A Framework for Indoor-Monocular Depth Estimation on UAVs Using a Thermal Camera

Carlos M. Perez-Rodriguez*[1], Cesar Martinez-Torres[1] , and Jose Martinez-Carranza[2]
[1] Universidad de las Americas Puebla, Puebla, Mexico
[2]Instituto Nacional de Astrofisica, Optica y Electronica, Puebla, Mexico

ABSTRACT

Monocular depth estimation is essential for autonomous UAV navigation, especially under low-visibility conditions. While thermal cameras enable perception regardless of lighting, their use for geometric reasoning is challenging due to intensity variations driven by object temperature and emissivity.

We propose a novel framework that transforms thermal images into a temperature-invariant structural representation using line segments. This preserves scene geometry while discarding volatile thermal data. The structural image is then processed by the AnyDepth neural network to predict dense depth maps.

Experiments on the OdomBeyondVision dataset, featuring indoor UAV flights with thermal sensors, show that our method enables accurate depth estimation. This demonstrates the viability of decoupling geometry from thermal intensity, paving the way for robust UAV navigation in challenging scenarios such as search and rescue.

1 INTRODUCTION

Unmanned Aerial Vehicles (UAVs) are increasingly employed in critical missions such as search and rescue, environmental monitoring, and infrastructure inspection. A fundamental requirement for enabling autonomy in these contexts is the ability to perceive depth from visual inputs. Monocular depth estimation, which infers 3D structure from a single image, provides a lightweight solution that is particularly suitable for small UAVs where payload capacity is limited [1, 2].

While RGB-based depth estimation has seen significant progress with deep learning approaches [3, 4], thermal cameras remain underexplored for this task, despite their inherent advantages. Unlike RGB sensors, thermal imaging enables perception in total darkness, fog, smoke, or other visually degraded environments [5]. These capabilities make thermal cameras ideal for operations in hazardous or low-visibility scenarios. However, their use for geometric reasoning is challenging, as thermal intensity is governed by object temperature and emissivity rather than stable photometric features. This introduces issues such as thermal crossover and spatial inconsistency, which undermine the performance of conventional depth estimation models [6].

To address these limitations, prior research has attempted to fuse thermal and RGB data [7], or apply domain adaptation techniques to transfer knowledge from visible to thermal domains [5]. Nonetheless, these approaches often require additional sensors or paired data, increasing system complexity and cost. An alternative strategy is to extract meaningful geometric information directly from thermal images, thereby decoupling geometry from thermal intensity altogether.

In this work, we propose a novel indoor-monocular depth estimation framework tailored for thermal images. Our approach converts the input thermal image into a temperature-invariant structural representation by extracting line segments. This representation captures the essential scene geometry while discarding unstable thermal cues. The resulting structural image is then processed by the AnyDepth-v2 deep neural network [8, 9] to predict a dense depth map.

We evaluate our framework using the OdomBeyondVision dataset [10], which contains indoor UAV sequences recorded with thermal cameras. Experimental results show that our method enables accurate depth estimation from thermal input alone, without requiring additional sensors or modalities. This opens the door to robust indoor UAV navigation, particularly in scenarios where thermal sensors are already deployed (such as search and rescue, firefighting, and industrial inspection).

The remainder of this paper is organized as follows: Section 2 reviews related work on monocular depth estimation and thermal image processing. Section 3 details our proposed framework, including preprocessing, structural extraction, and the depth prediction pipeline. Section 4 presents our experimental setup, dataset, and evaluation results. Section 5 discusses the implications and limitations of our approach, and Section 6 concludes the paper with directions for future research.

2 RELATED WORK

2.1 Depth Estimation

The field of monocular depth estimation was revolutionized by deep learning, which has largely superseded classical geometric methods. The pioneering work of Eigen et al. [11] was the first to successfully demonstrate that a multi-scale Convolutional Neural Network (CNN) could regress a dense

*Email address(es): carlos.perezro@udlap.mx

depth map directly from a single RGB image in a supervised fashion. This established the foundational paradigm for end-to-end depth learning.

Shortly after, a second major paradigm emerged with the work of Garg et al. [12], which introduced self-supervised learning for this task. By using photometric consistency between stereo image pairs as a supervisory signal, they eliminated the need for expensive ground-truth data from active sensors like LiDAR. This breakthrough paved the way for modern, robust self-supervised methods. Among them, Monodepth2 [13] became a fundamental and widely-used baseline, thanks to its carefully designed loss function that handles occlusions and moving objects. Further refinements like ManyDepth [14] improved consistency by leveraging information from multiple frames during inference.

In the supervised domain, architectures like AdaBins [15] have achieved exceptional performance. Its main innovation is a novel method to discretize the depth range into adaptive bins, allowing the network to capture both the global scene layout and fine-grained local details with high fidelity.

A key challenge in the field is generalization to unseen environments and recovering true metric scale. Architectures like MiDaS [16] decisively addressed this by training on a mixture of diverse public datasets, achieving unprecedented zero-shot generalization capabilities. More recently, models such as AnyDepth [8] have pushed the state-of-the-art in generalization even further.

However, the success of all these state-of-the-art methods fundamentally relies on the photometric consistency inherent in RGB images, a property that is absent in data from thermal cameras, thus motivating our work.

3 PROPOSED FRAMEWORK

Our proposed framework is designed to compute a dense depth map from a single thermal image by transforming it into a robust, temperature-invariant structural representation. The methodology consists of a multi-stage process, including image pre-processing, line segment extraction, and entropy-based filtering, before feeding the result into a deep neural network. The overall pipeline is depicted in Figure 1.

3.1 Thermal Image Preprocessing and Normalization

The process begins with the raw thermal image in 16-bit format, T. To reduce sensor noise while preserving significant edges, a **bilateral filter** is applied. The value of a pixel $\mathbf{x}$ is updated according to:

$$T_f(\mathbf{x}) = \frac{1}{W_p} \sum_{\mathbf{y}\in\Omega(\mathbf{x})} T(\mathbf{y}) G_{\sigma_s}(\|\mathbf{x}-\mathbf{y}\|) G_{\sigma_r}(|T(\mathbf{x})-T(\mathbf{y})|) \tag{1}$$

where T_f is the filtered thermal image, $\Omega(\mathbf{x})$ is a neighborhood around $\mathbf{x}$, G_{σ_s} and G_{σ_r} are Gaussian kernels for the spatial and range (intensity) domains, respectively, and W_p is a normalization factor. Subsequently, to handle outliers, pixel values are clipped and then rescaled using min-max normalization into a 8-bit single-channel intensity map, T_{cn}.

3.2 Line Segment Extraction

The core of our framework is the conversion of the thermal image into a structural representation that is robust to thermal variations. For this, we extract line segments using the **Fast Line Detector (FLD)** algorithm [17]. However, the quality of the detected lines is highly dependent on the algorithm's internal parameters θ. For this work, we select a configuration of these parameters θ based on previous experiments, to obtain the line-segment representation of thermal image $L_{T(\theta)}$

This simplified approach was deliberately chosen to clearly demonstrate the viability of our core hypothesis without introducing the complexity of an advanced optimization routine. We note, however, that this selection process could be further enhanced by employing more sophisticated techniques, such as swarm intelligence algorithms [18, 19], which remains a promising direction for future work.

3.3 Entropy-Based Filtering for Noise Suppression

To suppress noisy line segments in flat regions, we use a filtering step based on local entropy. First, a **local entropy map**, E_Ω, is computed from T_f. For each pixel $\mathbf{x}$, its local entropy is calculated over a neighborhood window $\Omega_k(\mathbf{x})$ of size $k \times k$:

$$E_\Omega(\mathbf{x}) = -\sum_{j=0}^{255} p_j(\mathbf{x}) \log_2(p_j(\mathbf{x})) \tag{2}$$

where $p_j(\mathbf{x})$ is the probability of intensity level j within the window $\Omega_k(\mathbf{x})$. This map highlights true edges with high entropy values. This map is then thresholded to create a binary mask, M:

$$M_E(\mathbf{x}) = \begin{cases} 1 & \text{if } E_\Omega(\mathbf{x}) > \tau_e \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where τ_e is a predefined entropy threshold. The final, clean line segment image, L_f, is obtained by an element-wise product:

$$L_f = L_{T(\theta)} \odot M_E \tag{4}$$

where $\odot$ denotes the Hadamard product [20, 21].

3.4 Depth Estimation Network

The final, filtered line segment image, L_f, serves as the input to the depth estimation network. For this work, we only use the inference power of **AnyDepth-V2** to obtain the relative dense maps $\hat{D}$. The entire process can be seen as a function composition where the final depth map $\hat{D}$ is produced by the network $\mathcal{N}$ from the output of our pre-processing framework $\mathcal{F}$:

$$\hat{D} = \mathcal{N}(\mathcal{F}(T)) \quad \text{where} \quad \mathcal{F}(T) = L_f \tag{5}$$

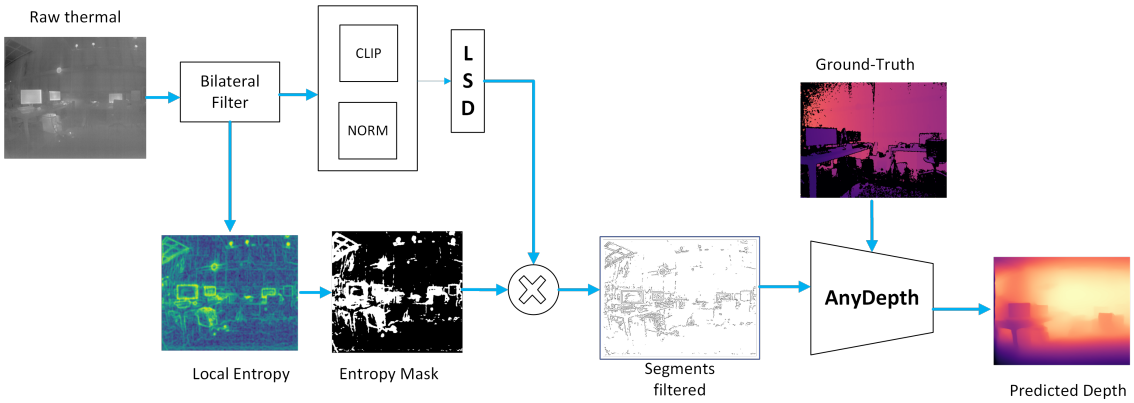


Figure 1: Pipeline of depth estimation from line-segment thermal representation.

4 Experimental and Results

This section presents the empirical validation of our framework. The main goal is to evaluate whether our proposed structural representation, derived from a thermal image, can serve as an effective input for a depth estimation network in the thermal domain, and to compare its performance against a standard RGB image input.

4.1 Experimental Setup

4.1.1 Dataset

For our experiments, we utilize the **OdomBeyondVision** dataset [22]. This indoor dataset provides sequences of RGB, depth and thermal images, captured from three platforms: Unmanned Ground Vehicle (UGV), Unmanned Aerial Vehicle (UAV) and Han-held. In this work, we use one UAV sequence, see details in Table 1. For the different frequency rates of devices, the number of images are different for the same sequence, so we made a synchronization of images using their time-stamps.

Sequence	Thermal	RGB	Depth
2022-01-20-00-14-36	18515	3176	6353

Table 1: UAV sequence details from OdomBeyondVision.

4.1.2 Implementation details

The pipeline was implemented in Python. The thermal images were first processed using a bilateral filter to reduce noise while preserving edges. We used the OpenCV library's implementation with a neighborhood diameter of `d=5`, and standard deviations of `sigmaSpace` (σ_s) = 100 and `sigmaColor` (σ_r) = 100.

Following this, the local entropy map was computed using the `entropy` function from Scikit-image, with a circular structuring element of radius 5 (`disk(5)`). An entropy mask was then generated by thresholding this map to filter out noisy line segments in low-texture regions. The threshold τ_e was selected automatically for every image using Otsu method [23].

The line segments were extracted using OpenCV's Fast Line Detector (FLD) implementation. The key parameters for the FLD were set as follows: `length_threshold=1`, `distance_threshold=0.1`, `canny_th1=70`, and `canny_th2=100`. Merging of lines was disabled to preserve raw structural details.

For the final depth estimation stage, we use the public implementation of the pre-trained **AnyDepth-V2 (Large)** model [8] in a zero-shot inference mode. Specifically, the model is configured with an output feature embedding size of $\mathbf{256}$ and utilizes multi-scale feature channels of $[\mathbf{256}, \mathbf{512}, \mathbf{1024}, \mathbf{1024}]$ in its decoder, as defined by the public repository. For inference, the structural input image $\mathcal{L}_f$ was resized to the model's default input resolution of $\mathbf{640} \times \mathbf{480}$ pixels. No fine-tuning or re-training was performed, as the purpose of the framework is to validate the effectiveness of the structural input alone.

4.1.3 Comparison Conditions:

To evaluate our method, we established two experimental conditions:

1. **Baseline (RGB):** The AnyDepth-V2 network receives the original RGB image from the dataset as input. This represents the expected performance of a standard vision system.

2. **Our Framework (Thermal-Structural):** The AnyDepth-V2 network receives the line segment representation generated from the corresponding thermal image, following the methodology described in Section 3.

This direct comparison allows us to evaluate the viability of a structural representation derived from a thermal sensor for the task of depth estimation when only thermal data is available.

To quantitatively evaluate the performance of both conditions, we employ the six standard metrics for monocular depth estimation proposed in [11].

4.2 Results

The results of our evaluation on the test set are summarized in Table 2. The table compares the performance of our thermal-structural framework against the RGB baseline. As expected, the baseline using direct RGB input outperforms our framework across all evaluated metrics [11]. This is attributable to the fact that the AnyDepth-V2 network was trained extensively on images from the visible spectrum.

However, the most significant finding is not the performance gap, but that our thermal-structural framework produces a coherent and quantitatively reasonable depth map without any re-training or fine-tuning. The ability of a network trained exclusively on RGB data to interpret such a radically different input modality—a binary line segment image derived from a thermal sensor—and still infer the 3D structure of the scene, validates our core hypothesis: the proposed structural representation is a viable and meaningful input for depth estimation. This demonstrates the framework's potential to decouple geometric perception from thermal information.

A visual comparison of the results is presented in Figure 2. We show: (a) the original RGB image, (b) the thermal image, (c) our line segment representation, (d) the depth map predicted from the RGB input, (e) the depth map predicted by our framework, and (f) the ground truth.

5 DISCUSSION

The quantitative results presented in Table 2 show that the baseline, using direct RGB input, outperforms our framework. This outcome is expected, given that the AnyDepth-V2 network was pre-trained extensively on images from the visible spectrum. However, the most significant finding of this work is not the performance gap, but the demonstrated viability of our approach. The ability of a network trained exclusively on RGB data to interpret a radically different input modality—a binary line segment image derived from a thermal sensor—and still infer a coherent 3D structure validates our core hypothesis. This confirms that the proposed structural representation is a meaningful input for depth estimation, showcasing the potential to decouple geometric perception from volatile thermal information.

The structural representation is intentionally dense and subject to extreme parameter settings in the FLD algorithm to maximize the capture of geometric details. This configuration is a deliberate design choice: the exceptionally low *distance threshold* parameter (0.1) forces the detected elements to adhere to a strict model of rectilinearity. This crucial geometric constraint filters out the curvature and irregularity inherent in simple gradient-based thermal edges, ensuring that the input L_f is composed of stable structural information.

This framework opens possibilities for robust UAV navigation using a single, lightweight sensor in environments where RGB cameras would fail, such as darkness or smoke-filled rooms. For platforms where size, weight, and power (SWaP) are critical constraints, leveraging an existing thermal camera for both high-level tasks like human detection and low-level navigation is a significant advantage.

It is important to acknowledge the limitations of this study. A portion of the performance gap can be attributed to the fact that the RGB and thermal images in the OdomBeyondVision dataset are not perfectly spatially aligned, which introduces a systematic error in the evaluation. Furthermore, this validation was conducted in a dataset with adequate lighting conditions; the performance in the target scenario of complete darkness has not yet been empirically tested, nor has the generalization to unstructured outdoor aerial environments, which present challenges of greater range and variable scale.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented a novel framework for monocular depth estimation on UAVs using a single thermal camera. Our core contribution is the transformation of thermal images into a temperature-invariant structural representation (line segments), addressing the lack of photometric consistency in thermal data. We successfully demonstrated that this structural representation serves as a viable direct input for the pre-trained AnyDepth-V2 network, validating the hypothesis that underlying geometric structure is a meaningful signal for depth perception. This opens a path for robust, lightweight, single-sensor navigation in challenging scenarios.

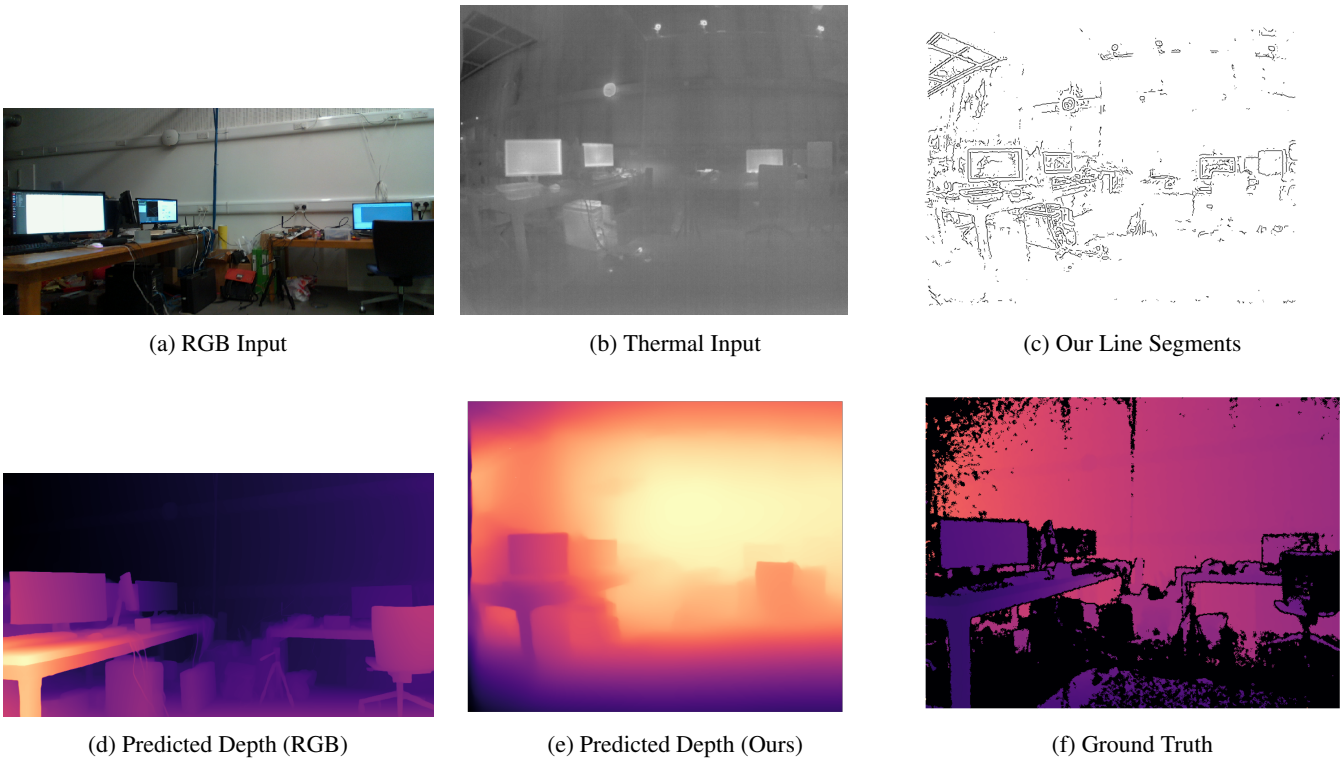
For future work, we identify three primary directions. First, the line segment extraction process can be enhanced; while the classic LSD algorithm is effective, we plan to explore modern deep learning-based methods for line detection, which could provide a more robust structural representation. Second, the most critical next step is to fine-tune or fully retrain a depth estimation network, such as AnyDepth, on our line segment representation. We hypothesize that this will significantly close the performance gap with the RGB baseline by allowing the network to specialize in this new modality. Finally, we will extend our validation to include datasets with challenging low-light and zero-light conditions to empirically demonstrate the framework's effectiveness in its primary target scenarios, such as search and rescue missions. This will include testing the framework's generalization capability to outdoor aerial environments.

ACKNOWLEDGEMENTS

The authors thankfully acknowledge computer resources, technical advice and support provided by Laboratorio Nacional de Supercómputo del Sureste de México (LNS), a

Method	Input Modality	REL ↓	RMSE ↓	log_{10} ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Baseline	RGB	0.3140	1.2442	0.3800	0.4955	0.7759	0.9015
Our Framework	**Lines (Thermal)**	0.4136	1.5156	0.4398	0.3981	0.6906	0.8635

Table 2: Quantitative comparison on the OdomBeyondVision dataset. Our framework uses a line segment representation from a thermal image as input to the AnyDepth-V2 network, while the baseline uses the corresponding RGB image.



(a) RGB Input (b) Thermal Input (c) Our Line Segments

(d) Predicted Depth (RGB) (e) Predicted Depth (Ours) (f) Ground Truth

Figure 2: Qualitative results. For a representative scene, we compare the depth map produced by the baseline using an RGB image against the depth map from our framework, which uses a structural representation derived from a thermal image. Our method successfully captures the main geometric structure of the scene.

member of the CONAHCYT national laboratories, with project No. 202402046.

References

[1] Mingyang Lyu, Yibo Zhao, Chao Huang, and Hailong Huang. Unmanned Aerial Vehicles for Search and Rescue: A Survey. *Remote Sensing*, 15(13):3266, June 2023.

[2] Carlos Osorio Quero and Jose Martinez-Carranza. Unmanned aerial systems in search and rescue: A global perspective on current challenges and future applications. *International Journal of Disaster Risk Reduction*, 118:105199, February 2025.

[3] Zhiwei Huang, Mohammed A.H. Ali, Yusoff Nukman, Hai Lu Xu, Shikai Zhang, Hui Chen, and Mohammad Alkhedher. A systematic review of monocular depth estimation for autonomous driving: Methods and dataset benchmarking. *Results in Engineering*, 26:105359, June 2025.

[4] Jiuling Zhang. Survey on Monocular Metric Depth Estimation, 2025.

[5] Muhamad Risqi U. Saputra, Pedro P. B. De Gusmao, Chris Xiaoxuan Lu, Yasin Almalioglu, Stefano Rosa, Changhao Chen, Johan Wahlstrom, Wei Wang, Andrew Markham, and Niki Trigoni. DeepTIO: A Deep Thermal-Inertial Odometry With Visual Hallucination. *IEEE Robotics and Automation Letters*, 5(2):1672–1679, April 2020.

[6] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep Depth Estimation from Thermal Image. In *2023 IEEE/CVF Conference on Computer Vision and Pat-*

tern Recognition (CVPR), pages 1043–1053, Vancouver, BC, Canada, June 2023. IEEE.

[7] Martin Brenner, Napoleon H. Reyes, Teo Susnjak, and Andre L. C. Barczak. RGB-D and Thermal Sensor Fusion: A Systematic Literature Review. *IEEE Access*, 11:82410–82442, 2023.

[8] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2, 2024.

[9] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, Seattle, WA, USA, June 2024. IEEE.

[10] Peize Li, Kaiwen Cai, Muhamad Risqi U. Saputra, Zhuangzhuang Dai, and Chris Xiaoxuan Lu. OdomBeyondVision: An Indoor Multi-modal Multi-platform Odometry Dataset Beyond the Visible Spectrum. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3845–3850, Kyoto, Japan, October 2022. IEEE.

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2366–2374, Cambridge, MA, USA, 2014. MIT Press. event-place: Montreal, Canada.

[12] Ravi Garg, Vijay Kumar B.G., Gustavo Carneiro, and Ian Reid. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, volume 9912, pages 740–756. Springer International Publishing, Cham, 2016.

[13] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3827–3837, Seoul, Korea (South), October 2019. IEEE.

[14] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, Nashville, TN, USA, June 2021. IEEE.

[15] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth Estimation Using Adaptive Bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4008–4017, Nashville, TN, USA, June 2021. IEEE.

[16] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12159–12168, Montreal, QC, Canada, October 2021. IEEE.

[17] Rafael Grompone Von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall. LSD: a Line Segment Detector. *Image Processing On Line*, 2:35–55, March 2012.

[18] Tareq M. Shami, Ayman A. El-Saleh, Mohammed Alswaitti, Qasem Al-Tashi, Mhd Amen Summakieh, and Seyedali Mirjalili. Particle Swarm Optimization: A Comprehensive Survey. *IEEE Access*, 10:10031–10061, 2022.

[19] Thounaojam Chinglemba, Soujanyo Biswas, Debashish Malakar, Vivek Meena, Debojyoti Sarkar, and Anupam Biswas. Introductory Review of Swarm Intelligence Techniques. In Anupam Biswas, Can B. Kalayci, and Seyedali Mirjalili, editors, *Advances in Swarm Intelligence*, volume 1054, pages 15–35. Springer International Publishing, Cham, 2023.

[20] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, New York, NY, second edition, corrected reprint edition, 2017.

[21] Grigorios G. Chrysos, Yongtao Wu, Razvan Pascanu, Philip H.S. Torr, and Volkan Cevher. Hadamard Product in Deep Learning: Introduction, Advances and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6531–6549, August 2025.

[22] Peize Li, Kaiwen Cai, Muhamad Risqi U. Saputra, Zhuangzhuang Dai, Chris Xiaoxuan Lu, Andrew Markham, and Niki Trigoni. OdomBeyondVision: An Indoor Multi-modal Multi-platform Odometry Dataset Beyond the Visible Spectrum, 2022.

[23] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979.