2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Thermal Image Refinement with Depth Estimation using Recurrent Networks for Monocular ORB-SLAM3

Hürkan Şahin[1], Huy Xuan Pham[2], Van Huyen Dang[1], Alper Yegenoglu[1], and Erdal Kayacan[1]

***Abstract*— Autonomous navigation in GPS-denied and visually degraded environments remains challenging for unmanned aerial vehicles (UAVs). To this end, we investigate the use of a monocular thermal camera as a standalone sensor on a UAV platform for real-time depth estimation and simultaneous localization and mapping (SLAM). To extract depth information from thermal images, we propose a novel pipeline employing a lightweight supervised network with recurrent blocks (RBs) integrated to capture temporal dependencies, enabling more robust predictions. The network combines lightweight convolutional backbones with a thermal refinement network (T-RefNet) to refine raw thermal inputs and enhance feature visibility. The refined thermal images and predicted depth maps are integrated into ORB-SLAM3, enabling thermal-only localization. Unlike previous methods, the network is trained on a custom non-radiometric dataset, obviating the need for high-cost radiometric thermal cameras. Experimental results on datasets and UAV flights demonstrate competitive depth accuracy and robust SLAM performance under low-light conditions. On the radiometric VIVID++ (indoor–dark) dataset, our method achieves an absolute relative error of approximately 0.06, compared to baselines exceeding 0.11. In our non-radiometric indoor set, baseline errors remain above 0.24, whereas our approach remains below 0.10. Thermal-only ORB-SLAM3 maintains a mean trajectory error under 0.4 m.**

***Index Terms*— Thermal imaging, Thermal-to-depth estimation, Recurrent neural networks, UAV, SLAM**

I. Introduction

In recent decades, research on autonomous UAVs has accelerated, broadening their range of applications, including environmental monitoring [1], [2], infrastructure inspection [3], [4], and disaster response [5]. In search and rescue missions, rescue robots and UAVs can rapidly access hazardous or confined spaces, reduce risks to responders, and use advanced sensors for monitoring and localization, thereby enhancing situational awareness and mission success [6]. Recent advances in autonomous aerial navigation favor lightweight, low-cost cameras over LiDAR, yet maintaining robustness in degraded or dark conditions remains a challenge. Thermal-infrared cameras thus provide key advantages under degraded conditions [7], as their working principle allows detecting infrared radiation without requiring light exposure, enabling penetration through smoke, dust, or haze.

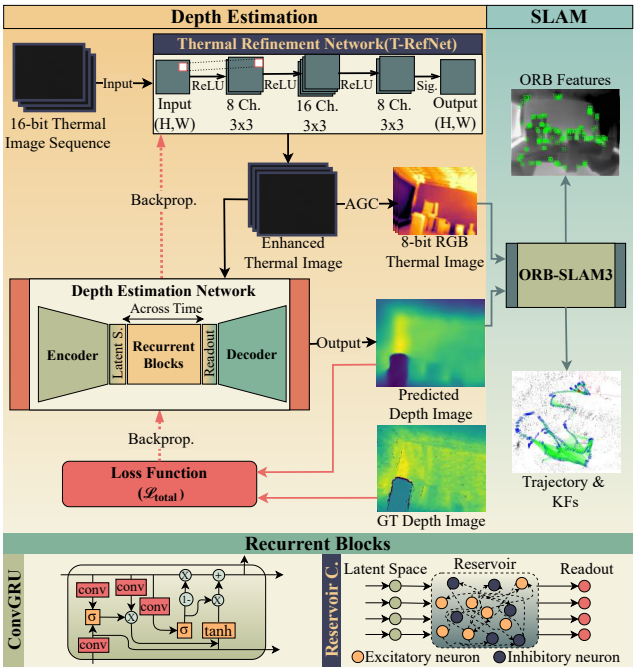


Fig. 1. Overview of the proposed thermal depth estimation pipeline. A raw 16-bit long-wave infrared (LWIR) image is first enhanced by the T-RefNet module, producing both an enhanced input for depth prediction and a color-mapped image for robust ORB-SLAM3 feature extraction. The encoder backbone extracts multi-scale features, which are processed by RBs (ConvGRU [9] or reservoir computing (RC) [10]) to enforce temporal consistency. Finally, the decoder outputs dense depth maps and enhanced thermal images integrated into ORB-SLAM3 [11] for robust feature extraction and metric-scale, temporally consistent tracking.

Although beneficial under degraded conditions, thermal cameras pose distinct challenges for reliable SLAM integration [8]. The 14/16-bit high dynamic range of thermal camera conflicts with 8-bit vision algorithms, automatic gain control (AGC) causes temporal inconsistencies, non-uniformity correction (NUC) interrupts streams, and low texture hampers feature detection. These factors necessitate specialized adaptation of thermal imagery for robust SLAM. To handle these challenges, we propose a novel framework (Fig. 1) that leverages recurrent thermal-to-depth modeling for monocular depth estimation from thermal imagery. The enhanced thermal images and reconstructed depth maps provide metric scale and temporally consistent priors that can be directly integrated into ORB-SLAM3, improving initialization, mapping accuracy, and real-time tracking for autonomous UAV navigation under extreme conditions.

*This work was partially supported by the Horizon Europe Grant Agreement No. 101136056 and No. 101070405, and Independent Research Fund Denmark, DFF-Research Project 1, with case number: 2035-00052B.

[1]Hürkan Şahin, Van Huyen Dang, Alper Yegenoglu, and Erdal Kayacan are with the Automatic Control Group (RAT), Paderborn University, 33098 Paderborn, Germany {hursah, van.huyen.dang, alper.yegenoglu, erdal.kayacan}@upb.de

[2]Huy Xuan Pham is with the Department of Electrical and Computer Engineering, Aarhus University, 8000 Aarhus C, Denmark, and also with Upteko ApS, Denmark huy.xuan@upteko.com

† Our dataset and source code are publicly available at https://hurkansah.github.io/thermal-depth-orbslam3/.

The key contributions of this paper are as follows:

- We propose a lightweight framework, T-RefNet, that leverages a recurrent unit to enhance thermal-to-depth conversion. Our framework can use ResNet [17], EfficientNet [18], and MobileNet [19] to serve as a backbone, combined with recurrent architectures: ConvGRU [20] and reservoir computing (RC) [10] to improve feature visibility and enforce temporal consistency in low-contrast and non-radiometric thermal imagery.
- We propose a non-radiometric thermal–depth UAV dataset to evaluate our framework, alongside existing radiometric public datasets such as VIVID++ [21].
- A comprehensive experimental study, including real-world experiments, is conducted to illustrate reliable performance in a thermal-only robot localization task across diverse trajectories and illumination settings, including fully dark environments where RGB-based SLAM typically fails.

The remainder of this paper is organized as follows. Section II summarizes related work, while Section III provides a brief overview of our methodology. Section IV details the experimental setup, dataset, and real-time results. Finally, conclusions are presented in Section V.

II. Related Work

The literature presents a wide range of thermal-based SLAM and visual odometry methods. As illustrated in Fig. 2, these approaches adopt diverse strategies to overcome the inherent challenges of thermal imaging, thereby enabling robust navigation in GPS-denied and visually degraded environments.

A feature-based monocular SLAM framework is proposed in [12] to address challenges in dynamic and visually degraded environments, combining thermal image denoising, semantic segmentation, and hybrid point-line tracking to improve robustness and accuracy. A fully self-supervised

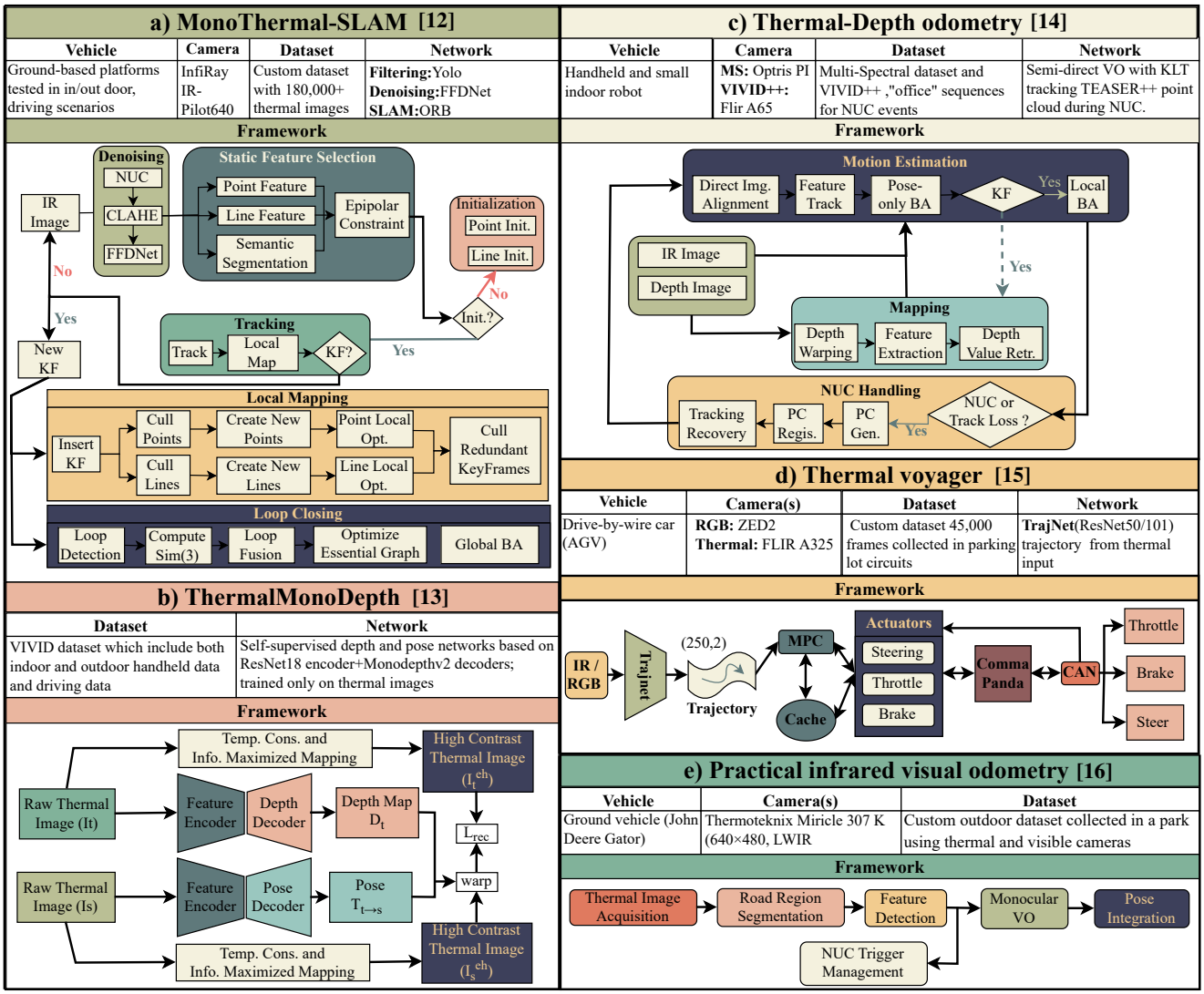


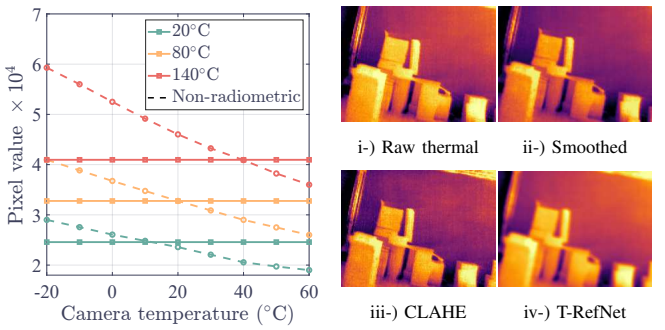
Fig. 2. Recent thermal navigation frameworks found in the literature span ground, handheld, and indoor platforms, across datasets from urban driving to parking-lot and outdoor road scenes. Representative approaches include feature/semantics-aware tracking and point–line SLAM [11], self-supervised depth–ego-motion [12], NUC handling [11,13,15], LWIR-based trajectory prediction with MPC [14], and road-segmentation–based scale recovery [15].

learning approach for estimating depth and ego-motion from monocular thermal video is proposed in [13], introducing a temporally consistent mapping technique to enhance contrast and structural information. A semi-direct VO system that fuses raw thermal and depth images with a dedicated NUC handling module for sensor disruption recovery is proposed in [14]. An end-to-end navigation pipeline using LWIR imagery and the deep learning model TrajNet for trajectory prediction under model predictive control is proposed in [15], enabling reliable nighttime operation. Finally, a monocular thermal visual odometry method for outdoor environments is proposed in [16], addressing scale ambiguity through road segmentation and mitigating NUC-induced pose estimation failures via a predictive trigger strategy.

The mentioned works highlight diverse strategies for addressing key challenges in thermal imaging, including low texture, NUC interruptions, and dynamic object interference, while extending navigation capabilities to low-light and GPS-denied environments. Our contribution diverges from these prior works in two ways. First, our approach simultaneously generates dense depth maps and enhances thermal images, enabling metric scale recovery and allowing for direct integration into existing SLAM frameworks without modification. Second, we present a non-radiometric thermal–depth UAV dataset, demonstrating robustness in fully dark indoor environments where conventional RGB-based methods fail. By incorporating recurrent modeling for temporal consistency and ensuring real-time deployment on embedded hardware, our framework emphasizes both the practical applicability and generalizability of thermal-based navigation.

III. Methodology

Non-radiometric thermal imagery presents unique challenges for depth estimation and SLAM, including low contrast, high dynamic range, and weak structural cues that hinder reliable feature extraction. To overcome these limitations, we propose a lightweight preprocessing network, T-RefNet, that refines thermal inputs and enhances their structural visibility. Integrated with a RB and a supervised depth decoder, the proposed pipeline enables temporally consistent and geometrically accurate depth predictions from thermal-only sequences, thereby facilitating robust SLAM performance. As illustrated in Fig. 1, the proposed system takes as input a raw 16-bit thermal image captured by an LWIR camera. To compensate for the inherently low contrast and high dynamic range of thermal data, a lightweight convolutional module, T-RefNet, is introduced to refine and normalize the input data. This module produces two complementary outputs: i) a normalized thermal image that serves as input to the supervised depth estimation backbone, and ii) an 8-bit color-mapped representation suitable for reliable ORB feature extraction within ORB-SLAM3. By providing both depth priors and texture-rich images, the system overcomes the limitations of raw thermal imagery, enabling robust SLAM operation with metric scale recovery.



(a) Radiometric and non-radiometric (b) Thermal enhancement techniques

Fig. 3. Comparison of the thermal image preprocessing methods. (a) Radiometric vs. non-radiometric thermal cameras at different TBB values. Solid lines represent radiometric outputs, while dashed lines indicate non-radiometric behavior.[1] (b) Thermal image enhancement techniques: i) Raw input suffers from noise that disrupts gradients; ii) Gaussian smoothing reduces noise but blurs edges; iii) CLAHE boosts local contrast but introduces spurious keypoints; iv) T-RefNet preserves edges while denoising, yielding stable features for SLAM.

A. Radiometric and non-radiometric thermal camera

Thermal imaging systems are either radiometric, delivering calibrated per-pixel temperatures for quantitative analysis, or non-radiometric, providing only relative contrast. Non-radiometric outputs are auto-scaled by frame content and internal temperature, so pixel values lack consistent physical meaning [22]. On the other hand, non-radiometric thermal cameras are low-cost, more accessible, and do not require continuous thermal calibration, while still providing sufficient relative contrast for navigation-focused tasks.

Figure 3a compares pixel responses of radiometric and non-radiometric cameras across varying target blackbody (TBB) temperatures. Radiometric thermal cameras produce consistent, near-linear outputs, enabling temperature-aware preprocessing such as contrast limited adaptive histogram equalization (CLAHE) or adaptive thresholds. By contrast, non-radiometric cameras re-map intensities frame by frame, causing histogram shifts, abrupt jumps with hot/cold regions, and unstable normalization.

To enhance thermal imagery for downstream vision tasks, different methods are compared in Fig. 3b. Raw 8-bit frames contain high-frequency noise, Gaussian smoothing reduces noise but blurs salient edges, and CLAHE improves contrast while amplifying spurious features. In contrast, the CNN-based T-RefNet produces denoised yet structurally consistent outputs, preserving contours and enabling stable feature extraction for SLAM.

B. Training flow of the depth estimation

The training procedure for the proposed T-RefNet-based sequence depth estimation model is described in Algorithm 1. At each timestep, the input thermal frame is first refined by the T-RefNet module and then encoded into multi-scale features. These features are passed through the RB to capture temporal context and finally decoded into a depth map.

[1]FLIR Boson: `https://oem.flir.com/products/boson`

Algorithm 1: Training flow of the refinement–sequence depth estimation.

Input: Thermal sequences $\{x_t\}_{t=1}^{T}$, and the ground-truth depths $\{z_t^{gt}\}_{t=1}^{T}$

1. Initialize parameters: θ (T-RefNet), ϕ (Enc–Dec), ψ (RB), lr η

for $t \leftarrow 1$ **to** T **do**

$y_t \leftarrow f_\theta(x_t)$;

$\{h_{t,\ell}^{enc}\}_{\ell=0}^{L} \leftarrow \text{Encoder}_\phi(y_t)$;

$h_t^{in} \leftarrow \text{F}_{\text{latent.S}}(h_{t,L}^{enc})$;

$h_t^{RB} \leftarrow \text{W}_\psi(h_t^{in}, h_{t-1}^{RB})$;

$h_{t,L}^{enc} \leftarrow \text{F}_{\text{readout}}(h_t^{RB})$;

$\hat{z}_t \leftarrow \text{Decoder}_\phi\big(\{h_{t,\ell}^{enc}\}_{\ell=0}^{L}\big)$;

2. Calculate loss function

$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{SIlog}} + \lambda_2\mathcal{L}_{\text{SSIM}} + \lambda_3\mathcal{L}_{\text{ord}} + \lambda_4\mathcal{L}_{\text{sm}}$;

3. Update the weights:

$\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\text{total}}$;

$\phi \leftarrow \phi - \eta\nabla_\phi\mathcal{L}_{\text{total}}$;

$\psi \leftarrow \psi - \eta\nabla_\psi\mathcal{L}_{\text{total}}$;

The model parameters are updated end-to-end with a composite loss designed to enforce both geometric consistency and perceptual accuracy. Building on the concept of combined loss formulations from prior work [13], we extend this idea by integrating multiple complementary objectives into a single framework. Specifically, the loss includes: i) a scale-invariant term $\mathcal{L}_{\text{SIlog}}$ [23], assigned the largest weight (0.9) to capture the global depth structure; ii) a perceptual similarity term $\mathcal{L}_{\text{SSIM}}$ [24], weighted 0.4 to preserve local structures and textures; iii) a depth-ordering term $\mathcal{L}_{\text{ord}}$ [25], weighted 0.1 to enforce correct relative ordering between pixels; and iv) an edge-aware smoothness term $\mathcal{L}_{\text{sm}}$ [26], also weighted 0.1, to regularize depth predictions while respecting image boundaries. This formulation improves stability and accuracy in thermal depth estimation.

To maintain efficiency and real-time capability, the encoder backbone is instantiated with lightweight architectures such as EfficientNet-B0, MobileNet, or ResNet-8, offering a favorable trade-off between accuracy and computational cost. For temporal modeling, the refined thermal sequence is further processed by either a ConvGRU bottleneck or a reservoir computing based network, both integrated into the depth estimation network. These recurrent architectures capture frame-to-frame dependencies and enforce temporal consistency across predictions, which is essential for stable SLAM operation in dynamic or texture-poor thermal environments. As a result, SLAM initialization becomes more reliable, mapping accuracy improves, and real-time tracking performance is enhanced.

C. Reservoir computing

RC constitutes a recurrent neural network paradigm that represents and processes temporal and sequential data [10]. Fundamentally, RC operates by embedding input signals into a high-dimensional state space via a randomly connected recurrent network of non-linear neurons. This state space inherently captures the temporal dependencies of the input, while the readout layer maps the reservoir dynamics onto the desired output.

Let $\mathbf{u}(t) \in \mathcal{R}^K$ be the input at time t with K input neurons. The internal state of the reservoir, $\mathbf{x}(t) \in \mathcal{R}^N$, is expressed as

$$\mathbf{x}(t+1) = f\left(\mathbf{W}_{in}\mathbf{u}(t+1) + \mathbf{W}\mathbf{x}(t)\right), \tag{1}$$

where f is a sigmoidal function, $\mathbf{W}_{in} \in \mathcal{R}^{N\times K}$ represents the input matrix, $\mathbf{W} \in \mathcal{R}^{N\times N}$ the weight matrix of the reservoir, and $\mathbf{y} \in \mathcal{R}^L$ is the output signal. Then the output is computed as

$$\mathbf{y}(t+1) = f^{out}\left(\mathbf{W}_{out}\mathbf{x}(t+1)\right), \tag{2}$$

with $\mathbf{W}_{out} \in \mathcal{R}^{L\times N}$. We base our reservoir implementation on [27], which uses a biologically realistic representation of neurons, namely the leaky-integrate and fire (LIF) neuron. The reservoir layer consists of a vector of membrane potentials of N of excitatory and inhibitory LIF neurons $\mathbf{v}(t) \in \mathcal{R}^N$. A differential equation describes the LIF neuron as [28]:

$$\tau_m \frac{dV(t)}{dt} = -V(t) + R_m I(t), \tag{3}$$

where $V(t)$ is the membrane potential at time t, τ_m is the membrane time constant R_m is the membrane resistance and $I(t)$ is the input current at time t.

IV. Experiments

In this section, we present the thermal-to-depth estimation results of the proposed model, followed by an evaluation of its integration into ORB-SLAM3 across various trajectories and scenes, using both UAV and handheld devices to highlight the advantages for robust localization.

A. Evaluation and baselines

To comprehensively evaluate the proposed approach, we conducted experiments on two different thermal–depth datasets: (i) the indoor-dark subset of VIVID++ [21], which is recorded with a radiometric thermal camera, and (ii) a custom dataset collected with a non-radiometric thermal sensor and a depth camera. This dual evaluation setup enables us to assess performance under both radiometric and non-radiometric conditions. The thermal data were captured with a Flir Boson+[1] non-radiometric shuttered camera (640×512). The dataset comprises approximately 65,000 samples, covering diverse lighting conditions—bright, dark, and semi-lit—and including scenes with both hot and cold objects to improve robustness across thermal distributions.

For comparison, we include both RGB-trained depth estimation networks (ZoeDepth [29], DepthAnything-V2 [30]) and thermal-specific approaches from the literature [13], [31]–[33]. Since ZoeDepth and DepthAnything-V2 were trained on RGB images, we pre-processed our thermal inputs by mapping them to RGB format before inference. Regarding [13], we retrained and evaluated the model on our non-radiometric dataset using the sequences we collected. We use

TABLE I

QUANTITATIVE COMPARISON OF DEPTH ESTIMATION ACCURACY ACROSS DIFFERENT ARCHITECTURES ON THE INDOOR-DARK SUBSET OF THE VIVID++ [21] DATASET. BEST VALUES ARE SHOWN IN BOLD.

Model	AbsRel	RMSE	a1	a2	a3
Shin (T) [31]	0.232	0.740	0.618	0.907	0.987
Shin (MS) [31]	0.166	0.566	0.768	0.967	0.994
Shin (Max.) [13]	0.149	0.517	0.813	0.969	0.994
ZoeDepth [29]	0.165	0.533	0.788	0.944	0.991
DepthAnything-V2 [30]	0.112	0.378	0.902	0.970	0.990
Ye et al. [33]	0.145	0.499	0.827	0.969	0.994
MSDFNet [32]	0.139	0.470	0.847	**0.980**	**0.996**
Eff-B0 noRB (ours)	0.139	0.497	0.839	0.945	0.984
Eff-B0+GRU noTRN (ours)	0.079	0.325	0.929	**0.980**	0.995
ResNet8+GRU (ours)	0.079	0.345	0.913	0.970	0.990
MobileNet+GRU (ours)	0.072	0.318	0.928	0.977	0.993
Eff-B0+GRU (ours)	**0.063**	**0.298**	**0.940**	**0.980**	0.993
Eff-B0+RC (ours)	0.069	0.313	0.931	0.976	0.993

Eff-B0: EfficientNet-B0 backbone; noRB: without recurrent block; GRU: ConvGRU; noTRN: without T-RefNet.

TABLE II

EVALUATION RESULTS OF THERMAL-TO-DEPTH NETWORKS ON A CUSTOM INDOOR DATASET ACQUIRED WITH NONRADIOMETRIC THERMAL AND DEPTH SENSORS. BEST VALUES ARE SHOWN IN BOLD.

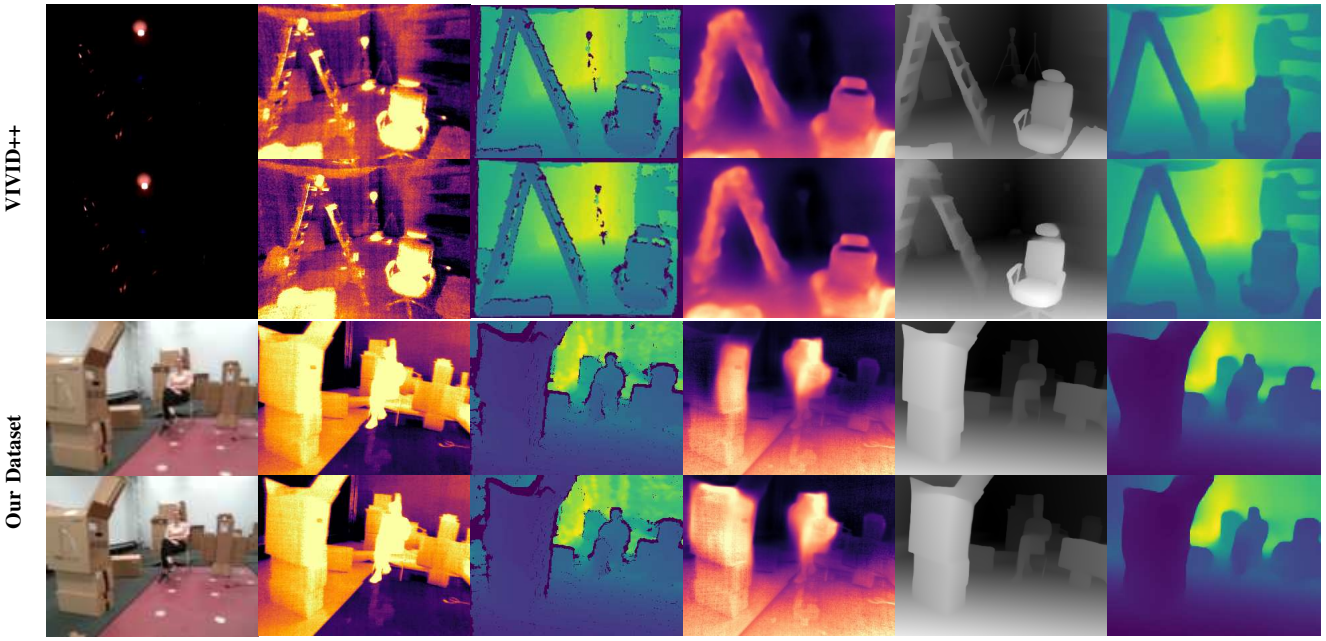
Model	AbsRel	RMSE	a1	a2	a3
Shin (Max.) [13]	0.262	1.273	0.589	0.890	0.960
ZoeDepth [29]	0.243	1.110	0.605	0.885	0.954
DepthAnything-V2 [30]	0.267	1.043	0.571	0.863	0.931
ResNet8+GRU (ours)	0.109	0.516	0.886	0.943	0.969
MobileNet+GRU (ours)	0.085	0.453	0.911	0.951	0.971
Eff-B0+GRU (ours)	0.079	**0.424**	0.920	0.955	0.971
Eff-B0+RC (ours)	**0.076**	0.439	**0.929**	**0.965**	**0.981**

key metrics to quantitatively evaluate our methods against the baselines, such as mean absolute relative error (AbsRel), root mean squared error (RMSE), and accuracy under thresholds $1.25, 1.25^2, 1.25^3$ (a1, a2, a3).

B. Thermal-to-depth estimation results

Table I shows the quantitative results on the VIVID++ indoor-dark subset. The methods demonstrate competitive performance, with MSDFNet [32] achieving the best $a2$ (0.980) and $a3$ accuracy (0.996). Among general-purpose RGB models, DepthAnything-V2 [30] yields strong results (AbsRel $= 0.112$, RMSE $= 0.378$). However, as illustrated in Fig. 4e, its predictions are not entirely consistent: Although it preserves sharpness in some static scenes, during motion it often blurs structures, removes objects, and hinders accurate depth analysis. In contrast, our architectures significantly outperform all baselines on most metrics. In particular, our model with the EfficientNet-B0 encoder achieves the best results with AbsRel $= 0.063$, RMSE $= 0.298$, and $a1 = 0.940$, highlighting the effectiveness of the combined ConvGRU and T-RefNet modules. In addition, the RC variant delivers results close to the best model in the radiometric data set, while using only about 50k parameters (including its latency space block and readout) with 32 reservoir neurons compared to over 800k parameters required for ConvGRU and its corresponding components, making it a lightweight yet competitive alternative.

Table II presents the evaluation on the proposed non-radiometric dataset, which is considerably more challenging due to fluctuating pixel intensities caused by auto-scaling and internal heating. The models trained purely on RGB inputs



(a) RGB (b) Thermal (c) Ground Truth (d) Shin et al. [13] (e) DepthAnything [30] (f) Ours (RC+Eff-B0)

Fig. 4. Qualitative comparison across two datasets. Top: VIVID++; bottom: our dataset. Each row shows two temporally adjacent frames. Columns: (a) RGB, (b) thermal, (c) thermal-aligned ground-truth depth, (d) Shin et al. [13], (e) DepthAnything-V2 [30] (RGB-only), (f) Our representative proposed model with RC.

(ZoeDepth, DepthAnything-V2) perform poorly, with higher AbsRel and lower accuracies. Regarding [13], as illustrated in Fig. 4d, the method employs radiometric-specific preprocessing and performs reasonably on the VIVID++ dataset, but on non-radiometric data its consistency degrades when hot or cold objects enter or leave the scene. In contrast, our models remain robust, with RC combined with EfficientNet-B0 giving slightly better results on the non-radiometric dataset (AbsRel $= 0.076$, $a1 = 0.929$), as also shown in Fig. 4. Although the difference compared to the GRU-based model is not large, it is notable that the RC variant, with a lighter architecture, performs better in the more variable non-radiometric conditions. These findings underscore the importance of radiometric invariance and demonstrate that our approach generalizes well across both radiometric and non-radiometric settings.

C. Localization results

To evaluate the practical applicability of our thermal-based preprocessing and depth estimation pipeline within a visual SLAM framework, we integrated the outputs of both the T-RefNet and the depth prediction network into the ORB-SLAM3 framework. The goal is to evaluate how well these outputs support localization and mapping under thermal-only input conditions.

All experiments are conducted offline using ROS bag files. Thermal and RGB-D images were captured and stored in real time, and the synchronized data streams were recorded into '.bag' files. The stored bags were then played back for evaluation to ensure consistent and reproducible conditions.

In low-light indoor scenarios, the quality of input images directly affects the ability of ORB-SLAM3 to maintain reliable tracking. As illustrated in Fig. 5, directly converting raw 16-bit thermal images into an 8-bit format results in frequent loss of structural features due to low contrast and high noise levels. The detected features are too sparse and unstable to support consistent tracking, making it impossible for ORB-SLAM3 to generate a meaningful trajectory. In dark illumination conditions, RGB images also fail to provide sufficient structure for reliable feature extraction, as shown in the top row of Fig. 5. In contrast, T-RefNet thermal frames maintain edge consistency and enhance feature visibility, facilitating stable keypoint detection and accurate pose estimation. Consequently, except for evaluations in bright environments, only the T-RefNet image was used in dark scenarios, while raw 8-bit thermal and RGB inputs were excluded from quantitative analysis.

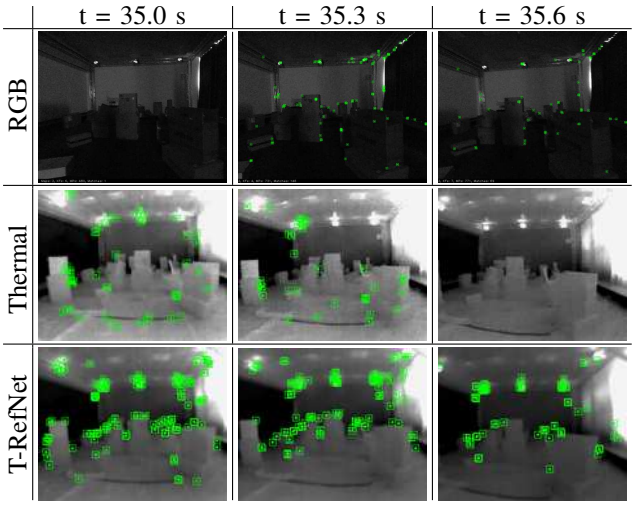
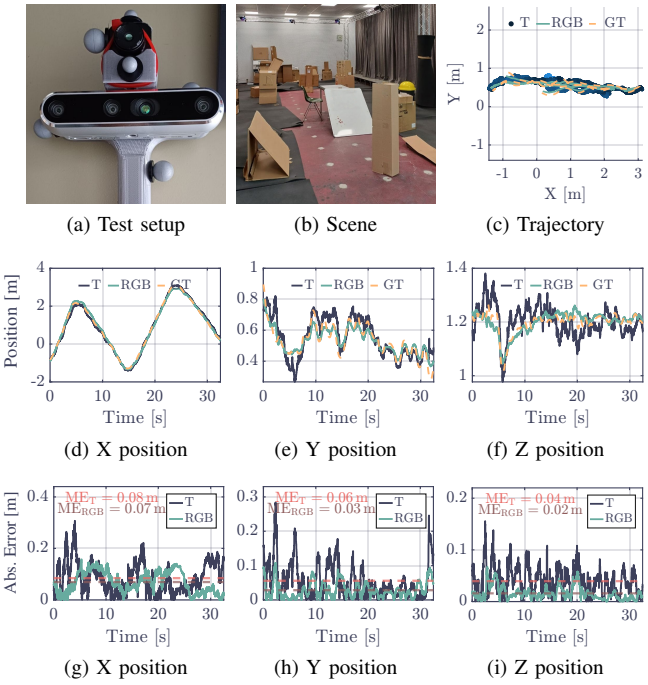


Fig. 5. Feature tracking results of ORB-SLAM3 using different image inputs: RGB images (top row), raw 8-bit thermal images (middle row), and T-RefNet enhanced thermal images (bottom row). While RGB features degrade under low-light indoor conditions, raw thermal inputs suffer from noise and low contrast. In contrast, T-RefNet outputs provide more stable and repeatable features, leading to improved tracking robustness.

Three evaluation scenarios are designed to assess the proposed pipeline under varying conditions: i) a bright environment with abundant features and linear motion, comparing RGB-D and estimated thermal depth; ii) a dark environment with circular motion and fewer features, evaluating thermal depth performance; and iii) a UAV-based test across corridors with varying illumination, where each corridor presents different lighting conditions.

In the bright environment scenario with linear back-and-forth motion (Fig. 6), RGB-D naturally achieves higher accuracy than thermal depth due to the abundance of detectable features. Nevertheless, since the motion is simple and the scene provides rich structural cues, both methods produce stable trajectories. The handheld setup further reduces motion jitter, resulting in consistent tracking across all axes.

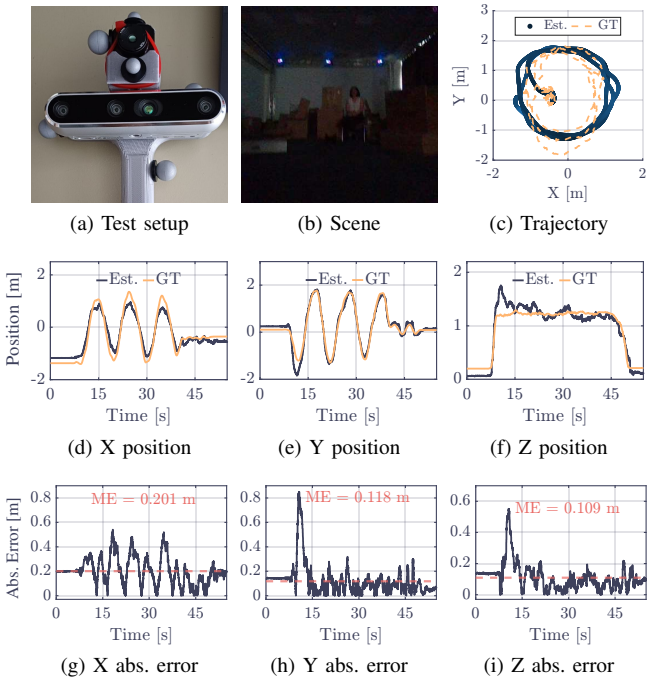


(a) Test setup (b) Scene (c) Trajectory

(d) X position (e) Y position (f) Z position

(g) X position (h) Y position (i) Z position

Fig. 6. Experimental results in a bright indoor scene with a handheld device. (a) Test setup with RGB-D and thermal cameras, (b) sample scene view, and (c) ground-truth trajectory. (d–f) Estimated trajectories along the X, Y, and Z axes from ORB-SLAM3 with RGB-D and T-RefNet refined thermal input are compared against ground truth. (g–i) Absolute position errors along each axis are shown, with mean error (ME) values highlighted for both RGB-D and T-RefNet inputs.

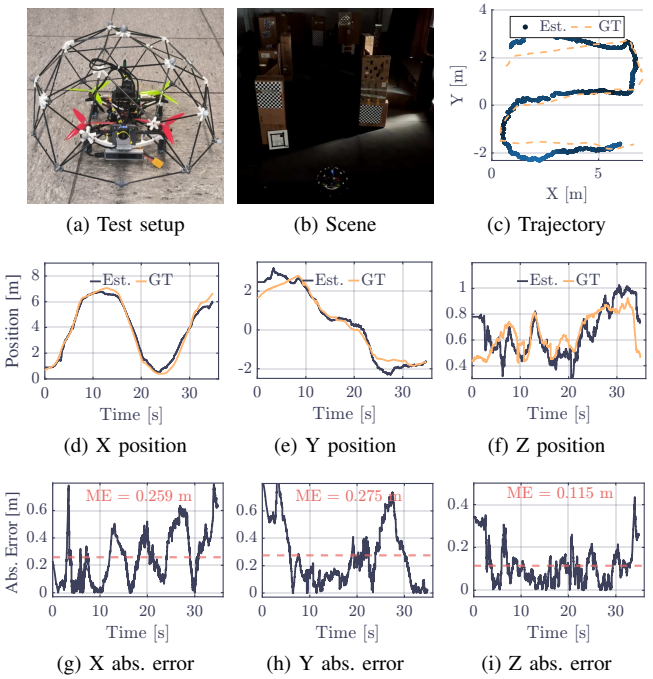


(a) Test setup (b) Scene (c) Trajectory

(d) X position (e) Y position (f) Z position

(g) X abs. error (h) Y abs. error (i) Z abs. error

Fig. 7. Experimental results in a dark indoor scene with a handheld device performing circular motion. (a) Handheld device, (b) sample scene view, and (c) ground-truth trajectory. (d–f) Estimated trajectories along the X, Y, and Z axes from ORB-SLAM3 with T-RefNet are compared against ground truth. (g–i) Absolute position errors are shown for each axis, with ME values highlighted.



(a) Test setup (b) Scene (c) Trajectory

(d) X position (e) Y position (f) Z position

(g) X abs. error (h) Y abs. error (i) Z abs. error

Fig. 8. Experimental results in a dark indoor scene with UAV. (a) UAV with thermal cameras, (b) sample scene view, and (c) ground-truth trajectory. (d–f) Estimated trajectories along the X, Y, and Z axes from ORB-SLAM3 with T-RefNet are compared against ground truth. (g–i) Absolute position errors are shown for each axis, with ME values highlighted.

Quantitatively, the Euclidean mean error is about 0.11 m for thermal depth and about 0.08 m for RGB-D indicating that while RGB-D has an advantage, the T-RefNet–enhanced thermal input with estimated depth still provides sufficiently accurate estimates for reliable localization.

In the dark circular motion scenario in Fig. 7, features cluster on one side of the scene, causing uneven keypoint coverage and degraded tracking. The X-axis shows the most significant deviations, with a mean error of approximately 0.20 m and frequent peaks of up to 0.70 m. Y and Z are more stable than X overall, with mean errors of about 0.12 m and 0.11 m; however, Y exhibits rare spikes up to 0.80 m, whereas Z's transients remain within about 0.50 m. These peaks are less frequent than in X, where fluctuations occur more consistently. Aggregated over all axes, the Euclidean mean error is about 0.26 m, highlighting the accuracy loss in circular motion under low light. Figure 5 further shows that ORB-SLAM3 with RGB input alone fails to maintain tracking in this scenario.

In the corridor experiment (Fig. 8), the UAV flies through three different corridors: two in complete darkness and one with mixed illumination, where certain regions were lit while others remained in shadow. Within and at the end of each corridor, distinctive structures were placed to ensure the presence of detectable features. Additionally, aluminium foil strips were attached to selected surfaces to create low-emissivity targets, which appear as cold objects in thermal images even when they are at room temperature.

ORB-SLAM3 with T-RefNet enhanced thermal input successfully tracked the UAV's trajectory, achieving a Euclidean mean error of approximately 0.39 m across the three axes. This error reflects the increased difficulty posed by uneven lighting, cold-object distractors, and narrow passages. Nonetheless, the system maintained a continuous trajectory estimate, demonstrating robustness under mixed illumination and in the presence of thermally deceptive objects.

In summary, the three evaluation scenarios demonstrate that the proposed pipeline delivers stable localization across diverse conditions. In the bright feature-rich environments, thermal depth achieves accuracy comparable to RGB-D. In the dark circular motion, it sustains tracking with moderate accuracy loss where RGB completely fails. In UAV corridor flights with mixed illumination and distractors, it maintains continuous trajectories within sub-0.4 m error. These results validate the robustness of the method under both handheld and UAV setups in challenging environments.

V. Conclusions and Future Work

This paper presents a thermal-based depth estimation and SLAM framework for navigation in GPS-denied, low-light environments. A lightweight thermal-to-depth network with recurrent blocks, including RC, was trained on radiometric VIVID++ and custom non-radiometric datasets, achieving state-of-the-art accuracy across both. Unlike prior methods that degrade on non-radiometric data, our recurrent design maintains temporal consistency under noise and low texture, requiring only ∼50k parameters versus ∼800k for

ConvGRU. For localization, ORB-SLAM3 with T–RefNet enhanced thermal inputs yielded robust trajectories in both the handheld and the UAV experiments. In contrast, raw thermal or RGB inputs failed in darkness, demonstrating that the proposed pipeline extends SLAM to conditions where conventional vision breaks down.

Despite its robustness, the framework can encounter challenges when the number of detected features is low, resulting in occasional tracking loss. Furthermore, artifacts inherent to thermal cameras, such as NUC, may cause temporary intensity changes, disrupting feature stability and tracking. While real-time operation is feasible on embedded hardware for short sequences and moderate motion, prolonged or more dynamic scenarios still pose difficulties. As future work, we aim to mitigate these limitations by improving robustness against NUC artifacts, optimizing the pipeline for embedded platforms, and integrating depth estimation with obstacle avoidance modules to enable autonomous drone navigation and full trajectory estimation.

REFERENCES

[1] J. G. Hansen, M. Heiß, D. Li, M. Kozłowski, and E. Kayacan, "Vessel inspection in-the-wild: Practical planning in large-scale industrial environments," in *2023 American Control Conference (ACC)*, 2023, pp. 812–817.

[2] J. G. Hansen, M. Heiß, M. Kozlowski, and E. Kayacan, "UAV trajectory evaluation in large industrial environments: A cost-effective solution," in *2022 European Control Conference (ECC)*, 2022, pp. 1336–1341.

[3] A. Amer, M. Mehndiratta, J. le Fevre Sejersen, H. X. Pham, and E. Kayacan, "Visual tracking nonlinear model predictive control method for autonomous wind turbine inspection," in *2023 21st International Conference on Advanced Robotics (ICAR)*, 2023, pp. 431–438.

[4] H. X. Pham and E. Kayacan, "FROST: Fusion and multimodal 3d reconstruction of icy surfaces for robotic exploration," in *2025 IEEE Symposium on Computational Intelligence on Engineering/Cyber Physical Systems Companion (CIES Companion)*, 2025, pp. 1–5.

[5] A. Q. Nguyen, H. T. Nguyen, V. C. Tran, H. X. Pham, and J. Pestana, "A visual real-time fire detection using single shot multibox detector for uav-based fire surveillance," in *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*, 2021, pp. 338–343.

[6] M. Lyu, Y. Zhao, C. Huang, and H. Huang, "Unmanned aerial vehicles for search and rescue: A survey," *Remote Sensing*, vol. 15, no. 13, 2023.

[7] Y.-S. Shin and A. Kim, "Sparse depth enhanced direct thermal-infrared SLAM beyond the visible spectrum," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2918–2925, 2019.

[8] J. Jiang, X. Chen, W. Dai, Z. Gao, and Y. Zhang, "Thermal-inertial SLAM for the environments with challenging illumination," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8767–8774, 2022.

[9] J. Yang, W. Jinxin, X. Li, and X. Qin, "Tool wear prediction based on parallel dual-channel adaptive feature fusion," *The International Journal of Advanced Manufacturing Technology*, vol. 128, pp. 1–21, 07 2023.

[10] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German national research center for information technology gmd technical report*, vol. 148, no. 34, p. 13, 2001.

[11] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: an accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[12] Y. Wu, L. Wang, L. Zhang, Y. Bai, Y. Cai, S. Wang, and Y. Li, "Improving autonomous detection in dynamic environments with robust monocular thermal SLAM system," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 203, pp. 265–284, 2023.

[13] U. Shin, K. Lee, B.-U. Lee, and I. S. Kweon, "Maximizing self-supervision from thermal image for effective self-supervised learning of depth and ego-motion," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7771–7778, 2022.

[14] X. Chen, W. Dai, J. Jiang, B. He, and Y. Zhang, "Thermal-depth odometry in challenging illumination conditions," *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 3988–3995, 2023.

[15] A. NG, D. PB, J. Shalabi, S. Jape, X. Wang, and Z. Jacob, "Thermal Voyager: a comparative study of RGB and thermal cameras for nighttime autonomous navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14 116–14 122.

[16] P. V. K. Borges and S. Vidas, "Practical infrared visual odometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 8, pp. 2205–2213, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[18] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[20] T. Ma, L. Zhang, X. Diao, and O. Ma, "ConvGRU in fine-grained pitching action recognition for action outcome prediction," 2020. [Online]. Available: https://arxiv.org/abs/2008.07819

[21] A. J. Lee, Y. Cho, Y. sik Shin, A. Kim, and H. Myung, "ViViD++: Vision for visibility dataset," 2022. [Online]. Available: https://arxiv.org/abs/2204.06183

[22] FLIR Systems. (2021) The benefits and challenges of radiometric thermal technology. Accessed: 2025-04-22. [Online]. Available: https://www.flir.com/discover/security/radiometric/the-benefits-and-challenges-of-radiometric-thermal-technology

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *CoRR*, vol. abs/1406.2283, 2014. [Online]. Available: http://arxiv.org/abs/1406.2283

[24] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," 2019. [Online]. Available: https://arxiv.org/abs/1806.01260

[25] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, "Structure-guided ranking loss for single image depth prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 608–617.

[26] C. Xu, B. Huang, and D. S. Elson, "Self-supervised monocular depth estimation with 3-D displacement module for laparoscopic images," *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 2, pp. 331–334, 2022.

[27] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu, "Liquid time-constant networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7657–7666.

[28] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press, 2014.

[29] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "ZoeDepth: zero-shot transfer by combining relative and metric depth," 2023. [Online]. Available: https://arxiv.org/abs/2302.12288

[30] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," 2024. [Online]. Available: https://arxiv.org/abs/2406.09414

[31] U. Shin, K. Lee, S. Lee, and I. S. Kweon, "Self-supervised depth and ego-motion estimation for monocular thermal video using multi-spectral consistency loss," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1103–1110, 2022.

[32] L. Kong, Q. Zheng, and W. Wang, "MSDFNet: multi-scale detail feature fusion encoder–decoder network for self-supervised monocular thermal image depth estimation," *Measurement Science and Technology*, vol. 36, no. 1, p. 016039, dec 2024.

[33] X. Ye, X. Mao, R. Xu, and H. Li, "Mining scene structural guidance for thermal images in self-supervised monocular depth estimation," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.