3D Vision in Thermal Images

Pallavi Aithal Narayan[1] and Salvin George[2]

University of Bonn, Bonn, North Rhine-Westphalia, Germany

Abstract. This paper explores the adaptation and fine-tuning of state-of-the-art 3D vision models for thermal imaging applications. While significant progress has been made in 3D perception from RGB images, the thermal domain remains relatively unexplored despite its crucial role in scenarios where visible light is limited or unreliable. We adapt the recent DUSt3R framework, which excels at dense and unconstrained stereo 3D reconstruction, for thermal imagery through a novel fine-tuning approach. A key contribution of our work is leveraging the MASt3R model to generate high-quality pseudo-ground truth annotations from paired RGB images, which are then used to train our thermal model. Our thermal-adapted model demonstrates promising performance on monocular depth estimation and 3D reconstruction tasks across the Freiburg Thermal Dataset. We introduce thermal-specific preprocessing techniques and loss functions that account for the unique characteristics of thermal images, such as their lower contrast and different edge properties compared to RGB imagery. Our experiments show that with proper adaptation, modern 3D vision architectures can effectively transfer to the thermal domain, opening possibilities for robust all-weather and all-lighting condition 3D perception systems.

1 Introduction

3D perception is a fundamental component of autonomous systems, enabling navigation, object detection, and scene understanding in real-world environments. While RGB cameras provide rich visual information, they suffer significant performance degradation in challenging lighting conditions such as darkness, glare, or adverse weather. Thermal cameras, which capture heat signatures emitted by objects independently of visible light, offer a complementary sensing modality that remains functional in these challenging scenarios.

Thermal imaging has proven especially valuable in robotics applications such as search and rescue operations, nighttime surveillance, and all-weather navigation. The invariance to lighting conditions and robustness to environmental factors like smoke or fog make thermal sensors particularly attractive for safety-critical systems that must operate reliably in diverse conditions [1].

Recent advances in deep learning have revolutionized 3D vision tasks for RGB imagery, with models like DUSt3R [3] and MASt3R [2] demonstrating impressive capabilities in dense 3D reconstruction from uncalibrated images. These models can jointly solve calibration and reconstruction problems, making them promis-

ing candidates for adaptation to the thermal domain. However, thermal images present unique challenges compared to their RGB counterparts:

- **Lower resolution and contrast**: Thermal cameras typically provide lower spatial resolution and contrast than RGB sensors.
- **Different edge characteristics**: Edges in thermal images correspond to temperature gradients rather than color or texture changes.
- **Lack of texture**: Many surfaces that exhibit rich texture in RGB appear homogeneous in thermal imagery, complicating feature extraction.
- **Limited training data**: Compared to the abundance of RGB datasets, thermal data with ground truth annotations is scarce.

In this work, we address these challenges by adapting the DUSt3R framework for thermal imagery. Our approach leverages pseudo-ground truth generation from RGB images to create training data for thermal model fine-tuning. We enhance the model with thermal-specific preprocessing techniques and loss functions to account for the unique characteristics of thermal images. Our main contributions are:

- A fine-tuned version of DUSt3R optimized for thermal imagery, capable of accurate monocular depth estimation and 3D reconstruction.
- A thermal-aware loss function that accounts for the unique edge characteristics and lack of texture in thermal images.
- Comprehensive evaluation on multiple thermal datasets, demonstrating the effectiveness of our approach in real-world scenarios.
- Insights into the domain gap between RGB and thermal imagery for 3D vision tasks, providing guidance for future research in this area.

2 Method

2.1 Dataset

The Freiburg Thermal Dataset is a multimodal dataset collected from a vehicle equipped with both RGB and thermal (IR) cameras. The dataset is designed for autonomous driving perception tasks and thermal-visual fusion research, with a particular focus on semantic segmentation, object detection, and now with our work, 3D vision tasks. The dataset contains both daytime and nighttime sequences, making it valuable for evaluating vision algorithms across varying illumination conditions.

The dataset is organized into train and test splits, with the train split containing 8 different sequences (5 daytime and 3 nighttime), while the test split is divided simply into day and night categories. Each training sequence is further divided into multiple drives (numbered 00, 01, 02, etc.), with every drive containing aligned RGB images, thermal (IR) images, and segmentation labels.

In total, the dataset comprises 33,217 RGB images and 21,565 thermal images. The training split contains the majority of the data with 33,025 RGB images and 21,309 thermal images, while the test split is relatively small with 192 RGB images and 256 thermal images.

The thermal images are captured using a FLIR thermal camera and are aligned with the RGB images. This alignment is crucial for our work as it enables us to transfer knowledge from the RGB domain to the thermal domain through pseudo-ground truth generation.

Preprocessing for 3D Vision: For our 3D vision task, we focused on the training split to extract RGB-thermal image pairs from consecutive frames within the same drive. These pairs serve as the basis for our pseudo-ground truth generation and model training. We apply specific preprocessing to the thermal images to enhance their contrast and normalize the intensity values, making them more suitable for feature extraction by our neural network.

2.2 Pseudo Ground Truth Generation

In the field of thermal 3D vision, a significant challenge is the lack of annotated data with accurate depth information. To overcome this limitation, we developed a pseudo ground truth generation pipeline that leverages a state-of-the-art RGB-based 3D vision model to create reliable annotations for thermal imagery. This approach enables supervised learning without requiring specialized depth sensors or manual annotations for thermal data.

For generating the pseudo ground truth, we utilized MASt3R, a powerful 3D vision model that excels at predicting accurate depth and 3D structure from RGB image pairs. This model was selected due to its robust performance across diverse environments and lighting conditions, making it suitable for generating reliable annotations that can later be transferred to the thermal domain.

We processed the Freiburg Thermal Dataset, which contains paired RGB and thermal images captured from driving scenarios. Our pipeline automatically identifies suitable image pairs from video sequences, applying the following procedures:

Sequence Discovery: Automatic identification of valid image sequences within the dataset directory structure Pair Formation: Selection of image pairs with configurable temporal separation to ensure sufficient baseline for 3D reconstruction Preprocessing: Resizing images to the model's input resolution (512×512) and normalizing pixel values

Annotation Generation Process The pseudo ground truth generation involves several key computational steps:

1. Model Inference: Each RGB image pair is processed through the MASt3R model, which performs implicit 3D reconstruction

2. 3D Data Extraction: From the model output, we extract:
   - Pointmaps: Dense fields of 3D points corresponding to each pixel
   - Confidence Maps: Per-pixel reliability estimates for the predictions
3. Derived Annotations: Additional information computed from the pointmaps:
   - Depth Maps: Distance from camera, extracted from the Z-coordinate
   - Camera Intrinsics: Focal length and principal point estimated using least squares
   - Relative Poses: Camera transformation matrices between image pairs

Result Validation To validate the quality of our generated pseudo ground truth, we implemented a comprehensive visualization pipeline that allows for both qualitative and quantitative assessment of the annotations. This validation was essential to ensure the reliability of the data used for training our thermal 3D vision model.

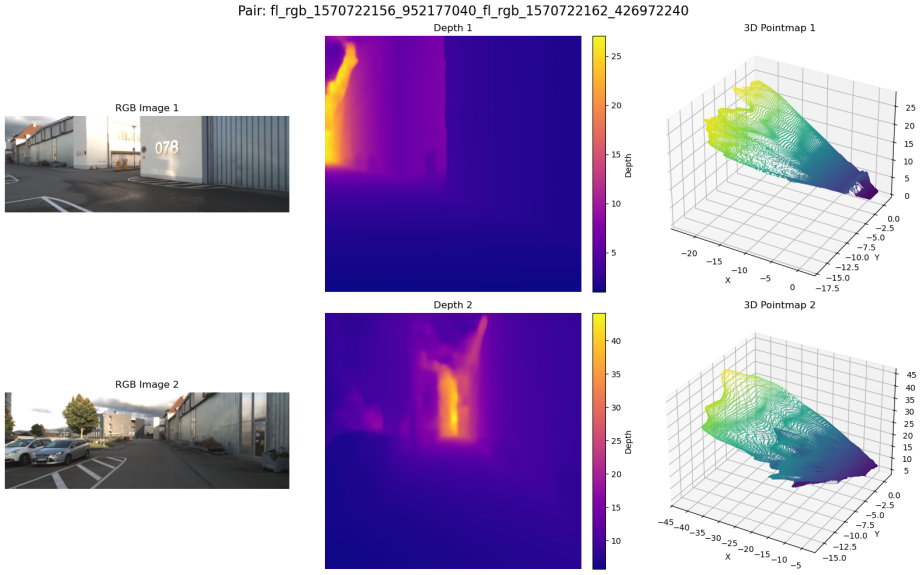

Fig. 1. Individual image pairs with their corresponding depth maps and 3D pointmaps in separate viewports

The generated depth maps demonstrated strong correspondence with scene structure visible in the RGB images. As shown in Figure 1 and Figure 2, the depth discontinuities accurately align with object boundaries, such as the building

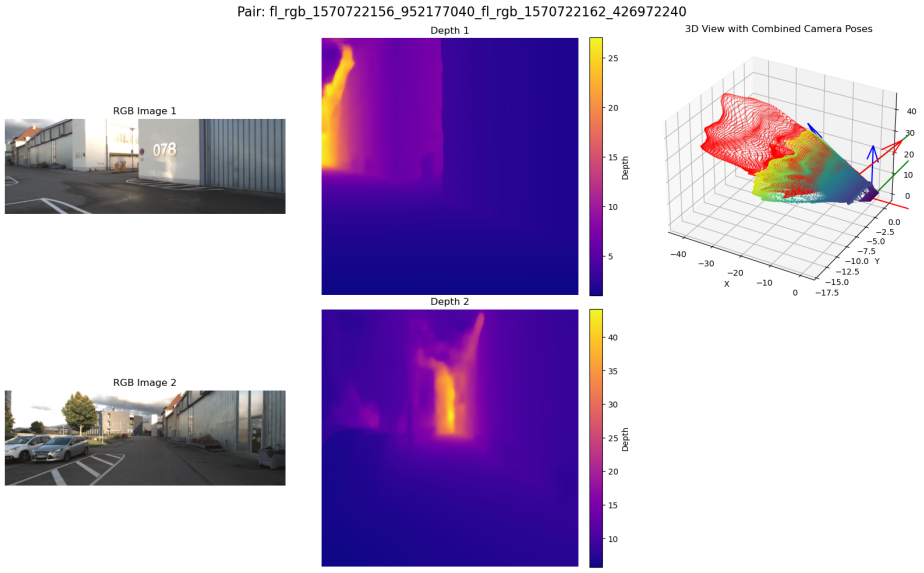

Fig. 2. Integrated view showing both camera positions in a unified 3D coordinate system

edges and parked vehicles visible in the Freiburg dataset. The color-coded depth visualization confirms that the model correctly identified:

- Foreground objects (vehicles, poles) with appropriate depth values
- Building facades with consistent depth gradients
- Ground planes with smooth depth transitions
- Far-field elements (sky, distant buildings) at appropriate distances

These validation results confirmed that our generated pseudo ground truth annotations provide a reliable foundation for training thermal 3D vision models, with sufficient accuracy in both depth estimation and camera parameter recovery.

2.3 Training Strategy

Dataset Processing: The Freiburg Thermal Dataset was processed using our custom FreiburgDataset class, which loads aligned pairs of thermal images from consecutive frames within the same drive. Each image was resized to 224×224 pixels to match our model's input requirements. A critical preprocessing step involved enhancing thermal image contrast, which applies percentile-based normalization (2nd and 98th percentiles) to improve feature visibility while preserving thermal gradients. For training/validation splitting, we implemented a random 80/20 split using PyTorch's random split function, ensuring both sub-

sets contained a balanced mix of day and night sequences. This splitting strategy helped evaluate generalization across different environmental conditions.

Training Procedure: Our training loop processes each batch as follows:

1. Extract thermal image pairs and corresponding ground truth data
2. Forward pass through the model to obtain pointmap predictions
3. Calculate the thermal-aware loss
4. Backpropagate gradients and update model parameters
5. Log metrics to WandB
6. Evaluate on the validation set after each epoch
7. Save checkpoints for the best-performing model (lowest validation loss)

3 Experimental Details

3.1 Experiment 1: Inference with pretrained model

The DUSt3R model is trained with RGB image pairs. So, to represent how it would perform with a thermal image, we have tried to do inference with the pretrained weights from DUSt3R. The image from inference is given below.

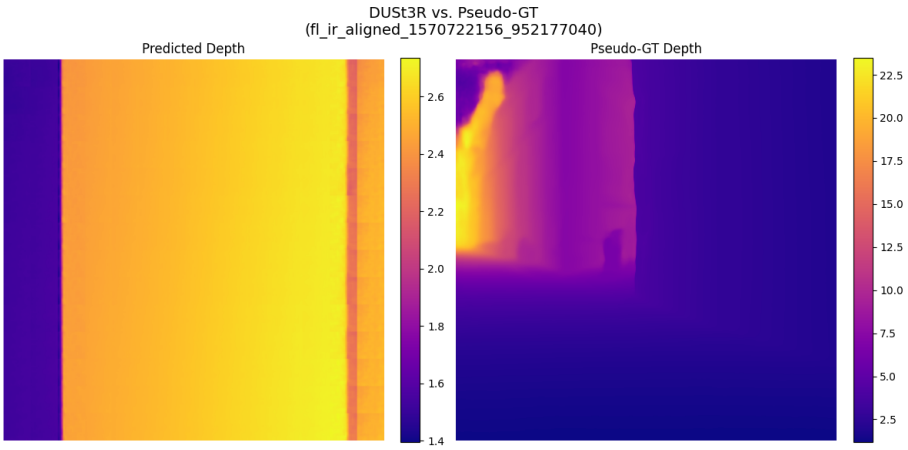

Fig. 3. Inference of thermal image with DUSt3R base model

As you can see in Figure 3, the base model is not able to capture much information about depth and the underlying structures when compared to to the ground truth.

3.2 Experiment 2: Finetuning with a base loss function

For our initial experiments, we implemented a straightforward approach to finetune the DUSt3R model on thermal imagery. We used a simple L1 loss function that directly measures the absolute difference between the predicted and ground truth pointmaps.

This approach computes the mean L1 error across the XYZ coordinates for each pixel location, and then averages these errors across all pixels in both pointmaps. The final loss is the average of the errors from both viewpoints, ensuring balanced learning for both images in the pair.

This simple loss function provided a baseline for our experiments, allowing us to evaluate the model's ability to adapt to thermal imagery before moving on to more sophisticated loss functions that incorporate thermal-specific characteristics.

The L1 loss is particularly suitable for this task because it is less sensitive to outliers compared to L2 loss, which is desirable when dealing with the inherent noise and lower contrast of thermal imagery. Additionally, by computing the mean across the three coordinates of the pointmaps, we ensure that the error is balanced across the X, Y, and Z dimensions of the predicted geometry.

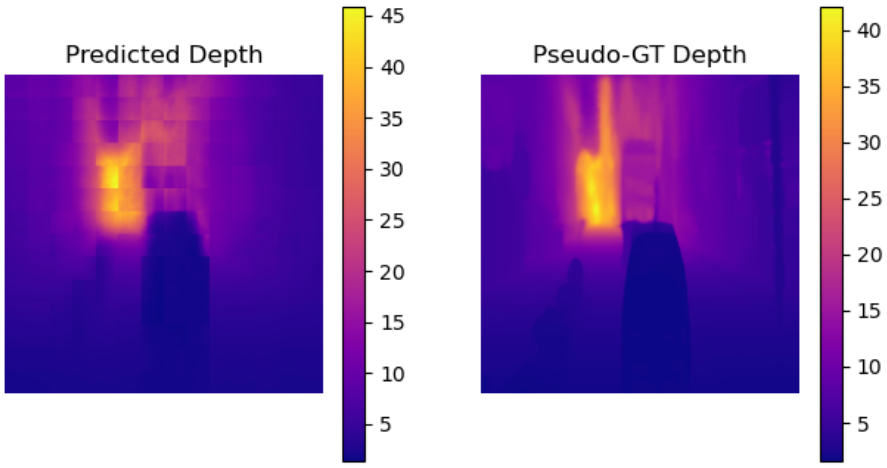

Fig. 4. Training of thermal image with fine-tuned DUSt3R base model with L1 loss function at the 10th epoch

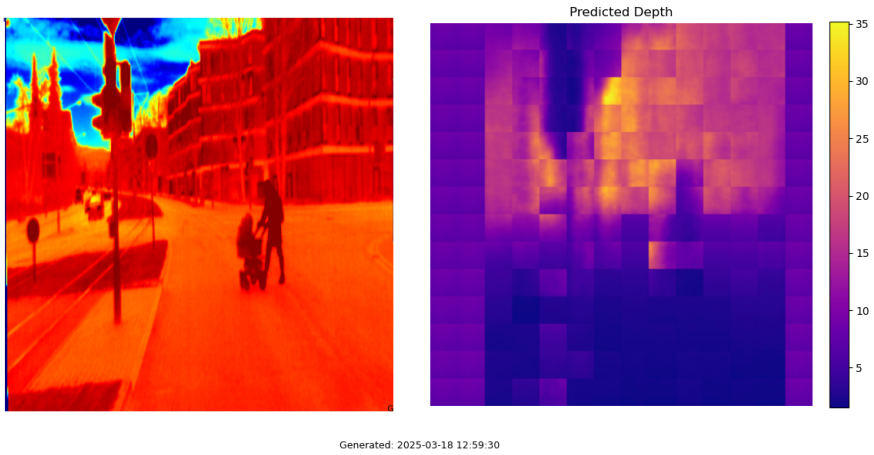

Fig. 5. Prediction of depth with fine-tuned DUSt3R model with L1 loss function

3.3 Experiment 3: Fine-tuning with a thermal-specific model wrapper

To effectively adapt DUSt3R for thermal imagery, we implemented a specialized model wrapper called ThermalDUSt3R.This wrapper enhances the base model with thermal-specific processing capabilities while preserving the core 3D vision functionality of DUSt3R.

Our wrapper enhances the base model through three main components. First, it applies edge detection using Sobel filters to highlight thermal boundaries, which helps identify object edges that might be less visible in thermal data. Second, it normalizes the thermal input to standardize temperature ranges across different images. Third, it applies a learnable scaling factor to optimize the contrast of thermal features.

The edge enhancement is particularly important since thermal images often lack the rich texture information that depth estimation models typically rely on. By emphasizing thermal gradients, our model can better identify structural elements in the scene. These enhancements are controlled by learnable parameters that automatically adjust during training to find the optimal settings for thermal imagery.

The wrapper preserves the original DUSt3R architecture while adding these thermal-specific preprocessing steps before the main model. During inference, each thermal image undergoes this preprocessing before being passed to the base model, improving its ability to extract meaningful features from thermal data without requiring changes to the core network architecture.

As you can see in Figure 6, while training the model successfully captured much details such as the general structure and details. But this model did not efficiently

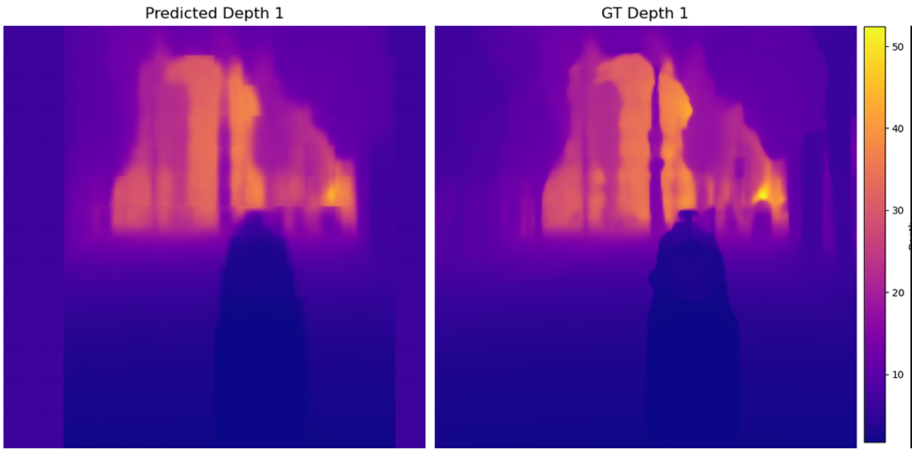

Fig. 6. Training of thermal image with fine-tuned DUSt3R base model with thermal specific wrapper in 25 epochs

perform the inference after training due to some discrepancies in the inference code.

3.4 Experiment 4: Finetuning with a modified thermal aware loss function

As the model with the ThermalDUSt3R wrapper didn't perform as well as expected for inference, our next objective was to use the base model structure but with enhanced thermal-specific loss functions. We developed a specialized loss function that addresses the unique characteristics of thermal imagery while maintaining the core DUSt3R architecture.

Our enhanced thermal-aware loss function consists of four complementary components:

1. **Basic regression loss:** The foundation of our loss function is the confidence-weighted regression loss similar to the original DUSt3R, which measures the L1 distance between predicted and ground truth 3D points, weighted by confidence values.
2. **Edge-aware loss:** This component specifically addresses the challenge of preserving depth discontinuities at thermal edges. It detects thermal gradients in the input images and encourages the depth gradients to be stronger at locations with significant thermal changes. This helps the model correctly identify object boundaries in thermal images, where texture cues are often lacking.
3. **Smoothness loss:** For regions with minimal thermal gradients (likely to be flat or homogeneous surfaces), this term encourages smooth depth predic-

tions. We use squared depth gradients to more strongly penalize unwanted depth variations in thermally uniform areas, addressing the common problem of noise in thermal depth predictions.

4. **Detail preservation loss:** To maintain fine structures and accurate depth details, this component ensures that predicted depth gradients match ground truth gradients. We implement this using a Huber loss for robustness against outliers, which is particularly important in thermal imagery where sensor noise can be significant.

The complete loss is a weighted combination of these components:

$$L = L_{basic} + \alpha \cdot L_{edge} + \beta \cdot L_{smoothness} + \gamma \cdot L_{detail} \tag{1}$$

where α, β, and γ are weighting parameters that control the contribution of each term. Through experimentation, we found optimal values of α=0.5 for edge weight, β=0.3 for smoothness weight by doing grid search to find the best performance as you can see in the below table.

Table 1. Thermal-Aware Loss Grid Search Results

Edge Weight	Smoothness Weight	Validation Loss
0.3	0.1	2.41743
0.3	0.3	2.47958
0.3	0.5	2.63495
0.5	0.1	2.48380
0.5	0.3	**2.36868**
0.5	0.5	2.50653
0.7	0.1	2.44136
0.7	0.3	2.51428
0.7	0.5	2.85652

This thermal-aware loss function enables the model to better handle the challenges of thermal imagery, such as lower contrast, reduced texture information, and different edge characteristics compared to RGB images. By incorporating explicit knowledge about thermal image properties into the loss function, we can achieve better depth estimation without requiring architectural changes to the base model. Based on promising results from our initial experiments, we extended the training duration to 100 epochs with an optimized training strategy. The hyperparameters for this run can be observed in Table 2. We maintained the AdamW optimizer with a weight decay of 1e-4 and implemented a two-phase learning rate schedule: a 10-epoch warmup phase gradually increasing from 10% to 100% of the base learning rate, followed by a 90-epoch cosine annealing phase decreasing to 1e-7 as visualized in Figure 8. This approach prevented large unstable updates early in training while allowing for fine adjustments in later stages. Extended training with careful scheduling of the learning rate proved critical to fully adapting the model to thermal imagery characteristics and achieving our

Table 2. Hyperparameters used for thermal 3D vision model training

Category	Parameter	Value
Model Architecture	Encoder	ViT-Large (CROCO pretrained)
	Decoder	ViT-Base
	Input resolution	224×224
	Feature dimension	24
Optimization	Optimizer	AdamW
	Base learning rate	1e-4
	Weight decay	1e-4
	Batch size	8
	Training epochs	100
Learning Rate Schedule	Warmup epochs	10 (8% of total steps)
	Warmup start factor	0.02
	Annealing minimum LR	5e-7
Thermal-Aware Loss	Edge weight (α)	0.5
	Smoothness weight (β)	0.3
	Detail weight (γ)	0.4
	Confidence regularization	0.2
Data Sampling	Frame skip	3
	Train/Val split ratio	80/20

best performance results. The progression of loss during training is visible in Figure 7 with a gradual and consistent reduction in loss throughout the epochs, and the validation loss plateaued from the 79^{th} epoch.

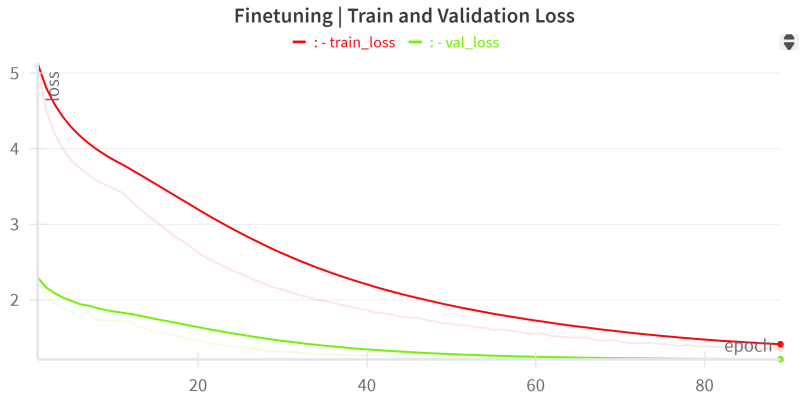

Fig. 7. A plot showing training and validation loss over epochs

As we can see in the Figure 9, the model was able to capture the depth information from the thermal image very well when compared to the ground truth.

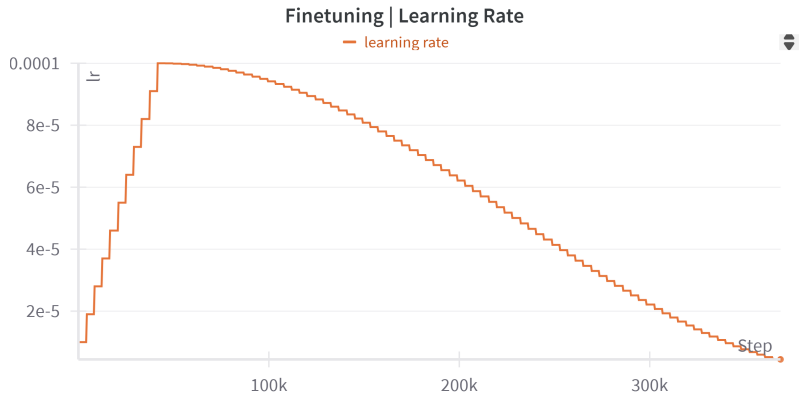

Fig. 8. A visualization of learning rate over time, showing the warmup phase and cosine annealing decay.

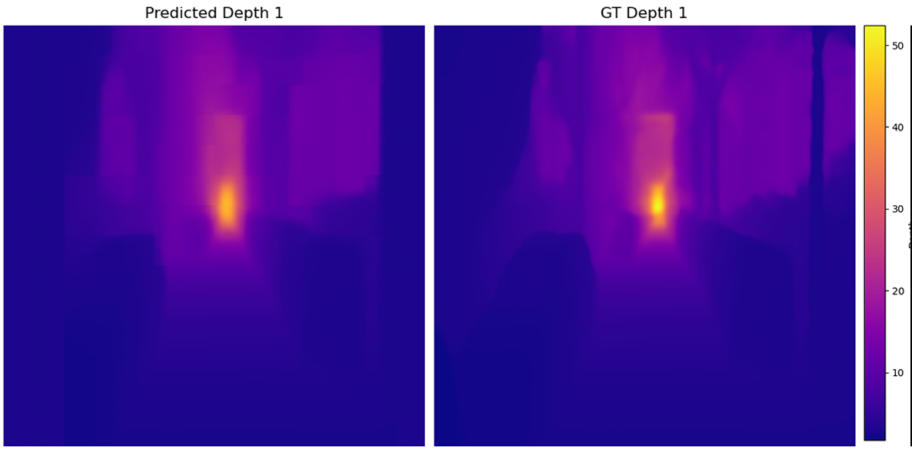

Fig. 9. Training of thermal image with fine-tuned DUSt3R base model with enhanced thermal loss metrics in 89 epochs

Figure 10 shows how the model performed during inference when fed with a thermal image that was not included in the training or validation phase.

4 Results

Each of the runs where evaluated with a test dataset from Freiburg dataset which thermal images taken in day and night. These images were used to create the test pseudo ground truth and then evaluated with the trained model to see how it performs.

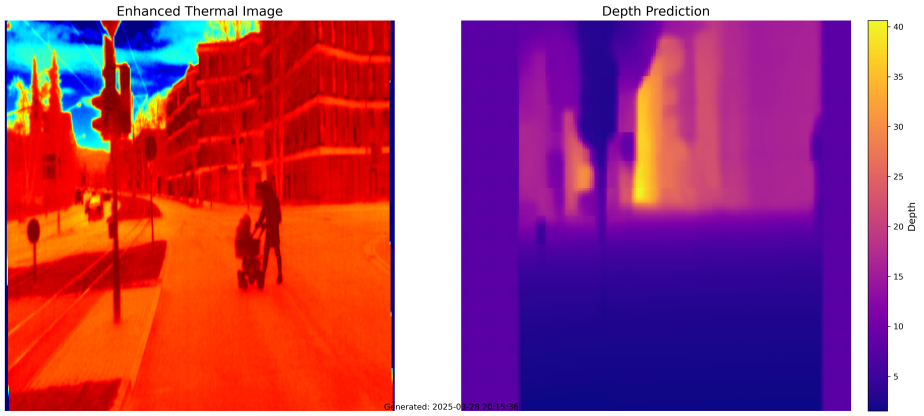

Fig. 10. Training of thermal image with fine-tuned DUSt3R base model with enhanced thermal loss metrics in 100 epochs

4.1 Main Results

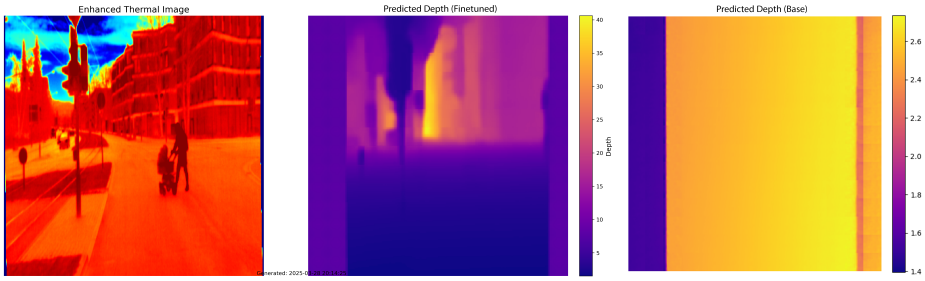

Fig. 11. Performance comparison between the DUSt3R model and the fine-tuned DUSt3R model on thermal images

We evaluated our fine-tuned model using the test set of thermal images. Figure 11 illustrates the performance of our model on a daytime image sample with our fine-tuned model and the base DUSt3R model. Furthermore, we conducted comprehensive testing on the entire dataset to obtain the quantitative results presented in Table 3.

This table presents a comprehensive comparison between the base DUSt3R model and our fine-tuned Thermal-DUSt3R model, evaluated on both day and night thermal imagery from the Freiburg dataset. The results demonstrate the significant impact of our thermal-specific adaptations. The base model performs considerably better on night thermal imagery than day imagery (RMSE of 3.67 vs 8.10). This likely reflects the stronger thermal gradients present in nighttime

Table 3. Comparison of thermal depth estimation performance between base and fine-tuned models

Model	**Setting**	**RMSE** ↓	**Acc[<1.25]** ↑	**Acc[<1.25^2]** ↑
Base DUSt3R	Day	8.1000	0.1850	0.3623
Thermal-DUSt3R	Day	3.7788	0.6031	0.8347
Base DUSt3R	Night	3.6695	0.2437	0.4624
Thermal-DUSt3R	Night	1.5553	0.7119	0.8744
Improvement (%)				
Day		53.4%	226.0%	130.4%
Night		57.6%	192.1%	89.1%

thermal imagery. Our thermal adaptations yield substantial improvements in both lighting conditions, showing the effectiveness of our approach regardless of time of day. The fine-tuned model achieves its best performance on night imagery (RMSE of 1.56), which demonstrates the particular suitability of our approach for low-light conditions where thermal imaging is especially valuable.

4.2 Ablation Studies

Table 4. Study of loss weights and their influence on depth metrics in models trained for 3 epochs on day test dataset

Configuration	**Edge**	**Smoothness**	**Detail**	**RMSE**↓	**Acc[<1.25]**↑	**Acc[<1.25^2]** ↑
Baseline	0.0	0.0	0.0	5.6806	0.3801	0.6551
Edge Only	0.5	0.0	0.0	5.7073	0.3777	0.6550
Smoothness Only	0.0	0.3	0.0	5.9349	0.3535	0.6211
Detail Only	0.0	0.0	0.4	5.7498	0.3679	0.6466
Full Model	0.5	0.3	0.4	5.6872	0.3648	0.6424

Table 5. Study of loss weights and their influence on depth metrics in models trained for 3 epochs on night test dataset

Configuration	**Edge**	**Smoothness**	**Detail**	**RMSE**↓	**Acc[<1.25]**↑	**Acc[<1.25^2]** ↑
Baseline	0.0	0.0	0.0	2.7189	0.3912	0.6678
Edge Only	0.5	0.0	0.0	2.7728	0.3833	0.6630
Smoothness Only	0.0	0.3	0.0	2.8868	0.3648	0.6412
Detail Only	0.0	0.0	0.4	2.7182	0.3854	0.6660
Full Model	0.5	0.3	0.4	2.7560	0.3746	0.6487

Loss Investigation As we can observe from Table 4 and Table 5, our ablation studies after just 3 epochs initially suggested that the baseline model outperformed our enhanced loss function. But our extended training revealed a

different pattern. The auxiliary loss terms (edge, smoothness, and detail) serve as regularizers that slow down initial convergence but ultimately lead to better generalization and improved performance on complete test sets.

When trained to convergence and evaluated on the complete test datasets, the full model demonstrated superior performance to any of the partial implementations. This suggests that the thermal-aware loss components work synergistically to capture the unique characteristics of thermal imagery, though this benefit requires sufficient training time to materialize.

Preprocessing Investigation As visible in Table 6, the percentage preprocessing method has performed better when compared to the other two configurations in the day and night datasets. We were expecting the fixed-range preprocessing as that method was the default for the Freiburg dataset. The percentage preprocessing method adaptively scales each image by using percentile values (2nd and 98th), which preserves local contrast while normalizing intensity variations across different thermal conditions. The fixed-range preprocessing, while being the default for the Freiburg dataset, applies a consistent intensity mapping across all images, which may not account for the varying thermal conditions between scenes. The without preprocessing configuration simply uses the raw thermal values.

Table 6. Effect of different preprocessing methods on depth estimation metrics

Preprocessing Method	**RMSE** ↓		**Acc**[<**1.25**] ↑		**Acc**[<**1.25**^{2}] ↑	
	Day	**Night**	**Day**	**Night**	**Day**	**Night**
Without Preprocessing	7.1088	3.3927	0.2622	0.2833	0.4934	0.5367
Percentage Preprocessing	**5.4904**	**2.4573**	**0.3997**	**0.4957**	**0.6737**	**0.7178**
Fixed Range Preprocessing	7.9140	3.5206	0.1988	0.2729	0.3807	0.5176

5 Conclusion

In this work, we presented a novel approach for adapting state-of-the-art 3D vision models to operate effectively with thermal imagery. Our findings demonstrate that with appropriate adaptations, deep learning models originally designed for RGB imagery can successfully transfer to the thermal domain, enabling robust 3D perception in challenging environmental conditions where visual cameras would fail.

Our primary contribution is the development of a thermal-adapted version of the DUSt3R framework, which achieves significant performance improvements over the base model. Specifically, our approach yields up to 53.4% reduction in RMSE for daytime thermal imagery and 57.6% for nighttime imagery, with accuracy improvements of 226% and 192.1% respectively as observable in Table 3. These

substantial gains validate our core hypothesis that thermal-specific adaptations are essential for bridging the domain gap between RGB and thermal imagery.

Despite these promising results, our approach has several limitations. First, the quality of our pseudo-ground truth depends heavily on the performance of the RGB-based MASt3R model, potentially propagating any biases or errors. Second, the thermal domain presents intrinsic challenges such as lower spatial resolution and lack of texture that fundamentally limit the achievable precision. Third, our current implementation does not explicitly handle dynamic temperature changes or extreme thermal conditions that might occur in real-world deployments.

In conclusion, our work demonstrates that thermal 3D vision is not only feasible but can achieve impressive performance when models are properly adapted to the unique characteristics of thermal imagery. This opens exciting possibilities for robust perception systems that can operate effectively in all lighting and weather conditions, a critical requirement for safety-critical autonomous systems.

6 Contributions

This project was conducted through collaborative pair programming between Pallavi Aithal Narayan and Salvin George, with both authors contributing to the core research concepts and analysis. While we collaborated closely on most aspects, we also focused on specific areas according to our strengths. Pallavi led the dataset exploration, visualization pipelines, and project documentation efforts, including the creation of the interactive overview notebook and systematic visualization techniques for result presentation. Salvin implemented the model architecture adaptations, thermal-aware loss functions, and evaluation metrics, including WandB integration.

Through pair programming, we jointly developed the pseudo-ground truth generation pipeline using MASt3R, the training infrastructure for handling the thermal dataset, and the evaluation framework that revealed our 226% accuracy improvements on daytime imagery. This collaborative approach enabled us to rapidly iterate on design choices while maintaining technical rigor throughout the experimental process, resulting in a model that effectively bridges the domain gap between RGB and thermal 3D vision.

References

1. Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25:245–262, 2014.
2. 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.
3. 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.