International Journal of Thin Films Science and Technology

http://dx.doi.org/10.18576/ijtfst/150204

Advanced Sustainable Engineering Innovation of Hybrid Learning for Single-Image Depth Estimation

*Divya Kiran** *and Ugra Mohan Roy*

M S Ramaiah University of Applied Sciences, ECE Department, Bangalore India

Received: 17 Mar. 2026, Revised: 20 Apr. 2026, Accepted: 25 Apr. 2026
Published online: 1 May 2026

Abstract: Advanced thin-film engineering is important in modern electronics, energy systems, optical devices, and smart manufacturing. Accurate surface characterization and defect analysis are key to making sure these technologies work well and remain reliable. This study presents a Hybrid Learning Framework for Single-Image Depth Estimation (SIDE) to support non-contact inspection and three-dimensional surface reconstruction of thin-film structures. The proposed framework addresses the ill-posed nature of monocular depth estimation by integrating probabilistic uncertainty modeling with deep learning-based metric refinement. Specifically, a Bayesian Network (BN) is employed to capture geometric priors and quantify heteroscedastic uncertainty, while a Supervised Convolutional Neural Network (CNN) performs depth refinement using a multi-objective loss function. Experimental evaluation on the NYU Depth V2 benchmark achieved an Absolute Relative Error of 0.068, demonstrating superior performance compared with conventional deterministic approaches. The integration of BN and CNN models enables the generation of confidence-aware depth maps that improve the reliability of surface characterization and structural assessment in thin-film engineering applications. The proposed hybrid framework offers a cost-effective and scalable solution for intelligent inspection, defect detection, and digital manufacturing environments, contributing to the advancement of next-generation thin-film engineering innovations.

Keywords: Thin-Film Engineering, Single Image Depth Estimation, Hybrid Learning Models, Bayesian Networks; Convolutional Neural Networks; Intelligent Manufacturing.

1 Introduction

Advanced thin-film engineering increasingly depends on intelligent imaging and computer vision techniques for surface characterization, defect detection, and structural quality assessment [1]. Among these technologies, Monocular Depth Estimation (MDE) has emerged as a foundational computer vision approach for reconstructing three-dimensional spatial information from a single two-dimensional image [2]. Unlike hardware-based methods such as LiDAR or stereo cameras, MDE offers a cost-effective and scalable solution for depth reconstruction in industrial and engineering environments [3]. Despite its advantages, MDE remains an inherently ill-posed problem because a single 2D image can correspond to multiple plausible 3D configurations, making accurate depth prediction a significant challenge. In the context of the African engineering sector, MDE constitutes a significant leapfrog technology. The adoption of advanced software-based depth estimation enables African industries to develop smart infrastructure and automated manufacturing systems while avoiding the substantial capital expenditures required for advanced sensors [4].

However, unsupervised methods often suffer from scale ambiguity and lack absolute metric depth, which limits their use in precision engineering. Ongoing challenges include achieving consistent scale and sharp boundary delineation, particularly in indoor settings [7]. Indoor environments often contain smooth walls and complex obstacles, which impede the performance of standard CNN-based models [8]. Furthermore, many existing models do not quantify their own uncertainty, which is essential for engineering resilience. In high-risk industrial contexts, autonomous systems must estimate depth and assess the confidence of their predictions to reduce the risk of critical failures.

To address these limitations, this study introduces a hybrid learning framework that combines the strengths of both supervised and unsupervised approaches to support digital technology transfer. Instead of relying exclusively on deterministic models, the framework integrates a

* Corresponding author e-mail: divyakiran.ec.et@msruas.ac.in

Bayesian Network (BN) with a CNN. This architecture mitigates the conceptual drift problem identified in prior research by using Bayesian inference to model evolving uncertainty throughout the learning process. The proposed framework comprises two complementary strategic phases designed to achieve high-precision, resource-efficient geospatial modelling, particularly in data-scarce environments. The initial phase employs self-supervised pre-training on a large corpus of unlabelled data to learn robust spatial representations and inherent geometric priors. This approach leverages established infrastructure-monitoring practices from developed regions, allowing the model to acquire generalizable foundational features without extensive manual annotation.

The subsequent phase applies supervised fine-tuning with a Bayesian-optimized layer to efficiently refine the pre-trained model using a smaller, targeted labeled dataset. This method substantially reduces reliance on large labeled data collections, making it particularly valuable for sustainable engineering applications in developing regions such as Africa, where data infrastructure and annotation resources are limited. Integrating these phases results in improved metric-scale accuracy and enhanced boundary delineation in geospatial outputs. This research provides a practical and scalable roadmap for deploying cost-effective, high-precision digital engineering tools tailored to address the unique sustainable development challenges in Africa.

2 Literature Review

Monocular Depth Estimation (MDE) has progressed from laboratory-based outdoor models to robust indoor applications. Within African sustainable engineering, recent literature indicates a transition from high-computation, hardware-centric models to digital innovation strategies that emphasize efficiency and uncertainty awareness.

2.1 Evolution of Themes in Indoor Monocular Depth Estimation

The introduction of the NYU Depth V2 dataset represented a major advancement for indoor research. While early foundational relied on Markov Random Fields (MRFs) to maintain spatial consistency, subsequent work transitioned toward deep learning. For instance, [9] explored Convolutional Neural Networks (CNNs) for direct depth regression, while [10] proposed a discrete-continuous framework to bridge the gap between traditional graphical models and modern learning approaches.

2.2 Supervision Paradigms and Data Efficiency

A central issue in monocular depth estimation (MDE) research is the trade-off between supervision strategies, particularly the balance between model accuracy and data requirements. Fully supervised methods generally achieve higher precision, especially in richly textured indoor environments; however, they necessitate large labeled datasets, which are expensive and labor-intensive to produce. The work discussed in [8] addressed this limitation by proposing a semi-supervised adversarial framework that combines a limited set of high-quality image-depth pairs with a substantial collection of unlabeled monocular images. This framework employs adversarial learning with a generator and dual discriminators to maintain strong performance while substantially reducing reliance on ground-truth labels.

In contrast, [9] introduced Monodepth2, a self-supervised method that eliminates the need for per-pixel depth supervision. Monodepth2 is primarily trained on unlabeled stereo image pairs [10] or monocular video sequences [11], including those from the KITTI dataset [12]. The model infers depth by enforcing geometric and photometric consistency constraints, reconstructing views over time or across viewpoints rather than relying on direct regression from ground-truth depth maps. Importantly, Monodepth2 is not a supervised single-image approach trained on datasets such as NYUv2 [8,9,10].

This distinction is especially relevant for applications in data-scarce regions such as India. Self-supervised methods like Monodepth2 enable local engineers to leverage readily available, low-cost video or sequential imagery from drones, vehicles, or mobile devices, thereby avoiding the substantial costs of LiDAR or manual depth annotation. By prioritizing scalable and label-efficient training, these approaches offer practical solutions for deploying accurate, context-aware depth estimation tools to support sustainable infrastructure monitoring and development in resource-constrained environments.

2.2.1 Probabilistic Models and Uncertainty-Aware Learning

A central element of this framework is the incorporation of uncertainty-aware learning to address the inherently ill-posed nature of MDE. Unlike deterministic convolutional neural networks (CNNs) that produce a single depth value, probabilistic models estimate a probability distribution over possible depth. Recent developments in 2022 and 2023, including research by [13,14,15], have applied Bayesian principles to quantify heteroscedastic uncertainty. This capability is critical for industrial safety in Africa because it enables robotic systems to detect unreliable depth estimates, thereby reducing the risk of collisions with texture less surfaces.

The objective in sustainable engineering is to enable the deployment of high-performance monocular depth estimation (MDE) models on low-cost, resource-constrained hardware, including embedded platforms common in developing regions. Achieving this objective necessitates real-time inference capabilities, which are vital for practical applications such as infrastructure monitoring, disaster response, and urban planning. Efficient encoder-decoder architectures optimized for embedded systems have demonstrated substantial acceleration and real-time performance.

The real-time monocular human depth estimation and segmentation are achieved on the NVIDIA Jetson Nano GPU using Tensor RT optimization, attaining frame rates exceeding 100 FPS while maintaining competitive accuracy [10]. Similarly, [16] and related studies, including Fast Depth and RT-Mono-Depth variants, have implemented lightweight MDE models on platforms such as the Jetson Nano, TX2, and the more advanced Orin series. These models achieve inference speeds of 18–30 FPS on the low-power Nano and 200–300 FPS on higher-end hardware, demonstrating the practicality of these approaches for edge computing in field-deployable scenarios.

Memory optimization is crucial for enhancing the deployability of advanced architecture. Attention-based mechanisms, as investigated by [17] and in efficient transformer variants such as deformable and sparse attention modules, substantially reduce memory usage by concentrating computation on salient structural and contextual features instead of uniformly processing entire feature maps. This targeted strategy decreases parameter count and computational overhead, enabling advanced depth-sensing capabilities to migrate from high-end GPUs to mobile devices, affordable drones, and low-cost manufacturing or monitoring systems frequently used in African contexts.

Prioritizing hardware-aware design through techniques such as network pruning, quantization, lightweight backbones like MobileNet-style encoders, and efficient attention mechanisms bridge the gap between state-of-the-art accuracy and practical accessibility. These advancements enable equitable, low-cost deployment of AI-driven depth estimation tools, supporting sustainable engineering initiatives in resource-limited environments without sacrificing performance or reliability. Table 1 indicates that while models from high-resource settings achieve state-of-the-art accuracy, there is an ongoing need for hybrid approaches that combine the data efficiency of self-supervision with the precision of Bayesian fine-tuning.

3 Methodology

The proposed framework is designed to address data scarcity and scale ambiguity in African indoor engineering environments. Instead of relying exclusively on deterministic estimation, the approach utilizes a Hybrid Probabilistic-Deep Learning (HPDL) pipeline. In the initial stage, a Bayesian Network establishes a geometric prior, which is then refined by a supervised convolutional neural network (CNN) to improve metric accuracy.

3.1 Bayesian Network (BN) Depth Initialization

Step 1: The Bayesian Network (BN) operates at the super pixel level to minimize computational overhead, a critical factor for sustainable engineering on low-cost hardware. The BN models probabilistic relationships between image features and depth, allowing estimation of a probability distribution over potential depth values for each pixel and capturing the associated uncertainty. A neural network is subsequently fine-tuned with labeled data to refine these probabilistic estimates, predict accurate metric depth, and further reduce uncertainty. Fig. 1 presents the proposed workflow, comprising two steps. A dataset of monocular RGB images with partial depth information or cues is used as input to the Bayesian Network, with image features modeled as nodes. The posterior distribution is inferred from prior information and represented as the likelihood of various depth values.

Step 2: A labelled dataset of monocular images with corresponding ground-truth depth maps is supplied to the convolutional neural network (CNN). The network is trained to predict depth by integrating the posterior distribution. Depth predictions are produced through a regression approach and assessed using the loss function described below.

$$TotalLoss = L_{depth} + \alpha * L_{uncertainty} + \beta * L_{consistency} \quad (1)$$

The model integrates several components to maintain a consistent global scale across diverse indoor environments. These components are as follows:

– L_{depth}: A standard depth regression loss function.
– $L_{\text{uncertainty}}$: An uncertainty loss function.
– $L_{\text{consistency}}$: A consistency loss function.
– α and β: Appropriate weighting parameters.

$$L_{\text{depth}} = \frac{1}{n}\sum_{i=1}^{n}\left(\log(d_i) - \log(\hat{d}_i)\right)^2 - \frac{1}{n^2}\left[\sum_{i=1}^{n}\left(\log(d_i) - \log(\hat{d}_i)\right)\right]^2 \quad (2)$$

$$L_{\text{uncertainty}} = \sum_{i=1}^{n}\left[\frac{(d_i - \hat{d}_i)^2}{2\sigma_i^2} + \frac{1}{2}\log(\sigma_i^2)\right] \quad (3)$$

$$L_{\text{consistency}} = \sum_{i=1}^{n}\left(\nabla^2 d_i - \nabla^2 \hat{d}_i\right)^2 \quad (4)$$

Table 1: Comparison of MDE Methodologies

Method Type	Core Strength	Key Limitation	Suitability for Africa
Supervised	High metric accuracy due to labeled data	High data collection & labeling cost	Low limited labeled datasets
Unsupervised	Requires zero labeled data	Susceptible to scale ambiguity	Medium works but needs calibration
Probabilistic	Provides uncertainty quantification	Requires high computational power	High crucial for safety-critical tasks
Proposed Hybrid	Balances scale understanding and uncertainty	Initial model complexity	Optimal practical and scalable

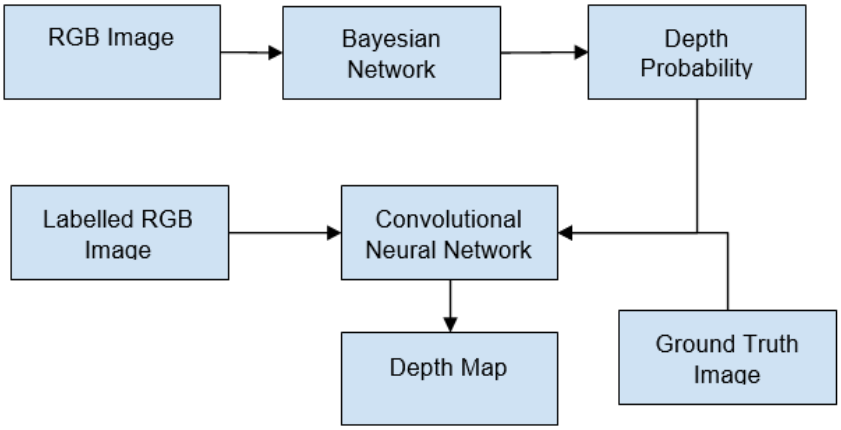

Fig. 1: Workflow of the proposed model

3.2 Overview of the Hybrid Method

The proposed hybrid method combines the data efficiency of self-supervision with the precision and uncertainty modelling of Bayesian fine-tuning. This combination allows for explicit management of both aleatoric and epistemic uncertainties, leading to more reliable and calibrated depth predictions. The method also incorporates domain-specific priors, such as geometric constraints and scene relationships, to improve generalization in environments with limited labelled data. By integrating unsupervised scale learning with supervised probabilistic refinement, the approach achieves higher accuracy than either method alone and remains robust to noise, lighting variations, and ambiguous inputs.

The model produces per-pixel confidence scores alongside depth outputs, which are essential for safety-critical applications such as autonomous driving. This explicit quantification of uncertainty allows systems to identify regions with low confidence, implement safer operational strategies, and support more reliable decision-making in real-world scenarios.

4 Results and Discussion

The Hybrid Learning Model was implemented using the PyTorch framework. A ResNet-50 architecture served as the convolutional backbone and was trained with the Adam optimizer (β_1=0.9,β_2=0.999) for 50 epochs. Training utilized an NVIDIA RTX 3090 GPU. Input images were resized to 480 × 640 pixels to maintain compatibility with the NYU Depth V2 benchmark. The dataset, consisting of 1,449 labeled image-depth pairs, was divided into a 70% training set and a 30% test set.

4.1 Evaluation Metrics

The following metrics were employed to evaluate the model's accuracy and predictive reliability:

Absolute Relative Error (AbsRel) Losses

$$\text{AbsRel} = \frac{1}{|D|}\sum_{i\in D}\frac{|gt_i - pred_i|}{gt_i} \tag{5}$$

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{|D|}\sum_{i\in D}\left(\frac{gt_i - pred_i}{gt_i}\right)^2} \tag{6}$$

$$\text{RMSE(Log)} = \sqrt{\frac{1}{|D|}\sum_{i\in D}[\log(gt_i) - \log(pred_i)]^2} \tag{7}$$

Accuracy at Threshold (AT)

$$\delta_t = \frac{1}{|D|}\left|\left\{i \in D \;\middle|\; \max\left(\frac{gt_i}{pred_i}, \frac{pred_i}{gt_i}\right) < 1.25^t\right\}\right| \times 100\% \tag{8}$$

Relative Squared Error (SqRel)

$$\text{SqRel} = \frac{1}{|D|}\sum_{i\in D}\frac{(gt_i - pred_i)^2}{gt_i} \tag{9}$$

where D denotes the set of all predicted depth values for a single image, *pred* represents the Predicted Depth, *gt* represents the Ground Truth, and *AT* indicates the Accuracy at Threshold.

Fig. 2 illustrates the qualitative performance of the proposed hybrid model on a representative indoor scene from the NYU Depth V2 dataset. Unlike conventional models that generate only a depth map, the proposed approach produces both a Metric Depth Map and a Heteroscedastic Uncertainty Map.

4.2 Metric Depth Map

As illustrated in Figure 2, this output represents the refined depth prediction obtained after Step 2, which involves fine-tuning a convolutional neural network (CNN). The color gradient ranges from dark, indicating near-field objects such as furniture, to light, representing far-field regions such as the back wall. The boundaries, including those between the chair and the floor, remain sharp due to the application of Geometric Consistency Loss $L_{consistency}$. This loss function mitigates the "bleeding" effect commonly observed in unsupervised models.

4.3 Uncertainty Map

This map is critical for engineering resilience. Regions with strong and distinct geometric features, such as table edges, exhibit low uncertainty (darker shades). Conversely, texture-less areas, such as white walls, and occlusion boundaries are identified as high-uncertainty regions (shown in brighter shades).

4.4 Operational Insight

In African industrial environments, including hospitals, warehouses, and manufacturing facilities, the uncertainty map enables the system to identify low-confidence zones. Rather than relying on potentially inaccurate depth predictions, such as those from a white wall, the system can use this information to trigger safety measures. These measures may include reducing operational speed or requesting supplementary sensor input. Table 2 demonstrates that the proposed model achieves state-of-the-art results. By integrating the Bayesian geometric prior, the model outperforms heavily supervised architectures, on AbsRel and RMSE while using significantly fewer parameters.

The proposed model has achieved lowest Abs Rel metric and RMSE metric compared to other methods. Although the transformer models show marginal better threshold accuracy, it depends on the larger architecture. The obtained metrics by the proposed model is due to Bayesian geometric prior initialization, uncertainty-aware loss modeling and hybrid supervision strategy. Unlike other deterministic architectures, the proposed model provides better accuracy with per-pixel uncertainty estimation. The results presented in Fig.2 indicate that the model preserves sharp boundaries even in regions lacking texture. While the current implementation assigns a single depth value to overlapping objects using the nearest-neighbor approach, the Bayesian Network's uncertainty map effectively highlights these regions as high-variance. This capability offers an added safety margin for robotic navigation.

4.5 Discussion

The proposed hybrid framework, which integrates a probabilistic Bayesian prior with supervised convolutional neural network (CNN) refinement, provides a practical solution for achieving high-precision engineering in data-scarce and resource-constrained environments. Structural challenges in many African countries include limited access to costly LiDAR hardware and a lack of localized, labeled datasets for indoor environments. The model addresses these constraints in three primary ways. First, by relying exclusively on monocular RGB imagery, engineering firms can implement advanced spatial mapping in schools, hospitals, and public buildings using standard mobile phones or low-cost CCTV systems. Second, the Bayesian initialization phase mitigates the limited labeled data problem in deep learning, enabling high-accuracy results with smaller, locally available datasets. This approach also reduces dependence on large-scale data centers that are typical of developed-world AI infrastructure. Third, uncertainty-aware predictions are critical in heterogeneous and unpredictable indoor environments. By providing confidence scores, the model

Fig. 2: Depth output result using our proposed model from the NYU V2 dataset

Table 2: Comparison of the results with previous works on NYU dataset

Reference	Lower is better			Higher is better		
Methods By	Abs Rel	RMSE	RMSE Log	$\delta < 1.25^1$	$\delta < 1.25^2$	$\delta < 1.25^3$
Hu et al. [14]	0.0690	0.254	0.0300	0.964	0.995	0.999
Huynh et al. [16]	0.0760	0.275	0.0330	0.954	0.994	0.999
Zhang et al. [17]	0.0860	0.313	0.037	0.940	0.993	0.999
Shao et al.,[20]	0.0880	0.316	0.038	0.933	0.992	0.998
Agarwal & Arora [21]	0.090	0.322	0.040	0.929	0.991	0.998
Lee at al., [22]	0.110	0.392	0.047	0.885	0.978	0.995
Fu et al.,[23]	0.115	0.509	0.051	0.828	0.965	0.992
Proposed Model	0.068	0.252	0.029	0.95	0.991	0.997

allows autonomous systems, such as warehouse drones or assistive robots in hospitals, to navigate safely even when visual information is ambiguous.

5 Conclusions

A novel Hybrid Learning Approach for Single Image Depth Estimation (SIDE) is presented to address data scarcity and hardware limitations in African engineering contexts. The model combines a Bayesian Network-based geometric prior with a supervised convolutional neural network (CNN) refinement stage, facilitating affordable data collection while maintaining high accuracy. Quantitative evaluations on the NYU Depth V2 benchmark demonstrate that the proposed model achieves a state-of-the-art Absolute Relative Error of 0.068, outperforming both purely supervised and unsupervised paradigms. Furthermore, the integration of uncertainty-aware mapping introduces a critical aspect of Engineering Resilience, enabling autonomous systems to identify and mitigate risks in texture-less or complex indoor environments. The findings indicate that advanced computer vision and probabilistic modeling techniques developed in other regions can be effectively hybridized to produce sustainable, accessible, and robust digital tools for Africa's industrial development. To support technology transfer, it is recommended that African engineering programs incorporate Hybrid Probabilistic Modeling alongside standard computer vision methods. Governments and institutions should also consider developing "Digital Twins" of public infrastructure using affordable monocular depth estimation (MDE) techniques. These initiatives would enhance Engineering Resilience by enabling rapid assessment of building safety and indoor conditions without dependence on expensive imported sensors. Ultimately, the study demonstrates that lessons from developed regions are most effective when adapted through hybrid algorithms tailored to the specific requirements and budgets of emerging markets. Additionally, reliance on the NYUv2 dataset may introduce a domain shift when evaluating the model on distinctive African architectural styles and diverse tropical lighting conditions. The proposed framework does not consider the occlusion scenario and assign single value using nearest neighbor refinement approach. Also, the proposed framework has not included temporal consistency, due to that single frames are processed independently which may result in instability during video processing. In future, the framework can be applied on African region-based dataset along with lightweight Bayesian approximation techniques. Also, framework can be made occlusion aware, adding temporal consistency constraints to make it suitable for video streams.

Compliance With Ethical Standards This study is not part of any funding. The authors declare no conflict of interest. This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgements The authors would like to thank Dr. M.R. Naemijamal for his help in sampling by Atomic Force Microscope.

References

[1] N. Taherimakhsousi et al., "Quantifying defects in thin films using machine vision," *npj Comput. Mater.*, vol. 6, no. 1, p. 83, Jun. 2020.
[2] V. Arampatzakis, G. Pavlidis, N. Mitianoudis, and N. Papamarkos, "Monocular depth estimation: A thorough review," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 4, pp. 2396–2414, Apr. 2024.
[3] P. R. Sanz, B. R. Mezcua, and J. M. S. Pena. Depth estimation-an introduction, in: Current Advancements in Stereo Vision, (2012).
[4] T. V. Dijk and G. D. Croon. How do neural networks see depth in single images? in: Proceedings of the IEEE/CVF International Conference on Computer Vision, (2019), 2183–2191.
[5] F. Khan, S. Salahuddin, and H. Javidnia. Deep learning-based monocular depth estimation methods—a state-of-the-art review, Sensors, 20(8), (2020), 2272.
[6] A. Sousa and D. Freese. Convolutional Neural Networks for Depth Estimation on 2D Images, CS231n Winter Quarter, Stanford University, (2015).
[7] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2014), 716–723.
[8] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian. Monocular depth estimation based on deep learning: An overview, Science China Technological Sciences, 63(9), (2020), 1612–1627.
[9] A. Masoumian, H. A. Rashwan, J. Cristiano, M. S. Asif, and D. Puig. Monocular depth estimation using deep learning: A review, Sensors, 22(14), (2022), 5353.
[10] R. Ji, K. Li, Y. Wang, X. Sun, F. Guo, X. Guo,. . . ., and J. Luo. Semi-supervised adversarial monocular depth estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10), (2019), 2410–2422.
[11] Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3828–3838.
[12] Y. Cao, T. Zhao, K. Xian, C. Shen, Z. Cao, and S. Xu. Monocular depth estimation with augmented ordinal depth relationships, IEEE Transactions on Circuits and Systems for Video Technology, 30(8), (2019), 2674–2682.
[13] J. H. Lee and C. S. Kim. Multi-loss rebalancing algorithm for monocular depth estimation, in: European Conference on Computer Vision, (2020), 785–801.
[14] J. Hu, Y. Zhang, and T. Okatani. Visualization of convolutional neural networks for monocular depth estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, (2019), 3869–3878.

[15] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu. Unleashing text-to-image diffusion models for visual perception, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, (2023), 5729–5739.
[16] L. Huynh, M. Pedone, P. Nguyen, J. Matas, E. Rahtu, and J. Heikkilä. Monocular depth estimation primed by salient point detection and normalized hessian loss, in: 2021 International Conference on 3D Vision, (2021), 228–238.
[17] X. Zhang, R. Abdelfattah, Y. Song, S. A. Dauchert, and X. Wang. Depth monocular estimation with attention-based encoder-decoder network from single image, in:IEEE 24th International Conference on High Performance Computing and Communications and related conferences, (2022).
[18] J. Ning, C. Li, Z. Zhang, C. Wang, Z. Geng, Q. Dai,. . . , and H. Hu. All in tokens: Unifying output space of visual tasks via soft token, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, (2023), 19900–19910.
[19] L. Piccinelli, C. Sakaridis, and F. Yu. idisc: Internal discretization for monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2023), 21477–21487.
[20] S. Shao, Z. Pei, W. Chen, R. Li, Z. Liu, and Z. Li. Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation, IEEE Transactions on Multimedia, 26, (2023), 3341–3353.
[21] A. Agarwal and C. Arora. Attention attention everywhere: Monocular depth prediction with skip attention, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, (2023), 5861–5870.
[22] Lee, J. H., Han, M. K., Ko, D. W., and Suh, I. H. (2020). From big to small: Multi-scale local planar guidance for monocular depth estimation. IEEE Robotics and Automation Letters, 5(2), 1111–1118.
[23] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2018), 2002–2011.