Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 33, Num. 2 (2026) 356-363

RESEARCH ARTICLE

Evaluating Monocular Depth Estimation on Embedded Platforms for Autonomous Navigation

Avaliação da Estimativa de Profundidade Monocular em Plataformas Embarcadas para Navegação Autônoma

Wilbur Naike Chiuyari Veramendi[1]*, Fernando Santos Osório[1]

Abstract: Deep learning-based monocular depth estimation has achieved significant advancements on urban benchmarks, but its embedded application remains limited by efficiency constraints. Vision Transformers (ViTs) and Foundation Models (FMs) show promising zero-shot generalization capabilities, yet their adaptation to resource-constrained hardware requires careful study. In this work, we investigate the development of the DepthAnything model on an NVIDIA Jetson Orin, analyzing the trade-off between accuracy and inference speed for different backbones (ViT-S, ViT-B, and ViT-L). We report quantitative metrics including *AbsRel*, δ_1, RMSE, and FPS on the KITTI dataset, along with qualitative results. Our experiments show that the ViT-S backbone offers the best balance of accuracy and real-time performance (44 FPS), whereas ViT-B suffers from degradation and ViT-L exhibits significant instability due to optimization artifacts. These findings highlight the viability of compact backbones for embedded visual perception and suggest future optimizations, such as quantization-aware training and pruning, in larger architectures.
Keywords: depth estimation — autonomous navigation — visual perception — self-driving car

Resumo: A estimativa de profundidade monocular baseada em aprendizado profundo tem obtido avanços expressivos em benchmarks urbanos, mas sua aplicação embarcada ainda é limitada por restrições de eficiência. Vision Transformers (ViTs) e modelos fundamentais (FMs) apresentam potencial de generalização zero-shot, porém sua adaptação a hardware restrito requer estudo cuidadoso. Neste trabalho, investigamos a execução do modelo DepthAnything em uma NVIDIA Jetson Orin, analisando o compromisso entre precisão e velocidade de inferência para diferentes backbones (ViT-S, ViT-B e ViT-L). As métricas *AbsRel*, δ_1, RMSE e FPS são reportadas sobre o dataset KITTI, juntamente com resultados qualitativos. Os experimentos mostram que o backbone ViT-S oferece o melhor equilíbrio entre acurácia e tempo real (44 FPS), enquanto ViT-B sofre degradação e ViT-L apresenta forte instabilidade devido a artefatos de otimização. Esses resultados reforçam a viabilidade de backbones compactos para percepção visual embarcada para futuras otimizações, como quantization-aware training e pruning, para arquiteturas maiores.
Palavras-Chave: estimativa de profundidade — navegação autônoma — percepção visual — veículo autônomo

[1] *Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP), Brazil*
*__Corresponding author__: wilburncv@usp.br
DOI: http://dx.doi.org/10.22456/2175-2745.150996 • **Received:** 16/10/2025 • **Accepted:** 16/10/2025

1. Introduction

Depth estimation in real-world environments requires understanding the 3D geometric structure from one or multiple images [1]. This task is fundamental in computer vision, with applications in autonomous navigation of mobile robots and self-driving vehicles [2]. In this context, several approaches for geometric understanding have emerged in computer vision, as the traditional methods, although useful, are often limited by the reliance on specialized sensors, precise calibration, or controlled acquisition conditions [3].

The advent of *Convolutional Neural Networks* (CNNs) has enabled significant progress, due to their ability to extract local patterns [4]. Through encoder–decoder architectures, CNNs integrate global contextual information via dense or MLP layers, learning to predict geometric representations of the scene and consequently estimating depth either relatively or metrically.

Following this progress, *Monocular Depth Estimation* (MDE) has emerged as a promising alternative, offering the advantages of collecting data in a simple and inexpensive

manner, while covering a wide variety of scenes [cite]. However, CNN-based MDE pipelines, such as those using ResNet backbones [5], tend to overfit when trained for long periods or on large-scale datasets, and their performance degrades as model complexity grows.

A more recent paradigm is represented by *Foundation Models* (FMs) [6]. These models overcome traditional limitations of large-scale supervised training and originated in the field of natural language processing, with pioneering works such as CLIP [7]. Their adoption in computer vision has been particularly successful thanks to transformer-based architectures [8], with *Vision Transformers* (ViTs) [9] standing out as key enablers.

FMs have demonstrated strong zero-shot and few-shot generalization capabilities [10], which are critical for handling unseen domains such as unstructured outdoor environments in autonomous driving [11]. Pioneering works in relative [12, 13] and metric MDE [14] have leveraged joint training on large-scale unlabeled monocular datasets, further enhanced by *self-supervised learning* and *knowledge distillation* [15]. Consequently, this paradigm has attracted increasing interest both in academia and industry, with contributions from Intel [13], Apple [16], TikTok [17], Bosch [18], and Meta [19].

Despite these advances, deploying FMs on resource constrained embedded hardware remains challenging. While specialized CPU/GPU-based embedded platforms have emerged [20], practical MDE implementations must balance accuracy and efficiency. CNN-based methods such as FastDepth [21] have achieved high inference speeds in FPS, while transformer-based variants demonstrate robustness due to their self-attention mechanisms [22]. However, studies focusing on FM-based MDE in embedded devices remain scarce [23], and most of them analyze either accuracy or inference speed in isolation, considering few times the trade-off between both.

To address this gap, this work evaluates the FM DepthAnything [17], which has shown strong zero-shot capabilities for MDE. We analyze its performance when deployed on an NVIDIA Jetson AGX Orin 64GB, reporting both quantitative, qualitative metrics and real-time efficiency.

The remainder of this paper is organized as follows: Section 2 presents the methodology, Section 3 discusses results, and Section 4 concludes the work with suggest for future research.

2. Methodology

The methodology of this work aims to evaluate the applicability of the DepthAnything in autonomous navigation scenarios, with an emphasis on MDE in urban environments. The FM model is based on ViTs and trained from self-supervised visual representations extracted from DINOv2. It is explored in zero-shot mode, that is, without additional adjustments to the evaluation dataset.

Unlike traditional approaches that treat depth estimation independently, the DepthAnything architecture adopts a multi-task strategy, guided by global semantic representations learned via knowledge distillation. This mechanism allows the model to integrate richer contextual information, resulting in greater generalization capacity to unseen domains.

Figure 1 presents the methodology pipeline, highlighting the flow of dataset preparation (KITTI), the deployment process, and depth inference developed on the NVIDIA Jetson AGX Orin 64GB platform.

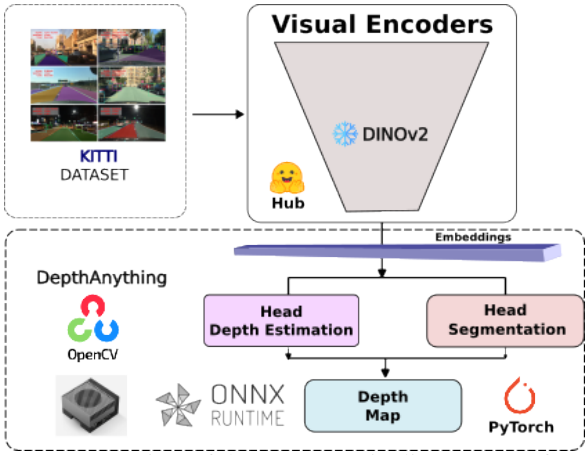

Figure 1. Overview of the DA-JOT pipeline. Images from KITTI are processed through DINOv2 visual encoders, generating embeddings used by DepthAnything heads for depth estimation. Source: Elaborated by author.

2.1 Experimental Configuration

To implement and evaluate the proposed architecture in a realistic mobile robotics and autonomous vehicle scenario, we used the NVIDIA Jetson AGX Orin 64GB platform, widely recognized as a benchmark for high-performance embedded systems. This platform offers up to 275 TOPS of computing power in INT8 mode, 64 GB of RAM, and full support for the CUDA and TensorRT ecosystem, integrated with the JetPack SDK 5.x. The system runs on Linux Ubuntu (20.04), with 512 GB of internal SSD storage.

The pipeline development was based on open source tools, including Torch (2.4.1) and Torchvision (0.19.1), in addition to support for libraries CUDA optimized with transformer operations (PyCUDA 2024.1). To maximize inference performance, we explored the TensorRT engine, which enables mixed precision (FP16), tensor compression, and quantization, and the ONNX Runtime as an intermediary for converting models trained in PyTorch.

2.2 Dataset

The KITTI Depth Prediction Benchmark [24], widely used for depth estimation tasks in urban settings, was used for the experiments. The dataset contains approximately 93,000 depth maps aligned with RGB images and raw LiDAR readings, obtained from the KITTI acquisition platform. The RGB images have a resolution of 1242×375 pixels, encoded at 8 bits per channel. The corresponding depth maps are stored in 16-bit format, representing metric depths in meters, with

values typically ranging from 0.1 m to 80 m, covering both nearby objects and distant structures in the urban scene. In this work, we employed a validation split consisting of 1,000 images with dense depth maps, used to calculate the quantitative metrics (AbsRel and δ_1). This choice ensures direct comparability with previous work in the literature.

As discussed previously, this work employs DepthAnything, a foundation model designed for monocular depth estimation. The model leverages a frozen encoder derived from DINOv2, which provides robust semantic features (semantic preservation) to guide knowledge transfer across large-scale labeled and unlabeled image collections through a knowledge distillation framework.

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input image processed by the shared visual encoder $E(\cdot)$, pre-trained on large-scale datasets with full and pseudo-supervision:

$$F = E(I) \tag{1}$$

where $F \in \mathbb{R}^{H'\times W'\times d_c}$ represents the latent feature representation extracted from the image .

The encoder derived from DINOv2, acts as a teacher model T to make predictions on unlabeled image sets $\hat{D}^u$ obtaining pseudo-labeled labels, according to the following Equation:

$$\hat{D}^u = \{(u_i, T(u_i)) \mid u_i \in D^u\}_{i=1}^{N} \tag{2}$$

Using a combination of labeled images (D^l) and unlabeled images (D^u), the authors customized a knowledge distillation model by finetuning the student model S based on the teacher model T. Robustness was addressed through three specialized loss functions: L_l, L_u and L_{feat}, as defined in the following Equations:

$$L_l = \frac{1}{HW}\sum_{i=1}^{HW} \rho\,(d_i^*, d_i) \tag{3}$$

$$L_u = \frac{\sum M}{HW} L_u^M + \frac{\sum (1-M)}{HW} L_u^{1-M} \tag{4}$$

$$L_{feat} = 1 - \frac{1}{HW}\sum_{i=1}^{HW} \cos\left(f_i, f_i'\right) \tag{5}$$

where Equation 3 is an affine-invariant loss based on a function ρ of disparity spaces (d_i^*, di), Equation 4 is a spatial interpolation loss of image pairs u_a and u_b together with a binary mask M, finally Equation 5 is a feature alignment loss from DINOv2 (f') with "semantic-assited" perception.

2.3 Efficient Implementation Flow

The backbone model used was DepthAnything, which combines a frozen encoder derived from DINOv2 with a dense prediction transformer decoder. The PyTorch checkpoint was first exported to the ONNX format (opset 14), ensuring compatibility with transformer-specific operations. The ONNX graph was then converted into a TensorRT engine optimized for FP16 execution. The inference pipeline was developed entirely in TensorRT, with PyCUDA managing GPU memory allocation, asynchronous execution streams, and data transfers.

For evaluation, we employed the KITTI single-image depth prediction benchmark, which provides RGB images paired with dense or semi-dense LiDAR-based ground-truth depth maps. A subset of 1000 images from the validation set was used to compute both accuracy metrics and efficiency. Images were preprocessed to the network input size (308×308 pixels) which is a multiple of 14 to meet the requirements of encoders headers, normalized with mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224. 0.225) , and transferred to GPU memory in FP16 format. Predictions were later resized back to the original resolution to allow pixel-wise comparison with ground truth.

Our methodology provides a reproducible framework for benchmarking FMs of depth estimation in embedded scenarios. By systematically evaluating the accuracy-speed trade-off across three ViT backbones (ViT-S, ViT-B, and ViT-L, with 24.8M, 97.5M, and 335.3M parameters respectively).

2.4 Metrics

The quantitative evaluation of monocular depth estimation requires metrics capable of capturing both geometric accuracy and the robustness of predictions in urban scenarios. In this work, we adopt three widely used metrics in the literature: the Absolute Relative Error (AbsRel – Equation 6), the Root Mean Squared Error (RMSE – Equation 7), and the threshold accuracy (δ_1 – Equation 8). These metrics quantify different aspects of model performance, including proportionality of error, penalization of larger deviations, and relative consistency between predicted and reference depth.

$$\text{AbsRel} = \frac{1}{|N|}\sum_{i\in N}\frac{|d_i - d_i^*|}{d_i} \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{|N|}\sum_{i\in N}||d_i - d_i^*||^2} \tag{7}$$

$$\delta_1 = \max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) < 1.25 \tag{8}$$

where N represents the set of all valid pixels.

These metrics provide a comprehensive view of performance: AbsRel captures the average relative error, RMSE

penalizes larger deviations, and δ_1 evaluates the proportion of estimates within a fixed accuracy threshold. Finally, in the context of computational feasibility, particularly for embedded devices such as the NVIDIA Jetson AGX Orin, performance is also measured in FPS, reflecting the efficiency of the pipeline. Thus, the set of indicators considered combines both geometric accuracy ($\uparrow \delta_1, \downarrow$ AbsRel, $\downarrow$ RMSE).

3. Results and Discussion

Table 1 summarizes the quantitative results obtained on the KITTI monocular depth estimation benchmark, comparing the baseline performance reported by DepthAnything with our embedded deployment, here referred as DA-JOT (DepthAnything on Jetson Orin TensorRT). The metrics reported include AbsRel, RMSE, δ_1 and inference speed measured in FPS.

Table 1. Quantitative evaluation on KITTI benchmark comparing baseline DepthAnything with our DA-JOT deployment across different backbones.

Model	AbsRel↓	RMSE↓	$\delta_1 \uparrow$	FPS↑
Depth Anything	0.046	1.896	0.982	-
DA-JOT (ViT-S)	0.524	13.768	0.204	44.25
DA-JOT (ViT-B)	2.079	22.139	0.105	40.68
DA-JOT (ViT-L)	11.329	111.767	0.021	24.03

The baseline DepthAnything achieves competitive performance with an AbsRel of 0.046 and $\delta_1 = 0.982$ demonstrating the effectiveness of large-scale pretraining and knowledge distillation. When deployed and developed on the NVIDIA Jetson AGX Orin, the small backbone (ViT-S) achieves a favorable balance, maintaining real-time inference at 44.25 FPS while only slightly degrading in accuracy (AbsRel 0.524). This trade-off highlights the suitability of lightweight ViT-based backbones for embedded applications where low latency is critical.

Conversely, larger backbones (ViT-B and ViT-L) exhibit a marked degradation in depth accuracy (AbsRel > 2.0 and AbsRel > 13.0, respectively) despite maintaining reasonable throughput (40.68 and 24.03 FPS). This behavior suggests that the increased model complexity, when compressed into TensorRT and executed under memory constraints, introduces optimization artifacts that compromise prediction quality.

Despite only a moderate increase in AbsRel, RMSE rises sharply from 1.896 to 13.768, indicating that certain regions (e.g., distant surfaces or textureless areas) suffer from severe depth misestimations. This discrepancy suggests that model compression and TensorRT deployment amplify localized errors, even when average performance remains competitive. The situation worsens for ViT-B and ViT-L, where RMSE values exceed 22 and 111, reflecting optimization artifacts and instability under memory constraint.

The observed degradation correlates strongly with the increase in model size—24.8M, 97.5M, and 335.3M parameters for ViT-S, ViT-B, and ViT-L, respectively—and reflects how backbone complexity directly impacts deployability in integrated inference environments.

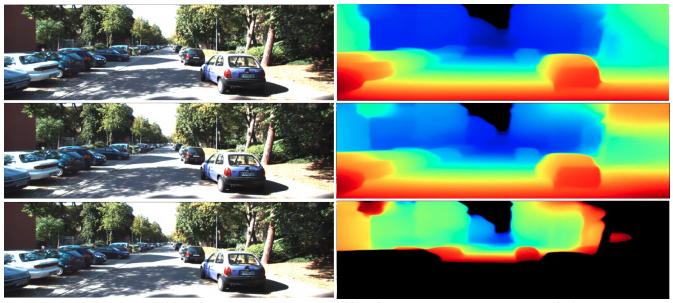

Figure 2. Qualitative depth predictions on KITTI with DA-JOT. ViT-S preserves overall structure (top), ViT-B introduces noise and blurring (middle), while ViT-L collapses with severe artifacts (bottom).

From an architectural perspective, the Vision Transformer backbone scales quadratically with both the token sequence length and feature dimensionality. As model capacity grows, the number of self-attention operations and intermediate tensor activations increases substantially, amplifying numerical sensitivity during FP16 quantization and TensorRT optimization. The ONNX-to-TensorRT conversion revealed systematic clamping of INT64 initializers and FP16 subnormal weights (e.g., 298 and 122 weights affected for ViT-S and ViT-L, respectively). While these truncations introduce negligible errors in smaller networks, their effects compound exponentially in deeper models—particularly within multi-head attention layers, where the scaling factor $1/\sqrt{d_k}$ yields extremely small activation magnitudes. Consequently, precision loss accumulates in ViT-B and ViT-L, destabilizing softmax normalization and linear projection layers, which in turn compromises spatial consistency in depth predictions.

Figure 2 shows the qualitative outcomes on the KITTI dataset and visually supports these quantitative observations. The ViT-S model (top) preserves the global geometric structure of the scene, producing smooth and coherent depth gradients. However, slight oversmoothing occurs near object boundaries, consistent with its moderate receptive field. The ViT-B backbone (middle) introduces local inconsistencies, visible as noisy transitions and blurred mid-range regions—an expected symptom of numerical instability during quantized attention computations. In contrast, the ViT-L backbone (bottom) collapses under embedded constraints, exhibiting severe artifacts, loss of depth continuity, and missing spatial information. These qualitative failures correspond directly to the weight clamping and underflow warnings observed in the TensorRT logs, providing visual confirmation of their cumulative effect on geometric prediction. These results highlight that smaller backbones yield more reliable predictions, while larger ones degrade significantly when deployed on Jetson.

Overall, these results demonstrate that DA-JOT (ViT-S) represents a promising configuration, providing a robust trade-off between accuracy and efficiency. While the raw accuracy

remains below the baseline values obtained in high-compute environments, the real-time inference capability on Jetson Orin validates the potential of deploying foundation models for autonomous navigation and mobile robotics.

4. Conclusion

In this work, we evaluated the deployment of the DepthAnything model on an NVIDIA Jetson AGX Orin (DA-JOT), analyzing the impact of different ViT backbones on MDE. Quantitative and qualitative results revealed that the ViT-S backbone provides the best balance between accuracy and real-time inference speed, whereas ViT-B shows noticeable degradation and ViT-L suffers from severe instability. These outcomes indicate that, under embedded hardware constraints, smaller backbones exhibit greater robustness, challenging the common assumption that increasing model size necessarily leads to better performance.

A deeper analysis of the TensorRT conversion showed that larger models experienced frequent weight clamping and FP16 underflow, especially within multi-head self-attention and normalization layers. This numerical instability propagates through the attention projections, compromising spatial consistency in depth predictions. The results suggest that architectural scaling in transformer-based depth networks must be accompanied by careful quantization and calibration strategies when targeting low-power platforms.

For future work, we plan to explore optimization strategies specifically designed for embedded deployment. In particular, we will investigate quantization-aware training (QAT), selective FP32 retention in normalization and projection layers, and structured pruning to remove redundant attention heads without compromising semantic fidelity. Moreover, post-training calibration techniques such as per-channel scaling and dynamic range equalization may effectively mitigate FP16 underflow and improve numerical stability during TensorRT inference. Beyond architectural optimization, we aim to integrate these models into multimodal perception pipelines and extend evaluation to more diverse urban and outdoor scenarios, advancing toward the operational requirements of real-world mobile robotics and autonomous driving.

Acknowledgements

The authors would like to thank the Institute of Mathematics and Computer Science (ICMC) of the University of São Paulo (USP) and the Mobile Robotics Laboratory (LRM) for the institutional support and infrastructure provided during the development of this work, as well as CAPES and the MOVER/FUNDEP program – SegCom Project for the financial support.

Author Contributions

Wilbur Naike Chiuyari Veramendi: Conceptualization, Methodology, Software (implementation on Jetson Orin), Validation (KITTI benchmarks), Formal analysis, Investigation, and Writing – Original Draft. Fernando Santos Osório: Conceptualization, Resources, Writing – Review & Editing, and Supervision.

References

[1] HORN, B. K. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970.

[2] BRUNO, D. R. et al. Carina project: Visual perception systems applied for autonomous vehicles and advanced driver assistance systems (adas). *IEEE Access*, IEEE, 2023. Disponível em: ⟨https://ieeexplore.ieee.org/abstract/document/10155133⟩.

[3] ZHANG, Z. et al. Review of monocular depth estimation methods. *Journal of Electronic Imaging*, SPIE, v. 34, n. 2, p. 020901, 2025. Disponível em: ⟨https://doi.org/10.1117/1.JEI.34.2.020901⟩.

[4] KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. v. 25. Disponível em: ⟨https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf⟩.

[5] HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

[6] WIGGINS, W. F.; TEJANI, A. S. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, Radiological Society of North America, v. 4, n. 4, p. e220119, 2022. Disponível em: ⟨https://pubs.rsna.org/doi/full/10.1148/ryai.220119⟩.

[7] RADFORD, A. et al. Learning transferable visual models from natural language supervision. In: PMLR. *International conference on machine learning*. 2021. p. 8748–8763. Disponível em: ⟨https://proceedings.mlr.press/v139/radford21a.html⟩.

[8] VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: ⟨https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf⟩.

[9] DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*. [s.n.], 2021. Disponível em: ⟨https://arxiv.org/abs/2010.11929⟩.

[10] RADFORD, A. et al. Learning transferable visual models from natural language supervision. In: MEILA, M.; ZHANG, T. (Ed.). *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021. (Proceedings of Machine Learning Research, v. 139), p. 8748–8763. Disponível em: ⟨https://proceedings.mlr.press/v139/radford21a.html⟩.

[11] ZHOU, X. et al. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, p. 1–20, 2024. Disponível em: ⟨https://ieeexplore.ieee.org/abstract/document/10531702⟩.

[12] RANFTL, R. et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 44, n. 3, p. 1623–1637, 2020. Disponível em: ⟨https://ieeexplore.ieee.org/abstract/document/9178977⟩.

[13] BHAT, S. F. et al. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. Disponível em: ⟨https://arxiv.org/abs/2302.12288⟩.

[14] YIN, W. et al. Metric3d: Towards zero-shot metric 3d prediction from a single image. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. [s.n.], 2023. p. 9043–9053. Disponível em: ⟨https://arxiv.org/abs/2307.10984⟩.

[15] JI, G.; ZHU, Z. Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. In: LAROCHELLE, H. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. v. 33, p. 20823–20833. Disponível em: ⟨https://proceedings.neurips.cc/paper_files/paper/2020/file/ef0d3930a7b6c95bd2b32ed45989c61f-Paper.pdf⟩.

[16] BOCHKOVSKII, A. et al. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. Disponível em: ⟨https://arxiv.org/abs/2410.02073⟩.

[17] YANG, L. et al. Depth anything: Unleashing the power of large-scale unlabeled data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [s.n.], 2024. p. 10371–10381. Disponível em: ⟨https://arxiv.org/abs/2401.10891⟩.

[18] GUO, Y. et al. Depth any camera: Zero-shot metric depth estimation from any camera. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. [s.n.], 2025. p. 26996–27006. Disponível em: ⟨https://arxiv.org/abs/2501.02464⟩.

[19] OQUAB, M. et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Disponível em: ⟨https://arxiv.org/abs/2304.07193⟩.

[20] MAN, A.; ABUAJWA, O. M. S.; AZIZ, N. H. A. A review of techniques for lightweight monocular depth estimation. In: *2024 IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE)*. [s.n.], 2024. p. 351–354. Disponível em: ⟨https://ieeexplore.ieee.org/abstract/document/10877399⟩.

[21] DAO, T.-T.; PHAM, Q.-V.; HWANG, W.-J. Fastmde: A fast cnn architecture for monocular depth estimation at high resolution. *IEEE Access*, v. 10, p. 16111–16122, 2022. Disponível em: ⟨https://ieeexplore.ieee.org/abstract/document/9690863⟩.

[22] LIU, X. et al. Real-time monocular depth estimation merging vision transformers on edge devices for aiot. *IEEE Transactions on Instrumentation and Measurement*, v. 72, p. 1–9, 2023. Disponível em: ⟨https://ieeexplore.ieee.org/abstract/document/10091191⟩.

[23] ŁUCKI, J. et al. Visual perception engine: Fast and flexible multi-head inference for robotic vision tasks. *arXiv preprint arXiv:2508.11584*, 2025. Disponível em: ⟨https://arxiv.org/abs/2508.11584⟩.

[24] UHRIG, J. et al. Sparsity invariant cnns. In: *2017 International Conference on 3D Vision (3DV)*. [s.n.], 2017. p. 11–20. Disponível em: ⟨https://ieeexplore.ieee.org/abstract/document/8374553⟩.