (Sec. 5), enables RGB-T scene reconstruction and camera pose estimation from unpaired multimodal RGB and thermal images, with negligible memory and inference overhead.

### 4.1 LoRA Integration

We fine-tune the RGB-pretrained VGGT model using LoRA to adapt it for mixed RGB+thermal inputs. LoRA modules are integrated into all linear and multi-head attention layers of the alternating-attention (AA) module. This approach preserves the model's pretrained RGB knowledge, since the original weights are

frozen; the original RGB and thermal data processing is left unchanged, pushing the model to focus on alignment between the modalities to reduce the loss during training. The features produced by the AA module are then processed by the prediction heads to estimate camera parameters and depth maps with their uncertainties. The adapter follows the default LoRA initialization scheme [12].

The original VGGT model uses DINOv2 [28] as a tokenizer and visual feature extractor for RGB images. For each frame, a camera token is appended to the image tokens, and the sequence is processed by the AA module. Prior work [6] has shown that while DINOv2 is able to extract meaningful thermal features without retraining, the feature distributions of the two modalities differ. Hence, we use DINOv2 as a tokenizer for both modalities and introduce two learnable thermal camera tokens—a counterpart to the two original VGGT camera tokens (now termed the RGB camera token)—to align both modalities' feature spaces before the AA module. These tokens encode the intrinsic/extrinsic features for thermal frames—the first token is for the first frame of the sequence, while the other is for all subsequent frames. RGB and thermal images are independently tokenized using DINOv2 before the RGB and learnable thermal camera tokens are respectively appended to the RGB and thermal tokens. Then, the augmented sequences are fed into the AA module. By assigning distinct camera tokens to each modality, the model learns modality-specific representations. Since these tokens are used in both frame-wise and global interactions in the AA module, they enable differentiated processing of RGB and thermal inputs. To preserve the pretrained model's behavior during early tuning, we initialize the thermal camera token with the RGB camera token's weights.

We optimize our model using the loss functions from VGGT, formulated as a multitask loss:

$$\mathcal{L} = \lambda_{camera}\mathcal{L}_{camera} + \mathcal{L}_{depth}. \tag{1}$$

Where $\mathcal{L}_{camera}$ is the Huber loss for camera pose estimation, $\mathcal{L}_{depth}$ is an uncertainty-aware loss for depth prediction, and $\lambda_{camera}$ is a weighting coefficient set to 5.0, as in the original VGGT model.

### 4.2 Data Pre-processing and Batching

We apply asymmetric augmentation pipelines for RGB and thermal inputs. For both RGB and thermal images, we apply random cropping, random aspect ratio adjustments (uniformly sampled from $[0.33, 1.0]$), Gaussian noise, Gaussian blur, and random sharpness adjustments. For RGB images, we further apply color jittering and grayscale conversion. For thermal images, we apply random linear transformations of pixel intensities followed by random exponential scaling of pixel values with degrees sampled uniformly in $[1/1.5, 1.5]$. Additionally, we apply random 90° rotations to both RGB and thermal images.

To ensure generalization to inputs of varying ratios and sizes, we construct batches of independent RGB and thermal images—i.e., without shared camera poses—sampled from a single RGB+thermal scene. This constraint prevents the model from relying on trivial RGB-thermal correspondences during training,

instead forcing it to learn inter-modal relationships across viewpoints. The ratio of thermal to RGB images $\tau$ within each batch is randomly sampled from a uniform distribution $U(0, 1)$, enabling the model to process arbitrary combinations of RGB and thermal inputs.

## 5 Datasets

### 5.1 Novel Challenging Cross-Spatial RGB-T Dataset

We introduce a novel multimodal RGB+thermal dataset of 1,890 paired images across nine diverse scenes, captured using a FLIR One Pro LT [37]. The dataset includes residential and outdoor environments (e.g., buildings, structures) as well as objects (e.g., metallic items, a telescope), which are visualized in Fig. 2. To assess cross-spatial reconstruction, each scene includes paired RGB-thermal images captured along two distinct trajectories. The dataset is publicly available [35].

The dataset consists of two subsets, differentiated by trajectory characteristics and environmental conditions. The first subset consists of six scenes (upper section of Fig. 2) captured under consistent, well-lit conditions (e.g., daytime or artificially illuminated indoor settings). Trajectories are non-intersecting, simulating independent data collection by separate sensors. Ground-truth poses are estimated using VGGT on all RGB images from both trajectories. The second subset (bottom section of Fig. 2) consists of three scenes captured under varying lighting (i.e., outdoor scenes spanning day/night). Since half of all scenes are low-illumination, RGB-based camera pose estimation is unreliable. Given the high uncertainty of pose estimation from thermal images, this subset is used for qualitative assessment due to the lack of reliable ground truth.

### 5.2 Publicly Available Datasets

Despite the scarcity of RGB-thermal datasets with ground-truth poses, we show that fine-tuning our model requires only limited real-world data. Our training set includes 66 scenes (around 15,000 RGB-thermal image pairs), aggregated from five publicly available datasets and randomly split into 48 training and 18 evaluation scenes: ThermoScenes [9] 17(train)/5(val), ThermalNeRF [20] 7/2, ThermalGaussian [23] 11/3, ThermalMix [29] 4/2, and the Radar Forest (RF) [15] Dataset 9/6—scene names per split are in the Supplementary Material (Sec. K). We exclude datasets not designed for scene reconstruction (e.g., autonomous driving datasets like [16, 34]).

For ThermoScenes, ThermalGaussian, and ThermalMix, RGB and thermal images are paired with identical extrinsics. We derive ground-truth extrinsics, intrinsics, and depth from VGGT reconstructions on RGB images, then transfer these to corresponding thermal frames. Since our goal is cross-modal reconstruction from unpaired data (not improving VGGT's pose estimation), using VGGT-derived RGB poses as ground truth does not bias validation. In ThermalNeRF [20], RGB and thermal images are paired via a known rigid transformation.

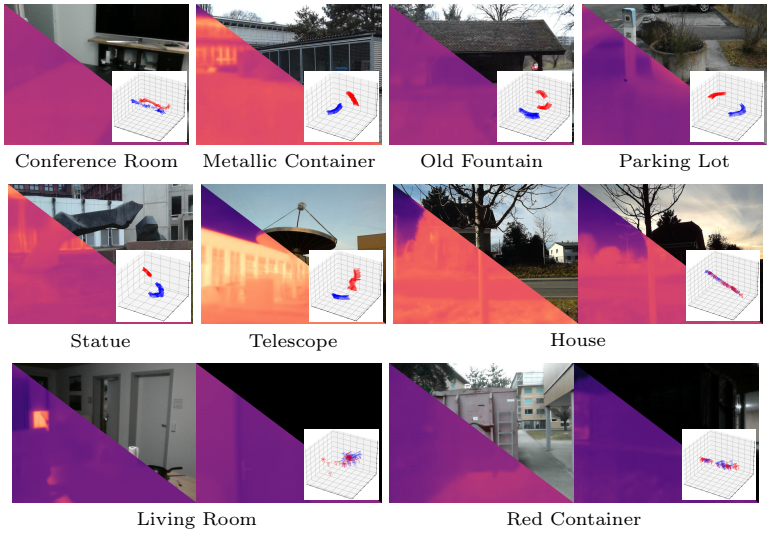



**Fig. 2:** The SEAR dataset includes 9 scenes, each with paired RGB-thermal images captured along two distinct trajectories. Ground-truth poses (red/blue for each trajectory) are estimated via VGGT on all RGB images. The top 6 scenes feature trajectories under similar lighting, while the bottom 3 have large lighting variations (some RGB images are near fully black).

We estimate RGB extrinsics using VGGT and derive thermal extrinsics by applying the rigid transform. Thermal depth maps are generated by projecting RGB depth estimates into 3D world points and reprojecting them onto the thermal frame. The RF dataset includes high-precision camera poses captured via a motion capture system and sparse LiDAR point clouds—both serving as ground truth in evaluation, eliminating the need for ground truth estimation using VGGT. We generate depth maps by projecting LiDAR points into the camera frame using the provided extrinsics and intrinsics.

### 5.3 Data Split of Evaluation Scenes

Validation uses a fixed thermal ratio of $\tau = 0.5$ (to avoid modality imbalance) and a batch size equal to the total number of frames per scene. As for training, we ensure that no RGB and thermal images in a batch share the same pose; given paired images, each scene is evaluated twice by swapping RGB and thermal image selections. This eliminates possible selection bias (since both runs are complementary to each other) and yields 36 evaluation cases (18 scenes $\times$ 2).

For evaluation on SEAR dataset, we use the dataset's split into two non-overlapping trajectories, pairing thermal images from one with RGB from the other. Each of the 6 scenes is evaluated twice (once per modality-trajectory pair), yielding 12 evaluation cases. We additionally include SmokeSeer [13], which provides two RGB-thermal drone scenes (both with and without dense smoke) but lacks ground-truth poses. We use SmokeSeer for qualitative evaluation only

since, unlike other datasets, VGGT-based pose estimation fails here due to high uncertainty on smoke-occluded RGB frames.

## 6 Experiments

### 6.1 Implementation Details

Our model's trained weights and implementation are available online [25]. We use the VGGT checkpoint from HuggingFace[4]. Training uses `bf-16` mixed precision, `AdamW` optimizer with learning rate of $5 \cdot 10^{-5}$ and weight decay of $10^{-2}$, linear warmup scheduling (first 10% of epochs, LR from 0.0 to $5 \cdot 10^{-5}$). We train the model for 100 epochs with a batch size of 24 (around 2 days on a single A100). To improve scalability and generalization to arbitrary input sizes, we follow VGGT's approach and partition the 24 frames of each batch into $N \in \{1, 2, 3, 4, 6, 12\}$ equal-length sequences. To reduce variations due to known numerical instabilities [47], we fine-tune 13 models with different seeds and report their average metrics.

To demonstrate that bridging the multimodal knowledge gap requires only minimal fine-tuning, the LoRA rank used is $r = 64$ and the scaling factor $\alpha = 128$, yielding ∼50M trainable parameters—less than 5% of the original model's size. Since, as established in the original LoRA paper, tuning $\alpha$ is equivalent to adjusting the learning rate, we fix $\alpha$ and optimize the learning rate empirically across all scenes in the ThermoScenes dataset.

### 6.2 Baselines and Metrics

We benchmark against traditional SfM (COLMAP [33] with SuperPoint + SuperGlue [32]) and deep-learning based models (DUSt3R [42], MASt3R [17], MapAnything [14], and pretrained VGGT). Since these methods lack multimodal support, we input RGB/thermal images as a single modality. We also benchmark against hybrid approaches mixing feature matching and deep learning. We use $\text{MA}_{\text{ELoFTR}}$ (ELoFTR [43] trained with MatchAnything [10]) and $\text{MINIMA}_{\text{ROMA}}$ (RoMA [4] trained with MINIMA [31]) to establish RGB-thermal correspondences, followed by DIM [26] for 3D reconstruction and pose estimation. We also use MP-SfM [30], which builds on COLMAP and includes depth and normal predictions, as well as correspondences from a learned matching model—we use $\text{MINIMA}_{\text{ROMA}}$ [31] as the matching model.

We use publicly available implementations and official pretrained weights for all baselines and use identical data splits and computational resources (one A100 GPU) for all methods. Classical approaches are evaluated using their default parameters. Due to computational constraints and the fact that this is not the best-performing method, we evaluate MP-SfM only on the Public Datasets and not on the more challenging SEAR dataset. Qualitative renderings for this method are provided in the supplementary materials.

[4] https://huggingface.co/facebook/VGGT-1B

Following VGGT [41], we evaluate the methods using the RRA@30 and RTA@30 (percentages of image pairs with relative rotation/translation errors $< 30°$) and AUC@30 (area under the accuracy-threshold curve of the maximum values between RRA and RTA with thresholds $[5.0, 15.0, 30.0]°$). We also report the processed frames per second (FPS) and registration rate (Reg.) [5] (percentage of successfully registered images). We compute the quality metrics on registered cameras only, excluding images without predicted poses. When LiDAR data is available, we report point cloud completeness (PCC) (the mean nearest-neighbor distance from reconstruction to LiDAR) and point cloud accuracy (PCA) (the mean nearest-neighbor distance from LiDAR to reconstruction). We also report the Chamfer distance, i.e., the average of PCC and PCA. To compute these metrics, we align estimated and ground-truth camera poses using Umeyama alignment [39], back-project estimated depth maps, and calculate the metrics between estimated and ground-truth point clouds. To mitigate bias toward methods that perform well on only a subset of scenes, we report the metrics over the errors across all datasets/scenes (for scene-agnostic evaluation) in Tab. 1.

### 6.3 Multimodal Camera Pose Estimation

Our method significantly outperforms all baselines (see Tab. 1) while maintaining a high registration rate. E.g., we achieve an AUC@30 of 70.0 on the Public Datasets and of 62.8 on our SEAR dataset, compared to 41.0/48.2 for $\text{MINIMA}_{\text{ROMA}}$, the method with the second-highest performance and a strong registration rate—though COLMAP achieves better metrics, its low registration rate artificially boosts those results. Non-multimodal approaches fail on thermal images due to the lack of discriminative features for cross-modal alignment, often producing disjoint reconstructions (see Fig. 3)—e.g., COLMAP only registers frames of one modality. $\text{MA}_{\text{ELoFTR}}$ produces sparse matches leading to low registration rates and poor 3D reconstructions. Interestingly, even methods trained with grayscale images (which could be construed as a substitute for thermal images), such as VGGT and MapAnything, fail to register or estimate depth in real thermal images, leading to incorrect 3D reconstructions: MapAnything exhibits the same failure mode as VGGT, reconstructing two separate point clouds. These failure cases underscore the necessity of both real data and specialized training strategies for successful RGB-Thermal reconstruction. Furthermore, our method is 200× faster (9.94 FPS vs. $\text{MINIMA}_{\text{ROMA}}$'s 0.05 FPS), nearly matching VGGT's speed (9.94 FPS vs. 10.46 FPS). Unlike other methods, SEAR achieves consistent RGB-T scene reconstruction.

### 6.4 Multimodal Point Cloud Reconstruction

As shown in Tab. 1, SEAR achieves the highest point cloud metrics: PCC of 0.06, PCA of 0.47, and Chamfer distance of 0.27. The next-best method, MASt3R, only has a PCC of 0.27, PCA of 0.66, and Chamfer distance of 0.46, while COLMAP+SPSG produces single-modality sparse point clouds. For hybrid methods, the quality of 3D reconstructions is dependent on the 2D-2D correspondences'

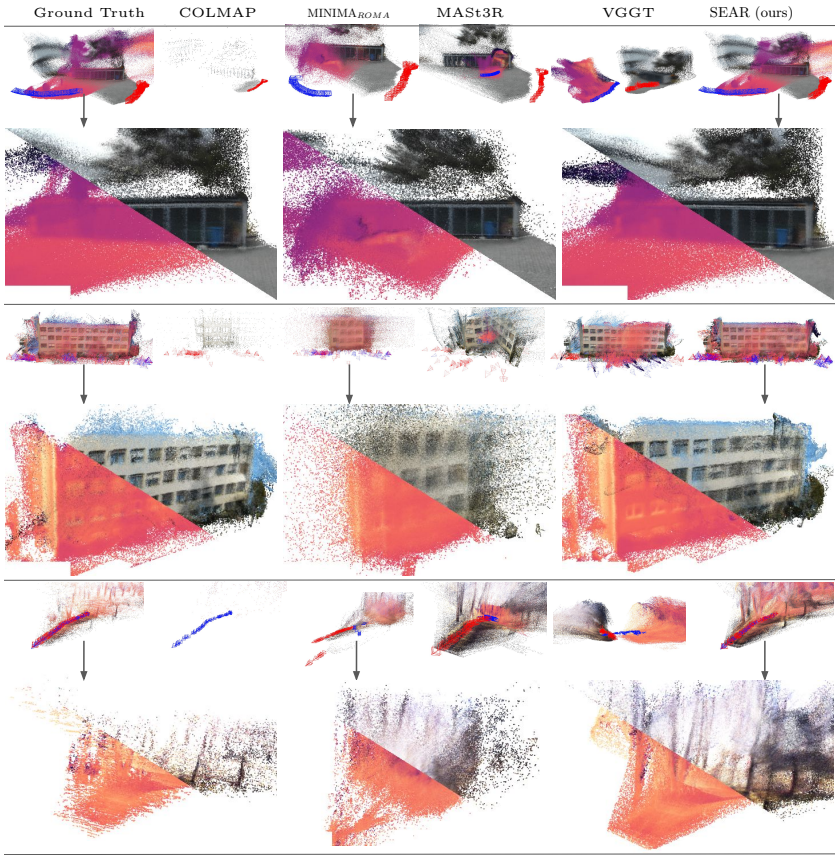



**Fig. 3:** Qualitative results comparing RGB/thermal reconstructions (camera poses in red/blue); we show results for 4 methods at rows 1, 3, and 5, and zoom in on more interesting reconstruction details in rows 2, 4, and 6. Our method (SEAR) achieves higher accuracy, consistency, and level of detail than other methods.

quality and density; $\text{MA}_{\text{ELoFTR}}$ only detected a small number of correspondences, leading to low metrics. $\text{MINIMA}_{\text{ROMA}}$ finds more correspondences, but their uncertainty results in low-quality reconstructions—its Chamfer distance is only 1.03 compared to 0.27 for SEAR. Similarly, MP-SfM achieves a Chamfer distance of only 1.08, due to the limited robustness of depth estimation on thermal images.

While point cloud metrics are only calculated against scenes with LiDAR ground truth, Fig. 3 provides additional qualitative comparisons across diverse scenes. DUSt3R, MASt3R, and VGGT frequently generate disjoint 3D representations for each modality, exhibiting inconsistencies in both scale and spatial alignment between modalities. $\text{MINIMA}_{\text{ROMA}}$ reconstructs less detailed point clouds than our method, which reconstructs consistent multimodal point clouds.

**Table 1:** Quantitative comparison of 3D reconstruction methods for all datasets. Metrics include `AUC`, `RRA`, and `RTA` @30 (higher is better, ↑), point cloud accuracy `PCA`, point cloud completeness `PCC`, `Chamfer` distance (lower is better, ↓), registration rate `Reg.` (higher is better, ↑), and frames per second `FPS` (higher is better, ↑). Best and second-best results are highlighted in blue and light blue, respectively.

| | Public Datasets | | | | | | | | SEAR Dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | AUC ↑ | RRA ↑ | RTA ↑ | PCA ↓ | PCC ↓ | Chamfer ↓ | Reg (%) ↑ | FPS ↑ | AUC ↑ | RRA ↑ | RTA ↑ | Reg (%) ↑ | FPS ↑ |
| COLMAP +SPSG | 57.6 | 82.5 | 74.6 | 1.64 | 1.20 | 1.42 | 44.7 | 0.44 | 74.4 | 99.9 | 95.9 | 27.9 | 0.66 |
| $MA_{ELoFTR}$ | 13.1 | 70.1 | 41.2 | 0.49 | 5.16 | 2.82 | 21.8 | 0.21 | 15.9 | 73.7 | 53.1 | 37.3 | 0.17 |
| $MINIMA_{ROMA}$ | 41.0 | 68.3 | 63.0 | 0.97 | 1.09 | 1.03 | 98.7 | 0.05 | 48.2 | 65.9 | 68.9 | 100.0 | 0.04 |
| MP-SfM | 31.4 | 78.0 | 52.5 | 1.59 | 0.58 | 1.08 | 100.0 | 0.04 | - | - | - | - | - |
| DUSt3R | 18.9 | 45.9 | 42.1 | 0.72 | 4.72 | 2.72 | 100.0 | 0.66 | 18.7 | 51.0 | 39.2 | 100.0 | 0.67 |
| MASt3R | 30.8 | 69.7 | 55.8 | 0.66 | 0.27 | 0.46 | 100.0 | 0.24 | 39.1 | 54.9 | 56.3 | 100.0 | 0.25 |
| MapAnything | 21.4 | 51.3 | 47.4 | 0.69 | 3.76 | 2.23 | 100.0 | 1.98 | 23.2 | 51.4 | 52.2 | 100.0 | 2.12 |
| VGGT | 22.9 | 50.7 | 48.5 | 1.22 | 2.60 | 1.91 | 100.0 | 10.46 | 23.3 | 50.5 | 56.4 | 100.0 | 10.51 |
| SEAR | 70.0 | 90.6 | 87.6 | 0.47 | 0.06 | 0.27 | 100.0 | 9.94 | 62.8 | 83.7 | 84.2 | 100.0 | 10.22 |

### 6.5 Qualitative Evaluation

We perform qualitative evaluation on SmokeSeer [13] and the subset of our new dataset with varying lighting conditions (see Fig. 4). Under challenging conditions, SEAR estimates well-aligned 3D camera poses across both RGB and thermal modalities, reconstructing the scenes with details, and demonstrating robust performance under challenging conditions (when both modalities are captured at different times or smoke is occluding the scene). In comparison, MASt3R and $MINIMA_{ROMA}$ misalign RGB/thermal reconstructions (rotated, improperly scaled thermal point cloud, with camera poses not consistent between modalities). On the other hand, VGGT sometimes achieves partial alignment (e.g., top image of Fig. 4), but the results are inconsistent and noisy (e.g., alignments are wrong on scenes from our new dataset), especially in the thermal domain. Additional qualitative results are provided in the Supplementary Material.

### 6.6 Varying thermal to RGB ratio

Prior experiments assumed a balanced RGB-thermal ratio ($\tau = 0.5$). To evaluate robustness to unequal ratios, we vary $\tau \in [0, 1]$ and report camera pose estimation in Fig. 5. For each $\tau$, we perform three random inferences per scene with randomly selected non-corresponding RGB-thermal pairs to reduce statistical variance. Fig. 5 shows a slow but constant performance decrease as $\tau$ increases, with a slight recovery after $\tau \approx 0.75$.

We hypothesize that the metrics decrease when $\tau \in [0.0, 0.75]$ is mostly due to thermal-specific characteristics—e.g., the ghosting effect [9], which reduces sharpness and contrast in thermal images, making the pose reconstruction more difficult. For $\tau \in [0.75, 1.0]$, quality improves because the model increasingly operates in a near single-modality regime, reducing multimodal alignment complexity, while the larger number of thermal views provides stronger scene coverage for

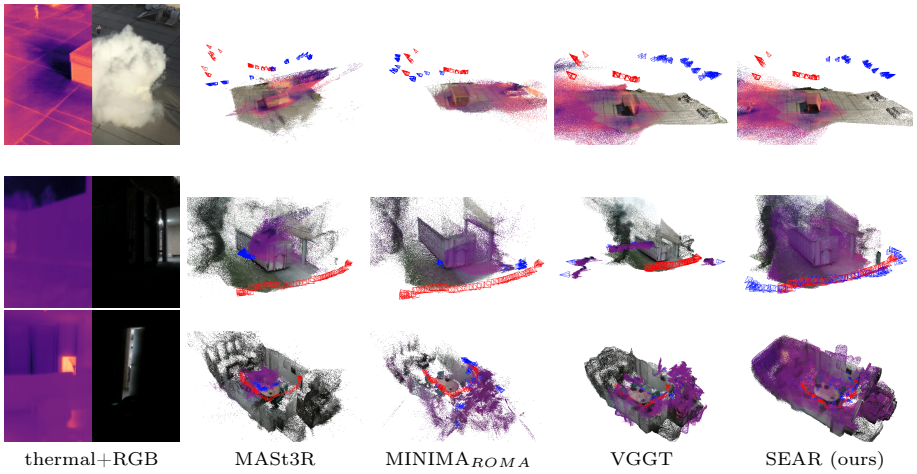
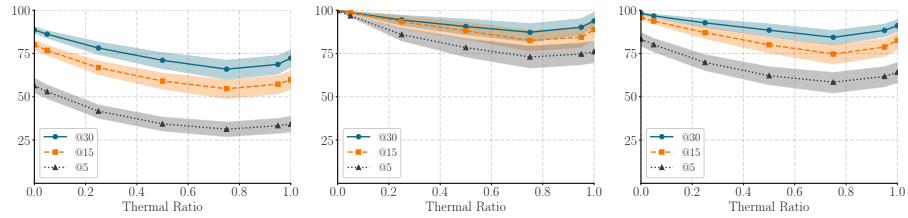



**Fig. 4:** Reconstructions from the SmokeSeer dataset (dense smoke, top) and our new dataset's scenes (lighting changes, middle and bottom rows). The first column shows cases where RGB images (right) are unreliable for localization, so thermal images (left) are used. Our method recovers the scene even with smoke (SmokeSeer) and aligns RGB and thermal camera poses in different lighting conditions (our dataset).




**(a)** AUC ↑ **(b)** RRA ↑ **(c)** RTA ↑

**Fig. 5:** The AUC, RRA, RTA (errors $<30°$, $15°$, $5°$) across varying thermal-to-RGB image ratios. The filled area represents the boundary from 0.25− to 0.75−quantiles estimated by bootstrapping scenes 2000 times.

thermal reconstruction. The minimum occurs at $\tau \approx 0.75$ (rather than $\tau = 0.5$) since, while more thermal frames weaken pose cues, a thermal-dominant input reduces cross-modal alignment and increases thermal view coverage.

### 6.7 Thermal to RGB Alignment

In contrast to VGGT, we believe that our method explicitly keeps the distributions of thermal and RGB tokens in the AA module aligned. To validate this hypothesis, we feed RGB-thermal image pairs into our model and extract intermediate outputs from the AA module's layers. We compute the cosine similarity [8] between same-level RGB and thermal tokens (excluding camera tokens). High similarity implies aligned distributions, while low similarity implies divergence. We compare RGB-to-RGB cosine similarity $x_{r2r}$ (the baseline) to the cosine similarity $x_{r2t}$ between

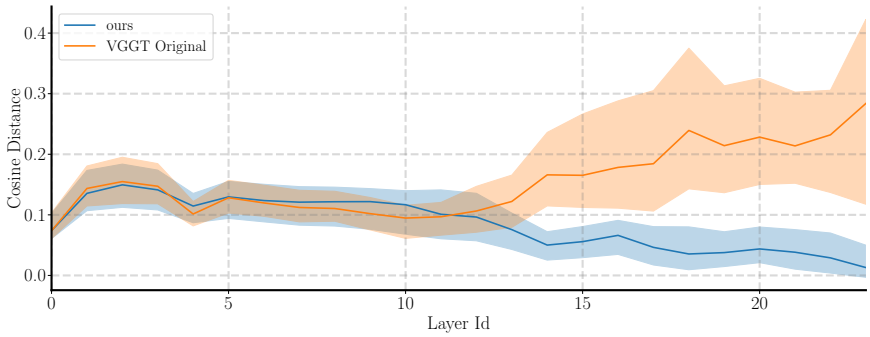



**Fig. 6:** The blue line represents the median difference between RGB-to-RGB and RGB-to-thermal cosine similarity dependence across layers for SEAR method. The orange line represents the same dependency for the VGGT model. The filled area represents the boundary from 0.25− to 0.75−quantiles.

RGB and thermal images. A large difference between $x_{r2r}$ and $x_{r2t}$ indicates a distribution mismatch, while a small difference suggests that the distributions are close. Experiments use a batch size of 12 and thermal ratios $\tau \in [0.25, 0.75]$.

Our results (Fig. 6) show that for the first 10 layers, both our model and VGGT show low RGB-thermal cosine similarity—interestingly, Bratulić et al. [1] demonstrated that VGGT reconstructs geometry post-`layer10` of the AA-module. Beyond `layer10`, when the geometric reconstruction starts, VGGT's difference between $x_{r2r}$ and $x_{r2t}$ increases, while ours remains consistently low, suggesting our model successfully aligned RGB and thermal token distributions for reconstruction in the whole AA module.

### 6.8 Ablation Studies

We perform extensive ablation studies to support our choice of architecture. We evaluate: 1) using the non-learnable rgb camera token instead of the learnable thermal camera token, 2) LLaVA [21]'s strategy (adding a learnable thermal projector of size 1024, initialized as the identity transformation, for thermal images), 3) adding a learnable thermal embedding, similar to positional embeddings in transformer [40] (initialized as zero to preserve the original model's performance at the beginning of training), 4) only adding LoRA to global-attention layers, and 5) only applying LoRA to frame-attention layers.

We compute metrics over the aggregated scenes of Public Datasets and SEAR dataset, and present their mean and standard deviation, as well as statistical analysis in Tab. 2 and Tab. 3. Our statistical analysis is a one-sided Welch's t-test evaluating the null hypothesis that a given model (row) outperforms another (column) in terms of AUC@30. Our study reveals that incorporating a learnable thermal projector or learnable thermal embeddings does not improve performance. Specifically, our method has p-values of 0.05 (trend) and 0.02 (statistically significant) when compared to those models, supporting our decision

**Table 2:** This table shows the camera pose estimation metrics average and 95% confidence interval for the different ablation studies of our paper.

| Method | AUC@30 | RRA@30 | RTA@30 |
|---|---|---|---|
| No Camera Token | $67.8 \pm 1.4$ | $89.1 \pm 1.0$ | $86.6 \pm 1.3$ |
| Thermal Projector | $67.2 \pm 1.1$ | $88.3 \pm 0.8$ | $86.0 \pm 0.8$ |
| Thermal Embedding | $66.1 \pm 1.8$ | $87.6 \pm 1.0$ | $85.1 \pm 1.2$ |
| Global Only | $69.1 \pm 1.5$ | $89.4 \pm 1.1$ | $87.3 \pm 1.6$ |
| Frame Only | $62.4 \pm 0.6$ | $85.6 \pm 0.9$ | $83.0 \pm 1.0$ |
| SEAR | $68.4 \pm 0.7$ | $89.1 \pm 0.5$ | $86.9 \pm 0.4$ |

**Table 3: Pairwise statistical significance (AUC@30).** Each entry at position $(i, j)$ reports the p-value from a one-sided Welch's t-test for the null hypothesis that the row method outperforms the column method in AUC@30. Darker blue indicates smaller p-values (stronger evidence against the null).

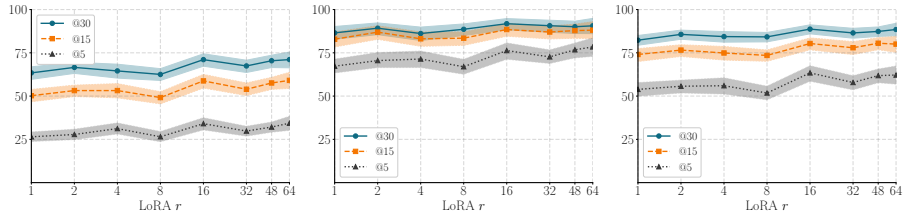
| | No Camera Token | Thermal Projector | Thermal Embedding | Global Only | Frame Only | Ours |
|---|---|---|---|---|---|---|
| No Camera Token | – | 0.72 | 0.91 | 0.12 | 1.00 | 0.24 |
| Thermal Projector | 0.28 | – | 0.83 | 0.04 | 1.00 | 0.05 |
| Thermal Embedding | 0.09 | 0.17 | – | 0.02 | 1.00 | 0.02 |
| Global Only | 0.88 | 0.96 | 0.98 | – | 1.00 | 0.78 |
| Frame Only | 0.00 | 0.00 | 0.00 | 0.00 | – | 0.00 |
| Ours | 0.76 | 0.95 | 0.98 | 0.22 | 1.00 | – |

not to include a learnable thermal projector or learnable thermal embeddings. For the no-camera-token model, the p-value of 0.24 does not invalidate the null hypothesis. However, when the no-camera-token model is evaluated against the thermal-embedding model, the p-value (0.09) only shows a trend, while our method achieves statistical significance ($p = 0.02$). Similarly, the no-camera-token model shows no significant difference from the thermal-projector baseline ($p = 0.28$), whereas our method reaches statistical significance ($p = 0.05$). These results indicate that, although our architecture does not significantly outperform the no-camera-token model, it provides some improvements and greater stability when compared to other models.

Our method and the global-only model achieved comparable results in terms of statistical significance. While our analysis does not indicate a clear superiority of one approach over the other, future work could explore applying LoRA layers exclusively to the global attention layer—rather than both the global and frame layers—to further reduce parameter count without compromising performance.

### 6.9 LoRA Parameters

We evaluate the impact of different values of $r$ for our LoRA layers; due to computational constraints, we run each experiment only once. We selected $r$ to keep the number of trainable parameters low (approximately 50M, less than




**(a)** AUC ↑ **(b)** RRA ↑ **(c)** RTA ↑

**Fig. 7:** The AUC, RRA, RTA (errors <30°, 15°, 5°) across varying LoRA $r$. The filled area represents the boundary from 0.25− to 0.75−quantiles estimated by bootstrapping scenes 2000 times.

5% of the original model size). As shown in Fig. 7, while increasing the rank $r$ slightly improves the final accuracy—suggesting that higher-rank adapters provide additional representativeness—the gains are modest and only a small $r$ value is enough. We hypothesize that this is due to VGGT's ability to process thermal images alone: the adapter only needs to estimate the RGB+thermal alignment, a comparatively smaller task.

## 7 Conclusion

We present SEAR, a simple and efficient framework to adapt visual geometric grounded transformers for RGB-T camera pose estimation and 3D scene reconstruction. By leveraging lightweight LoRA adapters and a carefully designed batching strategy, SEAR achieves state-of-the-art performance with minimal computational overhead and a modest amount of data, making it a practical solution for real-world applications. Our approach not only outperforms existing methods across all metrics but also demonstrates robustness in challenging real-world conditions such as low lighting, dense smoke, and asynchronous data capture. While this study focuses on RGB-Thermal reconstruction, the methodology is possibly extensible to other modalities—such as multi-spectral imaging—which remains a subject for future work. By showing that high-performance multimodal reconstruction is achievable with minimal adaptation and a relatively small amount of data, we hope to inspire broader adoption of parameter-efficient fine-tuning in multimodal geometric vision.

## Acknowledgements

We thank Ivan Skorokhodov for his assistance with writing, earlier drafts, and help with several experiments. We also thank Martin Magnusson for valuable discussions and for suggesting several evaluation metrics and datasets. Part of the experiments was run thanks to the facilities of the Scientific IT and Application Support Center of EPFL.

## References


1. Bratulić, J., Mittal, S., Brox, T., Rupprecht, C.: On Geometric Understanding and Learned Priors in Feed-forward 3D Reconstruction Models, (2026). arXiv: 2512.11508 [cs.CV]. http://arxiv.org/abs/2512.11508. Pre-published. https://doi.org/10.48550/arXiv.2512.11508
2. Chassaing, E., Forest, F., Fink, O., Mielle, M.: Thermoxels: A Voxel-Based Method to Generate Simulation-Ready 3D Thermal Models. J. Phys.: Conf. Ser. **3140**(4), 042003 (2025). https://doi.org/10.1088/1742-6596/3140/4/042003
3. Cordonnier, J., Xu, C., Fink, O., Mielle, M.: Unpaired RGB-Thermal Gaussian-Splatting Using Visual Geometric Transformers. In: ICRA 2026 Workshop on Multi-Modal Spatial AI for Robust Navigation and Open-World Understanding. IEEE (2026). https://doi.org/10.48550/arXiv.2606.05491. arXiv: 2606.05491 [cs.CV]
4. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust Dense Feature Matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19790–19800 (2024)
5. Elflein, S., Zhou, Q., Leal-Taixé, L.: Light3r-Sfm: Towards Feed-Forward Structure-from-Motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16774–16784 (2025)
6. Fan, R., Zhao, W., Lin, M., Wang, Q., Liu, Y.-J., Wang, W.: Generalizable Thermal-based Depth Estimation via Pre-trained Visual Foundation Model. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 14614–14621. IEEE (2024)
7. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM **24**(6), 381–395 (1981). https://doi.org/10.1145/358669.358692
8. Han, J., Hong, S., Jung, J., Jang, W., An, H., Wang, Q., Kim, S., Feng, C.: Emergent Outlier View Rejection in Visual Geometry Grounded Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 427–437 (2026)
9. Hassan, M., Forest, F., Fink, O., Mielle, M.: ThermoNeRF: A Multimodal Neural Radiance Field for Joint RGB-thermal Novel View Synthesis of Building Facades. Advanced Engineering Informatics **65**, 103345 (2025). https://doi.org/10.1016/j.aei.2025.103345
10. He, X., Yu, H., Peng, S., Tan, D., Shen, Z., Bao, H., Zhou, X.: MatchAnything: Universal Cross-Modality Image Matching with Large-Scale Pre-Training, (2025). arXiv: 2501.07556 [cs.CV]. http://arxiv.org/abs/2501.07556. Pre-published. https://doi.org/10.48550/arXiv.2501.07556
11. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-Efficient Transfer Learning for NLP. In: International Conference on Machine Learning, pp. 2790–2799. PMLR (2019)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank Adaptation of Large Language Models. In: International Conference on Learning Representations (2022)
13. Jain, N., Jong, A., Scherer, S., Gkioulekas, I.: SmokeSeer: 3D Gaussian Splatting for Smoke Removal and Scene Reconstruction. In: 2026 International Conference on 3D Vision (3DV), pp. 510–519. IEEE (2026)
14. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N.: MapAnything: Universal Feed-Forward Metric

3D Reconstruction; Map-Anything. Github. Io. In: 2026 International Conference on 3D Vision (3DV), pp. 499–509. IEEE (2026)
15. Kubelka, V., Kotlyar, O., Artan, U., Magnusson, M.: Viking Hill Dataset: A Lidar-Radar-Camera Dataset for Detection and Segmentation in Forest Scenes, (2026). arXiv: 2606.19154 [cs.RO]. http://arxiv.org/abs/2606.19154. Pre-published. https://doi.org/10.48550/arXiv.2606.19154
16. Lee, A.J., Cho, Y., Shin, Y.-s., Kim, A., Myung, H.: Vivid++: Vision for Visibility Dataset. IEEE Robotics and Automation Letters **7**(3), 6282–6289 (2022)
17. Leroy, V., Cabon, Y., Revaud, J.: Grounding Image Matching in 3D with MASt3R. In: Computer Vision – ECCV 2024. Ed. by A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, pp. 71–91. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-73220-1_5
18. Lester, B., Al-Rfou, R., Constant, N.: The Power of Scale for Parameter-Efficient Prompt Tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045–3059 (2021)
19. Li, J., Li, Y., Sun, C., Wang, C., Xiang, J.: Spec-Nerf: Multi-spectral Neural Radiance Fields. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2485–2489. IEEE (2024)
20. Lin, Y.Y., Pan, X.-Y., Fridovich-Keil, S., Wetzstein, G.: Thermalnerf: Thermal Radiance Fields. In: 2024 IEEE International Conference on Computational Photography (ICCP), pp. 1–12. IEEE (2024)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
22. Liu, Y., Chen, X., Yan, S., Cui, Z., Xiao, H., Liu, Y., Zhang, M.: Thermalgs: Dynamic 3d Thermal Reconstruction with Gaussian Splatting. Remote Sensing **17**(2), 335 (2025)
23. Lu, R., Chen, H., Zhu, Z., Qin, Y., Lu, M., Zhang, L., Yan, C., Xue, A.: ThermalGaussian: Thermal 3D Gaussian Splatting, (2025). arXiv: 2409.07200 [cs.CV]. http://arxiv.org/abs/2409.07200. Pre-published. https://doi.org/10.48550/arXiv.2409.07200
24. Mi, L., Wang, W., Tu, W., He, Q., Kong, R., Fang, X., Dong, Y., Zhang, Y., Li, Y., Li, M., Dai, H., Chen, G., Liu, Y.: Empower Vision Applications with LoRA LMM. In: Proceedings of the Twentieth European Conference on Computer Systems, pp. 261–277. ACM, Rotterdam Netherlands (2025). https://doi.org/10.1145/3689031.3717472
25. Mielle, M., Skorokhodov, V.: *Schindler-EPFL-Lab/SEAR: ECCV 2026*, version eccv2026 (2026). https://doi.org/10.5281/ZENODO.21077295. Zenodo.
26. Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F., Remondino, F.: Deep-Image-Matching: A Toolbox for Multiview Image Matching of Complex Scenarios. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **48**, 309–316 (2024). https://doi.org/10.5194/isprs-archives-XLVIII-2-W4-2024-309-2024
27. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE transactions on robotics **31**(5), 1147–1163 (2015)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A.: Dinov2: Learning Robust Visual Features without Supervision, (2023). arXiv: 2304.07193.
29. Özer, M., Weiherer, M., Hundhausen, M., Egger, B.: Exploring Multi-modal Neural Scene Representations With Applications on Thermal Imaging. In: Computer

Vision – ECCV 2024 Workshops. Ed. by A. Del Bue, C. Canton, J. Pont-Tuset, and T. Tommasi, pp. 82–98. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-92805-5_6
30. Pataki, Z., Sarlin, P.-E., Schönberger, J.L., Pollefeys, M.: Mp-Sfm: Monocular Surface Priors for Robust Structure-from-Motion. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 21891–21901 (2025)
31. Ren, J., Jiang, X., Li, Z., Liang, D., Zhou, X., Bai, X.: Minima: Modality Invariant Image Matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 23059–23068 (2025)
32. Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning Feature Matching with Graph Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4938–4947 (2020)
33. Schonberger, J.L., Frahm, J.-M.: Structure-from-Motion Revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)
34. Shin, U., Park, J., Kweon, I.S.: Deep Depth Estimation from Thermal Image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1043–1053 (2023)
35. Skorokhodov, V., Mielle, M.: *SEAR*. Dataset. Zenodo, Mar. 17, 2026. https://doi.org/10.5281/zenodo.19057885.
36. Sun, S., Mielle, M., Lilienthal, A.J., Magnusson, M.: High-Fidelity SLAM Using Gaussian Splatting with Rendering-Guided Densification and Regularized Optimization. In: 2024 IEEE International Conference on Intelligent Robots and Systems (IROS). IEEE, Abu Dhabi, UAE (2024). arXiv: 2403.12535
37. Teledyne FLIR, FLIR One® pro Thermal Imaging Camera for Smartphones, (2026). https://www.flir.com/products/flir-one-pro/
38. Tuzcuoğlu, Ö., Köksal, A., Sofu, B., Kalkan, S., Alatan, A.A.: Xoftr: Cross-modal Feature Matching Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4275–4286 (2024)
39. Umeyama, S.: Least-Squares Estimation of Transformation Parameters between Two Point Patterns. IEEE Transactions on pattern analysis and machine intelligence **13**(4), 376–380 (1991)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. Advances in neural information processing systems **30** (2017)
41. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual Geometry Grounded Transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 5294–5306 (2025)
42. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d Vision Made Easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20697–20709 (2024)
43. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient LoFTR: Semi-dense Local Feature Matching with Sparse-like Speed. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21666–21675 (2024)
44. Waseem, A., Mielle, M.: Physics-Informed Neural Networks for Thermophysical Property Retrieval, version 1. (2025). https://arxiv.org/abs/2511.23449. Pre-published. https://doi.org/10.48550/ARXIV.2511.23449
45. Wu, Y., Hu, Z.: PnP Problem Revisited. J Math Imaging Vis **24**(1), 131–141 (2006). https://doi.org/10.1007/s10851-005-3617-z

46. Xu, J., Liao, M., Kathirvel, R.P., Patel, V.M.: Leveraging Thermal Modality to Enhance Reconstruction in Low-Light Conditions. In: Computer Vision – ECCV 2024. Ed. by A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, pp. 321–339. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-72913-3_18
47. Yuan, J., Li, H., Ding, X., Xie, W., Li, Y.-J., Zhao, W., Wan, K., Shi, J., Hu, X., Liu, Z.: Understanding and Mitigating Numerical Sources of Nondeterminism in Llm Inference. Advances in Neural Information Processing Systems **38**, 169819–169851 (2026)