

Article

Survey on Monocular Metric Depth Estimation

Jiuling Zhang ^{1,*}, Yurong Wu ² and Huilong Jiang ^{1,3}¹ CRRC Technology Innovation (Beijing) Co., Ltd., 12th Research Institute, Beijing 100083, China² University of Chinese Academy of Sciences, Beijing 101408, China³ CRRC Dalian Co., Ltd., Dalian 116045, China

* Correspondence: zhangjiuling19@mailsucas.edu.cn

Abstract

Monocular metric depth estimation (MMDE) aims to generate depth maps with an absolute metric scale from a single RGB image, which enables accurate spatial understanding, 3D reconstruction, and autonomous navigation. Unlike conventional monocular depth estimation that predicts only relative depth, MMDE maintains geometric consistency across frames and supports reliable integration with visual SLAM, high-precision 3D modeling, and novel view synthesis. This survey provides a comprehensive review of MMDE, tracing its evolution from geometry-based formulations to modern learning-based frameworks. The discussion emphasizes the importance of datasets, distinguishing metric datasets that supply absolute ground-truth depth from relative datasets that facilitate ordinal or normalized depth learning. Representative datasets, including KITTI, NYU-Depth, ApolloScape, and TartanAir, are analyzed with respect to scene composition, sensor modality, and intended application domain. Methodological progress is examined across several dimensions, including model architecture design, domain generalization, structural detail preservation, and the integration of synthetic data that complements real-world captures. Recent advances in patch-based inference, generative modeling, and loss design are compared to reveal their respective advantages and limitations. By summarizing the current landscape and outlining open research challenges, this work establishes a clear reference framework that supports future studies and facilitates the deployment of MMDE in real-world vision systems requiring precise and robust metric depth estimation.



Academic Editor: Aditya Kumar Sahu

Received: 14 October 2025

Revised: 10 November 2025

Accepted: 17 November 2025

Published: 20 November 2025

Citation: Zhang, J.; Wu, Y.; Jiang, H. Survey on Monocular Metric Depth Estimation. *Computers* **2025**, *14*, 502. <https://doi.org/10.3390/computers14110502>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: computer vision; image processing; 3D vision

1. Introduction

Depth estimation aims to recover the three-dimensional structure of a scene from two-dimensional imagery, forming a foundational capability across computer vision applications. Accurate depth perception supports key tasks such as 3D reconstruction [1–3], autonomous navigation [4], self-driving vehicles [5], and video understanding [6]. Beyond traditional perception, depth estimation now drives emerging research in AI-generated content (AIGC), where depth maps guide image synthesis [7,8], video generation [9], and 3D scene creation [10–12]. Recent advances further extend its role to diffusion-based 3D reconstruction [13], dynamic scene generation [14], multimodal 3D reasoning [15], and motion understanding through geometric priors [16]. Depth cues extracted from monocular video have also been employed to model object affordances [17], reinforcing that depth information has become a core representation bridging perception, generation, and reasoning in visual intelligence.

Earlier methods primarily depended on parallax cues or stereo configurations, which required multiple calibrated cameras or active sensors. Such systems, including structured-light and Time-of-Flight setups, were effective under controlled conditions but limited by hardware complexity, environmental sensitivity, and cost. The rise of deep learning transformed this landscape by enabling monocular depth estimation (MDE), which infers scene geometry directly from a single RGB image. This paradigm shift eliminated the need for stereo calibration or auxiliary sensors and made large-scale deployment feasible in applications ranging from robotics to augmented reality. The growing research interest in MDE is exemplified by the recurring Monocular Depth Estimation Challenge (MDEC) at CVPR 2023–2025 (<https://jspenmar.github.io/MDEC/> (accessed on 15 November 2025)), demonstrating both sustained academic relevance and industrial competitiveness.

Within this field, monocular metric depth estimation (MMDE) has emerged as a crucial direction because many real-world systems require depth expressed in absolute metric units rather than relative scales. Metric estimation enables geometrically consistent outputs essential for scale-sensitive domains such as autonomous driving, robotic navigation, AR-based localization, and 3D content generation. Leading research groups and industrial laboratories—including Intel [18], Apple [19], DeepMind [20], TikTok [21,22], and Bosch [23]—have all contributed to advancing MMDE, highlighting the collaborative momentum across academia and industry.

The transition from relative to metric estimation introduces challenges related to domain generalization, structural precision, and resilience to real-world visual variability such as reflective surfaces and dynamic illumination. Discriminative models address these challenges by integrating metric constraints and geometric priors, whereas generative and diffusion-based approaches enhance structural realism and cross-domain robustness. Despite substantial progress, achieving consistent performance across diverse environments remains an open problem. The continued development of large-scale datasets, unified model architectures, and foundation-level priors is expected to further bridge the gap between relative and metric understanding, marking MMDE as a pivotal step toward unified, physically grounded visual perception.

1.1. Research Gap

Existing surveys fail to capture the recent acceleration of progress in MMDE. Earlier reviews [24–27] primarily focused on early MDE frameworks and thus lack coverage of modern architectures capable of predicting metric depth. More recent analyses [28–31] are limited in scope, concentrating on specific environments or relative estimation tasks [32–34]. Consequently, the field lacks a systematic synthesis of MMDE methodologies, datasets, and evaluation strategies that reflect current advances. A comprehensive and up-to-date review is therefore essential for understanding emerging trends and identifying open challenges.

1.2. Main Contribution

This review systematically analyzes the evolution of monocular depth estimation (MDE) toward monocular metric depth estimation (MMDE), focusing on developments presented in top-tier venues such as CVPR, ICCV, and ECCV. Only peer-reviewed publications from these conferences and a limited number of representative works from leading research institutions are included, ensuring technical reliability and reproducibility. Studies without publicly available code or model checkpoints, or those lacking sufficient experimental validation, are excluded to maintain a consistent standard of transparency and scientific rigor. Representative models such as ZoeDepth [18] and Marigold [35] are retained because they exemplify state-of-the-art practices and provide accessible implementations for empirical verification.

The selected literature reflects two dominant and interconnected research trajectories. The first centers on integrating metric constraints into monocular estimation, which addresses the long-standing issue of scale ambiguity inherent in conventional relative depth prediction. These approaches explore supervised and semi-supervised formulations that utilize geometric priors, scale calibration strategies, or sparse ground-truth data to produce depth values in real-world metric units. The second trajectory concerns the application of generative modeling frameworks, particularly diffusion-based and transformer-based architectures, that reconstruct high-fidelity scene geometry while maintaining semantic and structural consistency. These generative approaches demonstrate improved generalization across domains and reduced dependency on labeled data, thereby expanding the practical applicability of MMDE.

By comparatively synthesizing these developments, this review clarifies how modern MMDE frameworks achieve metric consistency, structural preservation, and robustness under varying illumination and scene conditions. The discussion further identifies emerging trends such as unified architectures that jointly learn scene semantics and depth, foundation models that generalize across datasets, and diffusion-based generative priors that enhance structural realism. Collectively, this work provides a coherent understanding of how MMDE research has evolved from conventional MDE methods and highlights the opportunities for future research toward fully generalizable, metric-consistent depth estimation.

1.3. Problem Statement

The task of depth estimation is to compute a dense depth map $D \in \mathbb{R}^{H \times W}$ from a single RGB image $I \in \mathbb{R}^{H \times W \times 3}$, where each pixel's depth value $d_{i,j}$ denotes the physical distance between the camera and the corresponding scene point [36]. This problem is inherently ill-posed because a two-dimensional projection inevitably discards spatial geometry, and monocular setups lack stereo disparity or temporal cues [37]. Accurate prediction therefore depends on learning robust geometric priors from large-scale datasets and leveraging multi-scale contextual features that reflect real-world structure and scale.

1.4. Objective

Depth estimation underpins a wide spectrum of vision applications spanning autonomous driving, robotics, and AR/VR systems [38]. In these scenarios, metric depth enables precise obstacle detection, motion planning, and environment reconstruction. In content creation and computational photography, high-quality depth maps enable multi-focus synthesis, spatially consistent video editing, and physically plausible 3D scene generation. Moreover, depth supervision enhances spatial reasoning in multimodal models by grounding visual tokens in geometric structure [15]. By capturing pixel-level geometry that encodes real-world scale, MMDE acts as a cornerstone for next-generation perception, simulation, and generative intelligence. Continued progress in this field is thus indispensable for advancing the broader goal of enabling machines to perceive, reason, and interact with three-dimensional environments in a physically grounded manner.

2. Background

2.1. Traditional Methods

Before the development of deep learning, depth estimation relied primarily on geometric principles and specialized sensors that implemented explicit physical and mathematical models. These approaches often achieved reliable results in controlled environments. However, the dependence on spatial analysis and additional hardware significantly reduced their flexibility in real-world scenarios.

Sensors: Early depth-sensing systems relied on dedicated hardware to directly acquire spatial information. For example, the original Microsoft Kinect used a structured-light technique: a projector cast a known pattern on the scene and the device inferred depth from the distortion of that pattern [39]. In contrast, Time-of-Flight (ToF) cameras determine distance by measuring the time delay or phase shift between emitted light and its reflection, giving a depth value for each pixel in dense form [40]. Although both modalities offer accurate depth in controlled settings, they face practical limitations: structured-light sensors become sensitive to ambient illumination and surface reflectance [41], while ToF devices suffer from multipath interference, temperature drift and reflectivity-dependent measurement error [42]. These constraints—in particular high cost, calibration complexity and reduced robustness in dynamic or unstructured scenes—have limited their deployment in portable and low-cost depth estimation systems.

Stereo: Stereo vision systems mimic binocular human perception by estimating scene depth from disparity—i.e., the difference between corresponding pixels in two images captured from slightly different viewpoints. The core challenge lies in accurate correspondence matching and calibration: as described in a seminal survey, stereo methods must balance local matching, global regularization and camera geometry to perform well across different conditions [43]. These techniques achieve excellent performance under favorable lighting and texture, but their reliability decreases significantly where texture is weak, illumination is poor or motion is dynamic [44]. In addition, stereo setups require dual-camera hardware and precise alignment, which increases system complexity and cost, thereby reducing practicality for large-scale deployment in constrained-budget or wearable applications [45].

Geometric Multi-Frame: Geometric multi-frame approaches form the classical foundation of three-dimensional reconstruction by exploiting geometric consistency across consecutive frames. These methods, including Structure from Motion (SfM) and Simultaneous Localization and Mapping (SLAM), estimate scene depth through parallax analysis and iterative optimization that jointly refines both camera motion and spatial structure. Indirect techniques identify and match visual features across frames, minimizing reprojection error to recover accurate camera poses and sparse 3D points. Direct techniques, in contrast, optimize photometric error based on pixel intensities, which allows them to retain fine structural details such as texture boundaries and shading transitions [46]. While both strategies achieve high geometric precision, their dependence on consistent illumination and stable textures makes them vulnerable in dynamic or low-contrast environments.

The integration of learning-based models with multi-frame geometry has mitigated many of these limitations by replacing handcrafted feature extraction with data-driven representations that adapt to visual variation. Deep geometric methods have been increasingly designed to couple temporal coherence with differentiable camera motion estimation, enabling dense and temporally consistent depth recovery from monocular videos [47]. The development of unified frameworks such as Vipe [48] further exemplifies this evolution. By jointly modeling camera pose, motion dynamics, and 3D geometry through an end-to-end differentiable pipeline, such systems provide robust geometric perception across varying motion scales and scene complexities. These advances demonstrate that geometric reasoning and neural representation learning are not competing paradigms but complementary mechanisms that together enhance depth estimation reliability.

2.2. Deep Learning

Deep learning has fundamentally shifted depth estimation from a geometry-driven task to a data-driven paradigm, enabling robust single-image inference without reliance on stereo cameras or specialized sensors. By leveraging large-scale datasets, neural networks capture both local textures and global semantic cues, which allows accurate depth

prediction even in ambiguous or low-texture regions. These capabilities extend practical applications to mobile augmented reality, drone navigation, and autonomous driving, where metric depth maps support path planning, obstacle avoidance, and scene understanding [49].

Recent advances in model architecture have further strengthened performance and generalization. Vision Transformer-based networks, particularly DPT, integrate multi-scale feature representations with global attention mechanisms, which enhances the capture of both fine-grained structures and long-range contextual dependencies. DPT models improve structural coherence in depth maps, reduce edge blurring, and provide a strong backbone for subsequent MMDE frameworks [50]. Complementarily, foundation models such as DINO v2 contribute high-quality visual representations that are pre-trained on massive, diverse image collections. When used as a backbone for depth estimation, DINO v2 facilitates zero-shot generalization and cross-domain adaptability, which is crucial for deploying MMDE models in varied real-world environments [51]. Together, these architectural innovations demonstrate that the combination of transformer-based global reasoning and large-scale pretraining substantially advances both accuracy and robustness of deep learning-based depth estimation.

2.3. Monocular Depth Estimation

Monocular depth estimation (MDE) uses deep learning to infer scene depth directly from a single RGB image, which eliminates the need for multi-view inputs or specialized hardware. Compared with traditional multi-frame methods, MDE provides a simpler system design and a lower deployment cost, which makes it highly suitable for real-world applications.

Early studies focused on supervised learning with datasets that contained paired RGB images and ground-truth depth maps. In 2014, Eigen et al. introduced a multi-scale convolutional neural network that predicted both global and local depth simultaneously [52]. The model combined coarse and fine predictions through multi-level feature fusion, which established the foundation for later developments. In 2015, the extended version incorporated surface normals and semantic labels through multi-task learning, which improved prediction accuracy and reduced overfitting [53].

As deep learning advanced, encoder–decoder architectures based on convolutional neural networks (CNNs) became the standard design for MDE. The encoder extracts global contextual features, while the decoder progressively upsamples and refines these features to generate high-resolution depth maps. Multi-scale feature fusion ensures consistency between large-scale structures and fine-grained details, which leads to more accurate depth estimation.

To address the inherent ambiguity of predicting depth from a single image, researchers integrated geometric priors such as perspective constraints and object size regularities. These priors act as guiding signals that encourage networks to produce more physically plausible depth estimates. The combination of geometric knowledge and learned representations enhances generalization, particularly in textureless regions and scenes with complex layouts.

Although early models achieved strong performance on specific benchmarks, they frequently struggled to adapt across domains. Recent approaches overcome this limitation by introducing universal feature extraction modules and domain-invariant learning strategies. These innovations extend the applicability of MDE, enabling robust performance in diverse and unstructured real-world environments.

2.4. Zero-Shot Depth Estimation

Zero-shot depth estimation has become an important strategy for improving generalization across diverse visual domains. Early models that directly regressed metric depth achieved high accuracy when the training and test sets shared similar distributions. However, these models transferred poorly to unseen domains because the predictions depended heavily on absolute scale and camera intrinsics.

To overcome this limitation, researchers reformulated the task as Relative Depth Estimation (RDE), which predicts ordinal relationships between pixels rather than exact metric values. This scale-free formulation removes the need for absolute distance estimation and increases adaptability across heterogeneous datasets. The development of scale-agnostic and scale-and-shift-invariant loss functions further enabled training on data from multiple sources, which significantly improved zero-shot generalization.

A major breakthrough in this direction was the MiDAS framework for zero-shot depth estimation [54]. MiDAS combined multi-source training with scale-invariant objectives to achieve strong cross-domain performance. Its architecture evolved from convolutional neural networks to Vision Transformer-based designs [55], which captured global context and multi-scale features more effectively. Although MiDAS produced only relative depth, its training strategies and architectural innovations provided a solid foundation for subsequent research in zero-shot estimation.

Despite these advances, relative depth estimation introduces a trade-off between generalization and precision. By discarding absolute scale, the task becomes a ranking problem that is robust across domains but insufficient for applications that require metric accuracy. Critical scenarios such as SLAM, augmented reality, and autonomous driving demand depth predictions that are both precise and temporally consistent, which purely relative methods cannot guarantee. The lack of a fixed-scale reference also creates inconsistencies across sequential frames, which reduces stability in dynamic environments.

Zero-shot depth estimation has reshaped the field of monocular depth estimation by directly addressing the challenge of cross-domain generalization. Current research seeks to unify relative and metric estimation by combining scale-invariant learning with mechanisms that preserve metric consistency. Achieving this balance is essential for reliable deployment in tasks such as SLAM, autonomous navigation, and three-dimensional scene reconstruction [56].

3. Materials and Methods

3.1. Monocular Metric Depth Estimation

Monocular metric depth estimation (MMDE) has received renewed attention in the deep learning community because many critical applications such as three-dimensional reconstruction, novel view synthesis, and simultaneous localization and mapping (SLAM) require precise geometric information. Relative depth methods cannot meet this requirement, especially in dynamic scenes where temporal consistency and geometric stability are essential. Advances in model architectures, including Vision Transformers, the scaling of networks to billions of parameters, and the availability of large-scale annotated datasets, have further strengthened the focus on metric depth prediction.

Unlike earlier approaches that were often restricted to narrow domains, modern MMDE seeks to generalize across diverse environments without strict dependence on fixed camera intrinsics or direct depth supervision during training. By predicting absolute depth in physical units, MMDE enables consistent perception across different settings and ensures temporal stability. This robustness in both indoor and outdoor scenarios makes MMDE particularly suitable for real-world deployment.

Early methods frequently relied on known camera intrinsics. Metric3D, for instance, addressed scale and shift variations by mapping images and depth maps into a canonical space and applying focal-length corrections [57]. ZeroDepth introduced a variational inference framework that incorporated camera-specific embeddings, although accurate intrinsic parameters were still required [58]. More recent approaches have reduced this dependency by estimating intrinsics through auxiliary modules or by predicting depth in geometry-aware spherical representations [59].

Another line of research has focused on binning strategies for depth representation. Instead of relying on fixed global distributions, AdaBins introduced dynamic bin allocation that adapts to image content, which improved accuracy in scenes with wide depth variation [36]. LocalBins extended this principle by modeling local depth distributions within spatial regions, which increased precision but also raised computational cost [60]. BinsFormer, which integrates a Transformer backbone, refined bin placement by combining global and local context, thereby producing more consistent predictions [61]. NeW CRFs further advanced this direction by combining deep neural networks with Conditional Random Fields, enforcing pixel-level consistency while explicitly handling uncertainty in depth prediction [62].

A major breakthrough was achieved by ZoeDepth [18]. This framework combined the MiDAS backbone with adaptive metric binning and a lightweight depth adjustment module to deliver accurate metric estimation. ZoeDepth also included an image classification module that dynamically selected the most suitable network head for each scene, which improved robustness across domains. By training on a wide collection of indoor and outdoor datasets, ZoeDepth demonstrated strong zero-shot generalization with minimal fine-tuning. Its unified architecture and multi-source training paradigm established a new benchmark for MMDE and provided a strong foundation for future research.

3.2. Challenges and Improvements

Despite significant progress in MMDE, reliable generalization to unseen scenes remains a critical challenge [63]. Models often lose accuracy and stability when deployed in environments that differ substantially from the training domain. Many single-inference architectures produce blurred geometry, fail to capture fine structural details, and adapt poorly to high-resolution inputs. These limitations reduce robustness and restrict deployment in practical applications.

To overcome these issues, researchers have advanced model architectures, training strategies, and inference designs. Architectural innovations aim to preserve structural integrity and recover fine-scale details. New training strategies incorporate diverse datasets and domain-invariant representations, which improve cross-domain generalization. Enhanced inference techniques reduce artifacts and improve scalability, which increases the practicality of MMDE in complex and dynamic environments.

This section reviews these recent developments and highlights improvements that strengthen geometric consistency, expand generalization capability, and enable reliable deployment in real-world scenarios.

3.2.1. Generalizability

Improving the generalization of MMDE requires two complementary strategies: dataset augmentation and model optimization. Dataset augmentation increases adaptability by exposing models to diverse training domains. Model optimization refines network architectures and training mechanisms, which enhances accuracy and strengthens cross-domain performance.

Dataset Augmentation: Dataset augmentation has become a central strategy for improving generalization and robustness in monocular depth estimation [64–67]. Depth Anything introduces a large-scale semi-supervised self-learning framework that generates more than sixty-two million self-annotated images, which substantially enhances performance across heterogeneous environments [21]. The framework leverages optimized training strategies that enable broad visual representations across multiple domains. In addition, auxiliary supervision incorporates semantic priors from pre-trained encoders, which reduces domain bias and increases robustness. Depth Anything demonstrates strong zero-shot performance in both indoor and outdoor scenarios, yet its effectiveness may still be limited by the quality and diversity of self-annotated labels, which can result in local structural inaccuracies in complex scenes.

Building on this approach, Depth Any Camera (DAC) focuses on adapting models trained on standard perspective images to unconventional imaging modalities such as fisheye and 360-degree cameras [23]. DAC combines Equi-Rectangular Projection (ERP), pitch-aware image-to-ERP conversion, field-of-view alignment, and multi-resolution augmentation. These techniques emphasize geometric consistency and resolution alignment across projections, which reduces errors caused by domain mismatch. Compared with Depth Anything, DAC prioritizes accurate cross-projection depth estimation over large-scale self-supervised diversity, which improves performance in specialized imaging conditions but does not directly address generalization across natural scene variations.

Together, these methods illustrate a progression in dataset augmentation strategies. Depth Anything emphasizes scale and diversity to improve general robustness, whereas DAC emphasizes geometric alignment and modality adaptation to ensure accurate predictions under unconventional imaging conditions. The distinction between approaches arises from their primary focus: one prioritizes self-supervised representation learning, and the other prioritizes geometric consistency. Understanding these differences highlights the complementary nature of augmentation strategies and the need to balance dataset scale, domain diversity, and geometric alignment for robust monocular depth estimation.

Model Improvements: UniDepth introduces a framework that directly predicts metric 3D point clouds without relying on explicit camera intrinsics or metadata [68]. The framework includes a self-promptable camera module that generates dense camera representations and adopts a pseudo-spherical output format, which decouples camera parameters from learned depth features. This design increases robustness to camera variation and improves cross-domain generalization. In addition, a geometric invariance loss stabilizes depth feature learning, while camera bootstrapping and explicit intrinsic calibration ensure accuracy and consistency. By disentangling camera attributes from metric depth prediction, UniDepth establishes a strong foundation for advancing the generalizability of MMDE.

Loss and Training Paradigm Innovation: Recent advances in MMDE have focused on refining training paradigms and loss formulations to enhance structural fidelity, robustness, and generalization. Many approaches address the scarcity of high-quality labels and the presence of noisy supervision by leveraging consistency-based fine-tuning frameworks. For example, unsupervised consistency-regularization strategies exploit augmented views of unlabeled data, which allows models to maintain relational consistency among local image patches and preserve semantic boundaries and fine geometric details. This approach strengthens both cross-domain adaptability and prediction robustness under real-world conditions characterized by incomplete or noisy supervision [69].

Complementing self-supervised consistency, the integration of sparse external signals has been shown to guide metric depth predictions effectively. Incorporating sparse LiDAR inputs as conditional prompts enables pre-trained models to refine predictions adaptively,

which balances the efficiency of purely vision-based approaches with the accuracy of sensor-assisted estimation [70]. This strategy is particularly advantageous in high-resolution settings where structural fidelity and absolute scale consistency are critical. However, the reliance on external sensor signals introduces potential limitations in scenarios where sensor data are unavailable or unreliable, which requires careful trade-offs in deployment.

Advancements in training paradigms have also explored explicit probabilistic modeling to improve metric accuracy and structural sharpness. MoGe employs a multi-stage generative framework that enforces metric consistency through probabilistic reconstruction [71]. By modeling the conditional distribution of depth given an RGB image and progressively refining latent representations, the framework captures global scene context while preserving fine-grained geometric structures. Metric alignment losses further constrain predictions to maintain absolute scale consistency, and adaptive regularization schedules enhance training stability. Unlike purely deterministic regression approaches, this probabilistic paradigm improves generalization across unseen domains and enhances the model's ability to recover sharp and coherent depth structures.

StableDepth complements these directions by integrating uncertainty-aware supervision into training, which addresses the domain gaps arising from synthetic and real data mixtures [72]. Confidence-weighted consistency constraints regulate prediction stability and reduce overfitting caused by noisy labels or domain shifts. This approach enhances robustness across diverse lighting conditions and scene types, demonstrating that uncertainty-guided supervision can effectively complement self-supervised and sensor-conditioned strategies.

Auxiliary Priors & Prompts: MMDE traditionally relies on single-view RGB images, which introduces fundamental challenges such as scale ambiguity, structural blurriness, and limited generalization. Incorporating auxiliary priors and prompts provides external geometric or metric cues that complement image-based predictions and improve both accuracy and robustness [70,73,74]. These strategies differ in the type and timing of the auxiliary information while sharing the goal of enhancing metric reliability and cross-domain adaptability.

One line of research leverages pre-existing metric priors derived from sources such as LiDAR, structure-from-motion reconstructions, or sparse depth sensors. These priors are fused with dense monocular predictions through coarse-to-fine pipelines and distance-aware alignment, which explicitly enforce metric consistency [73]. By conditioning the depth estimation network to integrate both prior and predicted features, this approach improves zero-shot generalization across tasks including depth completion, super-resolution, and inpainting. The main limitation of this strategy lies in the reliance on pre-filled priors, which may not always be available or sufficiently diverse for all deployment scenarios.

Another strategy introduces sparse or low-resolution depth prompts during inference to guide the monocular model toward accurate metric reconstruction [70]. These prompts act as conditioning signals that constrain the latent representations of the model, which enhances sharpness, scale consistency, and robustness in textureless or challenging lighting conditions. Unlike pre-filled priors, prompt-based conditioning provides flexibility by allowing configuration changes at test time, which enables a trade-off between computational efficiency and prediction accuracy. However, the effectiveness of this method depends on the quality and availability of prompt signals, which may vary across devices or environments.

Cross-modal auxiliary signals offer a complementary perspective by incorporating non-visual metric information directly into the learning process. Radar-camera fusion provides explicit range cues that serve as strong metric constraints for monocular RGB estimation [74]. One-stage fusion architectures align radar and image features within a single forward pass, which maintains geometric accuracy while preserving inference

efficiency. This approach demonstrates robust performance under adverse weather, low-light conditions, and partial occlusions, where purely vision-based models typically fail. Compared with pre-filled priors or prompt-based guidance, cross-modal fusion introduces additional sensor complexity but provides complementary structural cues that significantly enhance resilience in challenging scenarios.

3.2.2. Blurriness

Detail loss and edge smoothing remain persistent challenges in dense prediction tasks such as small object detection, depth estimation, and image segmentation. Regression-based depth models often fail to capture accurate geometry and fine-grained features, particularly along object boundaries and in regions with complex textures such as hair or fur. These problems are especially severe at occlusion boundaries and in high-frequency regions, which significantly limits the applicability of depth estimation in real-world environments. Moreover, balancing high-resolution input processing with the preservation of both global consistency and local detail remains difficult. The consequence is frequently blurred edges and structural degradation.

To address these challenges, researchers have proposed new model architectures, training strategies, and data resources. The following subsections outline representative approaches, each of which is designed to improve edge fidelity and structural sharpness while maintaining computational efficiency.

Patching: Patch-based strategies improve depth estimation by combining localized predictions with global scene understanding, which proves particularly effective in visually complex and high-resolution environments. These methods exploit the complementarity between fine-grained local detail and overall scene context, which allows models to capture subtle geometric structures while maintaining structural coherence across the image [19,75,76].

A common approach divides input images into patches for independent depth prediction and subsequently aligns these patches to enforce global consistency. Multi-resolution fusion modules, together with consistency-aware training and inference frameworks, ensure that patch boundaries are smooth and geometric cues are preserved [75]. While this approach improves local accuracy, the multi-stage pipeline increases computational complexity, and small textures can be misinterpreted as depth, which occasionally introduces structural artifacts.

Refinement-oriented strategies reframe high-resolution depth estimation as a post-processing task, where initial coarse predictions are progressively corrected [76]. Loss functions that disentangle detail and scale information sharpen object boundaries while preserving overall depth coherence. Pseudo-labeling and knowledge transfer from synthetic to real datasets further enhance prediction accuracy. Compared with purely patch-wise fusion, refinement methods reduce pipeline complexity, improve inference efficiency, and achieve more reliable global consistency.

Efficiency-focused strategies prioritize fast and practical depth inference while maintaining acceptable accuracy [19]. By leveraging multi-scale Vision Transformers trained on mixed real and synthetic datasets, these methods divide images into minimally overlapping patches, which mitigates context loss and accelerates inference. Unlike earlier patch-based or refinement techniques, such approaches can directly predict absolute depth without requiring camera intrinsics, simplifying deployment. However, focusing on near-object accuracy may reduce depth consistency in distant regions, which highlights the trade-off between efficiency and global structural fidelity.

Synthetic Datasets: Real-world datasets for depth supervision often suffer from label noise, missing values in reflective or transparent regions, inaccurate annotations, and blurred object boundaries. These limitations result from hardware constraints and manual annotation processes, which reduce the reliability of fine structural details during training. Synthetic datasets provide pixel-accurate depth maps generated through rendering engines, which deliver precise supervision even under challenging conditions such as reflections, transparency, and complex lighting [76].

Although synthetic data improves structural fidelity and scene diversity, directly training on synthetic datasets can introduce overfitting and biases due to domain gaps between rendered and real images. Depth Anything V2 [22] addresses these issues by combining synthetic data with pseudo-labeled real images generated by a teacher model, which reduces discrepancies in color distribution and scene composition. The framework also employs a gradient matching loss that sharpens predictions while excluding regions with high loss, which mitigates overfitting to ambiguous or noisy samples. Compared with raw synthetic-only training, this strategy balances the benefits of precise synthetic supervision with the necessity of adaptation to real-world visual complexity, although the diversity achievable through rendering engines still imposes limitations.

StableDepth [72] further enhances synthetic dataset utilization by integrating large-scale synthetic scenes with real-world data within an uncertainty distillation framework. Confidence-weighted consistency constraints regulate prediction stability during training, which prevents synthetic data from introducing overfitting or amplifying biases. Unlike Depth Anything V2, which primarily relies on gradient-guided pseudo-labeling to bridge domain gaps, StableDepth emphasizes uncertainty-aware supervision, which ensures stable and accurate metric depth predictions across diverse domains and lighting conditions. This approach demonstrates that synthetic datasets, when combined with uncertainty-guided learning, can provide both structural precision and robust cross-domain generalization.

Generative Methods: Generative diffusion techniques have introduced effective solutions for reducing edge smoothing and enhancing structural fidelity in depth estimation. By simulating image degradation and iteratively restoring missing information, these models can reconstruct high-frequency structures more accurately than traditional discriminative approaches [35,77–79]. Early work such as Marigold applied diffusion to depth estimation, producing sharper edges and improved structural consistency compared with baseline models [35]. However, its performance declines in scenes with multiple objects or complex spatial layouts, revealing limitations in dense environments.

Enhancements in subsequent frameworks address these limitations by explicitly modeling scene characteristics. GeoWizard, for instance, introduces a decoupler that separates scene distributions during training, which reduces blurring and ambiguity from mixed data [80]. It further incorporates a scene classification vector that improves foreground-background separation and enables more accurate modeling of complex geometries. Joint estimation of surface normals and pseudo-metric depth provides additional structural cues, which enhances 3D reconstruction fidelity compared with single-output diffusion models.

Logarithmic depth parameterization and explicit field-of-view (FOV) conditioning, as employed in Diffusion for Metric Depth (DMD) by DeepMind, further improve cross-domain and scale consistency [20]. By simulating diverse FOVs and providing vertical FOV as input, this approach mitigates depth ambiguity caused by varying camera intrinsics. Single-step denoising with parameterization accelerates inference while preserving structural fidelity, offering a balanced trade-off between runtime efficiency and depth accuracy across indoor and outdoor scenarios.

SharpDepth advances this line of research through a diffusion distillation framework that refines pre-trained MMDE [81]. A teacher–student setup enables the teacher to generate

high-fidelity depth maps via multi-step diffusion, while the student learns a compact representation that retains fine structural details and allows faster inference. The integration of a scale-consistency loss ensures alignment with metric ground truth, which improves absolute depth estimation across diverse environments. Compared with prior generative methods, SharpDepth demonstrates superior preservation of edges, fine structures, and object boundaries, particularly in high-density or complex scenes.

4. Results

4.1. Research Trend

Figure 1 illustrates the development trajectory of monocular depth estimation (MDE) from 2014 to 2025, showing the gradual evolution from narrow-domain regression models to cross-domain generalization and ultimately to MMDE. Early methods, from Eigen et al. (2014) [52] to MiDaS (2020) [82], primarily focused on single-domain or dataset-specific regression. These approaches relied on scale-invariant losses, ordinal regression, and self- or unsupervised image reconstruction objectives, which effectively captured relative depth but lacked metric consistency and generalization capability. Their scope remained constrained to controlled environments, limiting applicability across diverse domains.

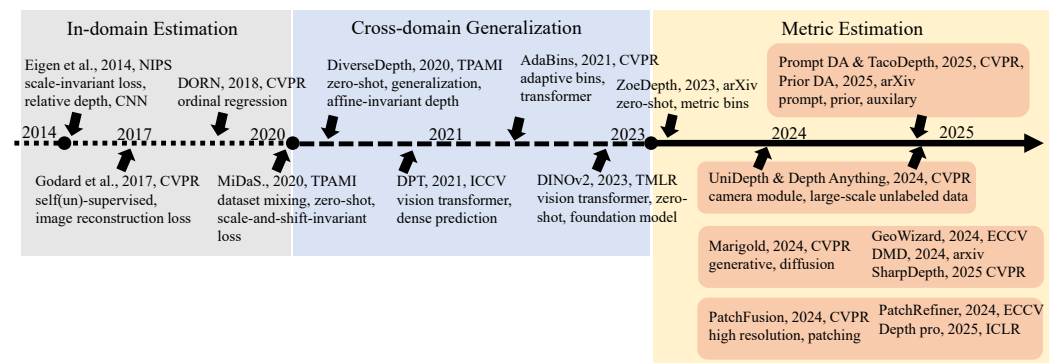


Figure 1. Visualization of key developments in MDE from 2014 to 2025. The timeline is divided into three stages: in-domain relative depth estimation (2014–2020) [52,56], cross-domain generalization (2020–2023) [36,82], and MMDE from 2023 onward [18,20,21,35,68,70,74–76,80,81]. Major milestones such as DPT [50] and DINO v2 [51] are also marked to show their influence on the landscape.

After 2020, the introduction of MiDaS marked a turning point toward cross-dataset generalization. Research began emphasizing scale-and-shift-invariant loss formulations, adaptive depth binning, and the adoption of Vision Transformers (ViTs), which enhanced spatial reasoning and feature abstraction. Building on this foundation, DPT (2021) extended ViT-based architectures to dense prediction tasks, becoming a backbone for numerous subsequent depth and segmentation models. The release of DINO v2 (2023) further reinforced this trend, providing a powerful pretrained foundation model that significantly improved feature transfer and representation quality across visual tasks.

From 2023 onward, with the emergence of ZoeDepth, MDE entered the metric estimation era. ZoeDepth introduced unified frameworks for zero-shot metric prediction across domains, establishing a conceptual bridge to MMDE. By 2024–2025, research converged on developing universal and diffusion-based MMDE models, such as UniDepth and Marigold, that pursue both metric accuracy and generative adaptability. This shift reflects the field's broader transition from data-specific supervision to foundation-level perception, driven by advances in model architecture, loss design, and large-scale pretraining.

4.2. Criteria

Two established evaluation systems are adopted to ensure comparability and transparency across datasets. The first, used in Booster, ETH3D, Middlebury, NuScenes, Sintel, and SUN-RGBD, assesses performance through the proportion of inlier pixels δ_i under thresholds 1.25^i (higher is better) and the scale-invariant logarithmic error (SIlog), defined as $100p \cdot \text{Var}(\varepsilon_{\log})$. These indicators jointly capture relative accuracy and scale consistency, enabling fair evaluation across heterogeneous scenes without relying on absolute scale. The aggregated average rank further summarizes cross-dataset robustness, where lower values denote stronger generalization.

The second system, widely applied to NYU Depth V2 and KITTI, employs the Absolute Relative Error (A.Rel) to quantify metric fidelity, reflecting the mean ratio between predicted and ground-truth depths under true physical scale (lower is better). Given the dense and metric-aligned ground truth in these datasets, A.Rel provides a direct measure of scale-aware precision crucial for real-world perception and navigation tasks. Overall, the δ_i /SIlog scheme focuses on scale-invariant generalization, while A.Rel emphasizes absolute metric reliability—together forming a coherent and complementary framework for evaluating MMDE.

4.3. Comparison

Table 1 provides a qualitative summary of representative MMDE approaches, outlining their key design characteristics and data dependencies. The comparison reveals a clear shift toward sensor-assisted and knowledge-conditioned paradigms, where models incorporate priors, prompts, or cross-modal cues to enhance scale consistency and contextual understanding. Traditional regression-based networks relied solely on monocular cues and image statistics, whereas recent conditioned models leverage pretrained vision-language features or auxiliary sensors to infer depth in complex or ambiguous scenes. This evolution underscores the field's growing emphasis on semantic grounding and context-driven metric reasoning, which extends beyond geometric reconstruction alone.

Table 2 reports quantitative results across both zero-shot and non-zero-shot evaluation settings. For consistency, performance on ETH3D, Middlebury, NuScenes, Sintel, and SUN-RGBD is aligned with the standardized metrics presented in DepthPro (ICLR 2025), while NYU Depth V2 and KITTI results are taken directly from the respective publications of each method. The cross-dataset comparison highlights that models integrating synthetic and real-world data achieve stronger generalization under zero-shot conditions, whereas those trained exclusively on real metric datasets maintain superior scale fidelity in structured environments. Together, these results demonstrate a balance between adaptability and metric precision that reflects distinct architectural priorities across MMDE frameworks.

Figure 2 provides qualitative comparisons showing that patch-based and generative models significantly mitigate the edge blurring and depth smoothing artifacts characteristic of single-pass inference. The visual evidence confirms that, while lightweight architectures favor real-time performance, the incorporation of local refinement or generative priors yields perceptually superior results, particularly in complex or fine-grained regions. Figure 3 visualizes the trade-off between accuracy and computational efficiency by comparing inference time and memory usage for three primary paradigms—single-pass, patch-based, and generative methods—on a 400-frame 1080p video using an RTX 3090 GPU. The logarithmic x-axis illustrates that single-pass models maintain a substantial advantage in runtime and memory efficiency, while patching and generative approaches introduce additional computation to improve structural detail and contextual depth continuity.

Table 1. Recent advances in MMDE are summarized chronologically. Category distinguishes between discriminative models that directly regress depth, generative models that synthesize depth via diffusion or reconstruction processes, and disc + condition methods that extend discriminative models with auxiliary priors, sensors, or prompts for conditional adaptation. Inference indicates whether prediction is performed in a single forward pass or through multiple iterative processes such as diffusion or patch-based reconstruction. Dataset specifies the type of training data, including real, synthetic, or hybrid combinations with sensor-based priors. Output refers to whether the model predicts metric or relative depth, and Source denotes code availability. Recent work demonstrates that while generative diffusion-based approaches improve fine structural fidelity, discriminative and condition-aware frameworks (e.g., StableDepth, Depth Any Camera, Prompting DA) have achieved strong generalization with lower computational cost. The combination of synthetic supervision, sensor priors, and conditional adaptation highlights an emerging trend toward unified, scalable depth estimation frameworks.

Method	Publication	Category	Inference	Dataset	Output	Source
Zoedepth [18]	arXiv	discriminative	single	real	metric	open
Depth Anything [21]	CVPR '24	discriminative	single	real	metric	open
Patch Fusion [75]	CVPR '24	discriminative	multiple (patching)	real	metric	open
Unidepth [68]	CVPR '24	discriminative	single	real	metric	open
Marigold [35]	CVPR '24	generative	multiple (diffusion)	synthetic	relative	open
DMD [20]	arXiv	generative	multiple (diffusion)	real	metric	close
Depth Anything v2 [22]	NeurIPS '24	discriminative	single	real + synthetic	metric	open
GeoWizard [80]	ECCV '24	generative	multiple (diffusion)	real + synthetic	relative	open
Patch Refiner [76]	ECCV '24	discriminative	multiple (patching)	real + synthetic	metric	open
Depth pro [19]	ICLR '25	discriminative	multiple (patching)	real + synthetic	metric	open
Depth Anything AC [69]	arXiv	discriminative	multiple (patching)	real + synthetic	metric	open
MoGe [71]	CVPR '25	discriminative	single	real + synthetic	metric	open
Depth Any Camera (DAC) [23]	CVPR '25	discriminative	single	real	metric	open
SharpDepth [81]	CVPR '25	generative	single	real	metric	open
StableDepth [72]	ICCV '25	discriminative	single	real + synthetic	metric	open
Prompting DA [70]	CVPR '25	disc+condition	single	real + sensor	metric	open
TacoDepth [74]	CVPR '25	disc+condition	single	real + sensor	metric	open
Prior DA [73]	arXiv	disc+condition	single	real + sensor	metric	open

Table 2. ZeroDepth fails to complete evaluations on some datasets due to storage limitations. Metric3D relies on camera parameters, limiting its general applicability. Although Depth Anything provides a flexible framework, its current performance falls short of zero-shot generalization requirements. The evaluation table shows significant performance variation across models and domains, indicating that MMDE still faces major challenges in achieving robust generalization. The six datasets on the left side of the table use higher-is-better metrics ($\uparrow$) to evaluate zero-shot performance, based on test results reported by Depth Pro [19]. In contrast, the two datasets on the right use Absolute Relative Error (AbsRel), where lower values indicate better performance ($\downarrow$), to assess non-zero-shot tasks. While the table offers useful comparisons, the MDE field still lacks a unified benchmarking standard. Inconsistencies in training data, training protocols, model sizes, and inference costs hinder fair and comprehensive evaluation.

	Dataset	Booster $\uparrow$	ETH3D $\uparrow$	Middlebury $\uparrow$	NuScenes $\uparrow$	Sintel $\uparrow$	Sun-RGBD $\uparrow$	NYU v2 $\downarrow$	KITTI $\downarrow$
Method		Indoor	Outdoor	Outdoor	Outdoor	Outdoor	Indoor	Indoor	Outdoor
DepthAnything [21]		52.3	9.3	39.3	35.4	6.9	85.0	4.3	7.6
DepthAnything V2 [22]		59.5	36.3	37.2	17.7	5.9	72.4	4.4	7.4
Metric3D [57]		4.7	34.2	13.6	64.4	17.3	16.9	8.3	5.8
Metric3D v2 [83]		39.4	87.7	29.9	82.6	38.3	75.6	4.5	3.9
PatchFusion [75]		22.6	51.8	49.9	20.4	14.0	53.6	-	-
UniDepth [68]		27.6	25.3	31.9	83.6	16.5	95.8	5.78	4.2
ZeroDepth [18]		-	-	46.5	64.3	12.9	-	8.4	10.5
ZoeDepth [18]		21.6	34.2	53.8	28.1	7.8	85.7	7.7	5.7
Depth Pro [19]		46.6	41.5	60.5	49.1	40.0	89.0	-	-

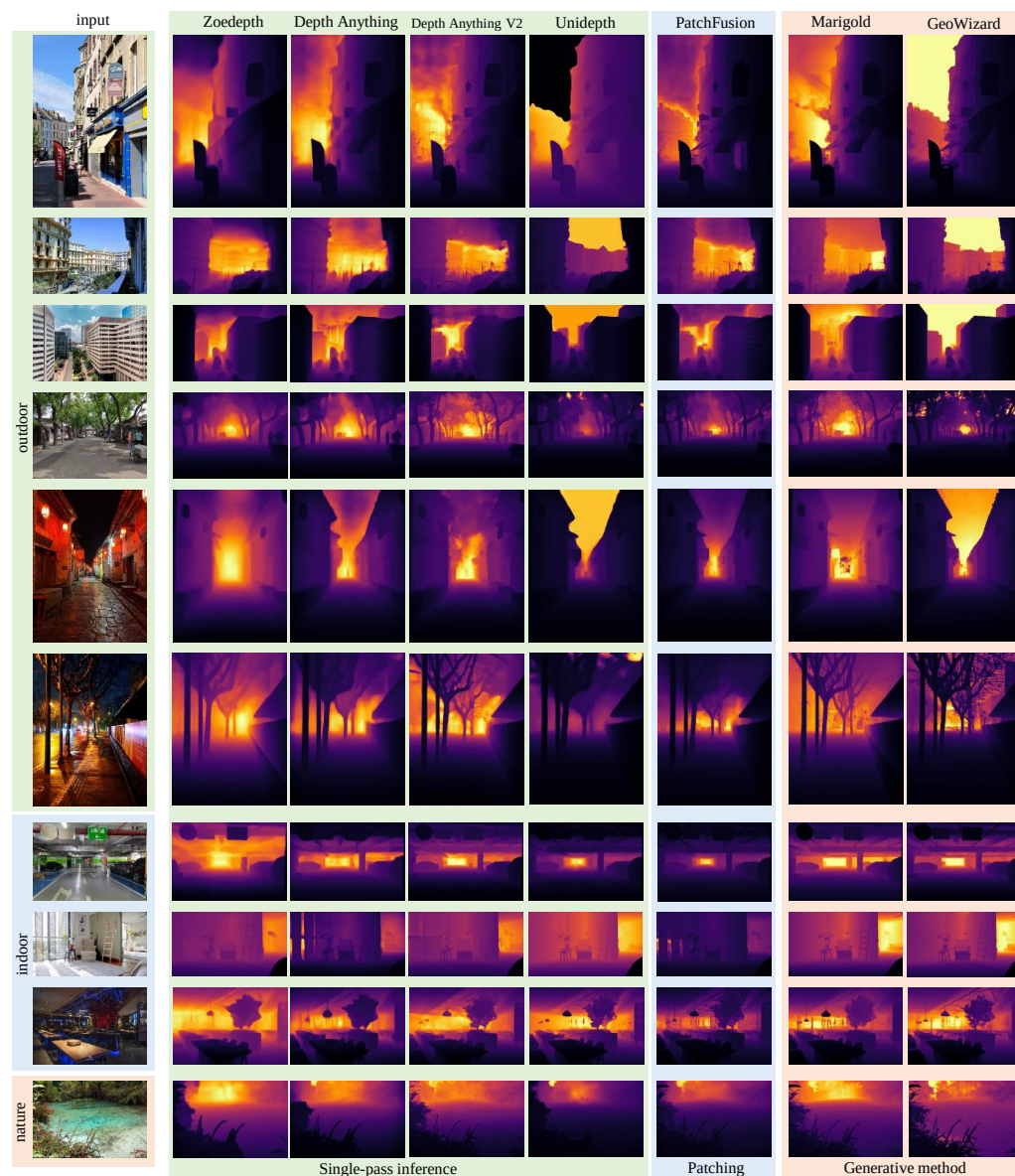


Figure 2. The analysis compares model performance across diverse scenarios, including outdoor and indoor scenes, streets and buildings, large- and small-scale environments, urban and natural settings, and varying lighting conditions. Colors in the figure distinguish scene types and method categories. Generative approaches yield relative depth, while other methods produce absolute depth values.

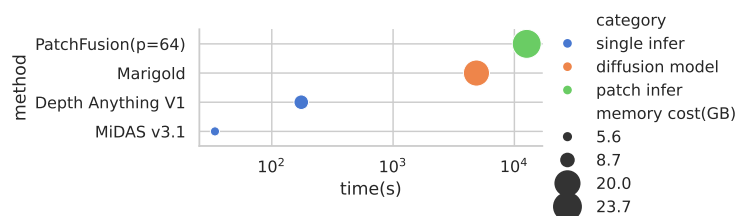


Figure 3. Inference time and memory usage for different model types are shown on a logarithmic scale in seconds.

4.4. Datasets for MMDE

Table A1 provides a comprehensive overview of datasets used for MMDE and identifies several key trends. Among the 38 datasets reviewed, 21 were collected in real-world outdoor environments, which highlights the strong emphasis on outdoor scenes, particu-

larly those designed for autonomous driving. Landmark datasets such as KITTI, Waymo Open Dataset, and nuScenes contain extensive multi-sensor data that include RGB images, LiDAR, and radar. These datasets provide highly accurate metric depth, which is essential for large-scale training and rigorous benchmarking.

Although fewer in number, indoor datasets remain critical for applications in robotics, augmented reality, and indoor navigation. Examples such as ScanNet, NYU Depth V2, and SUN RGB-D offer high-quality RGB-D imagery with reliable metric ground truth. These resources are indispensable for developing models that must perceive and interact with indoor spaces with both accuracy and robustness.

Driving-oriented datasets constitute a substantial proportion of the available resources, with at least twelve collected directly from moving vehicles. This subset includes all major autonomous driving benchmarks, which support the training of models that must operate reliably under dynamic and unpredictable road conditions.

The distinction between synthetic and real-world datasets has become increasingly important. Although most datasets are collected in real environments, synthetic resources such as TartanAir, Hypersim, and vKITTI have gained prominence. Synthetic datasets provide perfectly aligned and noise-free metric depth, which makes them valuable for augmenting real-world data or for evaluating models under controlled conditions that simulate complex scenarios. The complementary use of synthetic and real-world datasets mitigates limitations in real data acquisition, such as missing values, sensor noise, or incomplete annotations.

Datasets also differ widely in sensor configurations and data modalities, ranging from RGB-only captures to multi-sensor collections that integrate LiDAR, GPS, and IMU. This diversity supports both purely monocular approaches and more advanced fusion-based methods. The availability of metric depth ground truth, whether measured directly by sensors or derived through accurate calibration, remains the decisive factor for MMDE. Among the 38 datasets analyzed, 32 provide true metric depth, while six contain only relative depth. Relative depth remains useful for tasks in which absolute scale is unnecessary, whereas metric depth is indispensable for applications such as 3D reconstruction, robotics, and autonomous navigation.

5. Discussion

The evolution of monocular MMDE reflects a continuous effort to balance structural precision, computational efficiency, and generalization across domains. Three paradigms—single-pass, patch-based, and generative approaches—represent distinct trade-offs along this spectrum. Single-pass models perform inference through one forward computation, which ensures high speed and low latency. This property makes them particularly effective for real-time tasks such as autonomous navigation and interactive rendering. However, their reliance on global representations often leads to the loss of high-frequency details and structural fidelity, especially in scenes with fine textures or complex boundaries. The dependence on large-scale, well-labeled datasets further limits their adaptability in diverse or weakly annotated environments.

Patch-based strategies attempt to overcome these weaknesses by processing localized image regions independently before fusing them into a coherent depth map. This mechanism allows finer spatial resolution and more accurate edge delineation. However, the linear increase in inference time with the number of patches introduces scalability concerns. While improved fusion algorithms—such as those used in PatchFusion—mitigate cross-patch discontinuities, the requirement for repeated inference cycles still constrains their practicality in high-resolution or real-time applications. The trade-off between local precision and computational cost remains a core limitation.

Generative diffusion models redefine the problem by iteratively refining predictions through denoising processes, thereby capturing complex geometric relationships and multi-scale scene dependencies. These models achieve superior structural coherence and realistic depth reconstruction compared with discriminative methods. For example, Marigold demonstrates strong performance in maintaining spatial continuity in indoor layouts, while Diffusion for Metric Depth (DMD) employs logarithmic depth parameterization and field-of-view conditioning to deliver scale-aware zero-shot predictions. Despite their advantages, diffusion-based models are still constrained by multi-step inference, stochastic variation, and high computational demand, which hinder their integration into real-time systems. Moreover, most diffusion-based pipelines remain oriented toward relative rather than metric depth, limiting their use in applications such as SLAM or precise 3D mapping.

The increasing integration of synthetic and real datasets has also transformed MMDE training paradigms. Synthetic datasets provide pixel-perfect annotations that capture reflective, transparent, or occluded regions, while real-world datasets preserve authentic visual statistics. Hybrid training regimes, supported by uncertainty-guided supervision as in StableDepth, improve the balance between synthetic precision and real-world adaptability. Nevertheless, domain gaps persist, particularly in texture, illumination, and color distribution, which continue to challenge robust generalization.

In parallel, zero-shot generalization has emerged as a defining direction for the next generation of MMDE models. Architectures such as ZoeDepth, UniDepth, and VGGT [84] demonstrate that large-scale pretraining and multi-task learning can yield models capable of handling unseen domains without explicit scale calibration. This capability, combined with innovations in loss formulation and cross-modal supervision, marks a significant step toward universal metric depth estimation that functions across diverse sensors, environments, and scene types.

Looking ahead, three core challenges remain: achieving computational efficiency comparable to single-pass models, ensuring geometric consistency across multi-view scenes, and improving domain transfer between synthetic and real data. Generative frameworks offer a pathway to richer spatial understanding but require simplification and optimization to meet real-time constraints. MMDE thus stands at a convergence point, where progress in architecture design, loss modeling, and data synthesis is driving the field toward robust, scalable, and generalizable depth perception systems.

6. Conclusions

MMDE has evolved from narrow-domain, relative depth prediction toward scalable, cross-domain, and metric-aware frameworks that integrate generative modeling and multi-source supervision. The field now balances efficiency, structural fidelity, and generalization, with approaches ranging from single-pass models to patch-based and diffusion-based methods. The SWOC analysis (Table 3) highlights the current strengths of MMDE, including structural coherence, zero-shot adaptability, and data flexibility, while also identifying weaknesses such as high computational cost, domain gaps, and multi-view inconsistency. Opportunities lie in accelerated generative inference, improved data alignment, and unified perception pipelines, whereas remaining challenges focus on real-time deployment, geometric consistency, and robust cross-domain performance.

Future research is expected to emphasize hybrid strategies that combine efficiency with fine-detail reconstruction, leveraging synthetic-real data integration, advanced loss formulations, and generative refinement. As these developments mature, MMDE will increasingly serve as a foundational technology for 3D scene reconstruction, autonomous navigation, AR/VR applications, and interactive perception systems, providing accurate and generalizable metric depth across diverse environments.

Table 3. SWOC analysis of different MMDE method types, highlighting their strengths, weaknesses, and potential opportunities or challenges for real-world deployment.

Method Type	Strengths	Weaknesses	Opportunities/Challenges
Single-pass	Efficient real-time inference; low memory footprint; simple pipeline	Loss of fine structural details; limited ability to model complex geometries; performance heavily depends on large-scale annotated datasets	Opportunity to combine with patch- or diffusion-based refinements; challenge to maintain high-frequency details while preserving efficiency
Patching	Improved edge and structure preservation; scalable to high-resolution inputs; supports localized refinement	Increased inference time; cross-patch consistency issues; diminishing returns with excessive patch division	Opportunity to optimize fusion strategies; challenge to balance resolution, consistency, and computation for real-time deployment
Generative diffusion	Captures complex geometries and fine structures; reduced reliance on dense supervision; strong zero-shot adaptability	High computational cost; iterative inference slows deployment; limited metric depth progress in some frameworks	Opportunity for accelerated diffusion and hybrid architectures; challenge to enforce geometric consistency and achieve real-time performance
Conditioned sensor-assisted	Leverages priors and additional modalities to enhance depth accuracy; improved robustness in difficult scenes	Dependency on sensor quality or prompt design; limited generalization without carefully curated conditions	Opportunity for unified perception pipelines in robotics and autonomous systems; challenge to balance additional input modalities with model complexity

Author Contributions: Conceptualization, J.Z.; Methodology, J.Z.; Software, J.Z.; Validation, Y.W.; Formal analysis, J.Z.; Investigation, J.Z.; Resources, Y.W.; Data curation, J.Z.; Writing—original draft preparation, J.Z.; Writing—review and editing, J.Z. and Y.W.; Visualization, J.Z.; Supervision, H.J.; Project administration, J.Z. and H.J.; Funding acquisition, H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Conflicts of Interest: Author Jiuling Zhang and Author Huilong Jiang have received research grants from CRRC Technology Innovation (Beijing) Co., Ltd. and CRRC Dalian Co., Ltd.

Appendix A

Table A1. This table presents a comprehensive overview of datasets that are commonly used for MMDE and related computer vision tasks. Each dataset is described in terms of scene type, modality, and annotation quality, which directly determine its suitability for depth estimation. Datasets labeled as Metric in the “Relative/Metric” column, such as NYU-D, KITTI, and ApolloScape, provide RGB images paired with accurate ground-truth depth maps, which makes them essential for training and evaluating depth models under absolute distance supervision. Datasets labeled as Relative, such as DIW, Movies, and WSVD, offer only ordinal relationships rather than metric distances, and therefore serve better for methods focusing on relative depth prediction. Synthetic or mixed datasets, including BlendedMVS and TartanAir, supply noise-free annotations and perfectly aligned labels, which makes them valuable for domain adaptation and augmentation of real-world data. The table is organized to assist researchers in selecting datasets according to specific goals, whether for autonomous driving, indoor scene understanding, or multi-task learning. Each column contains specific information: Name specifies the dataset title; Indoor/Outdoor identifies the scene type; Driving Data indicates whether data were collected from a moving vehicle; Synthetic/Real clarifies the source of the dataset; Tasks lists the supported computer vision applications; Data Categories describe available sensor modalities and annotations; Relative/Metric specifies the type of depth supervision; and Description provides a concise summary of key features.

Name	Indoor Outdoor	Driving Data	Synthetic Real	Tasks	Data Categories	Relative Metric	Description
Argoverse2	Outdoor	Yes	Real	Trajectory Prediction, Object Detection, Depth Estimation, Semantic Segmentation, SLAM	RGB, LiDAR, GPS, IMU, 3D BBoxes, Labels	Metric	A large-scale autonomous driving dataset that provides 360-degree LiDAR and stereo data, supporting long-term tracking and motion forecasting.

Table A1. Cont.

Name	Indoor Outdoor	Driving Data	Synthetic Real	Tasks	Data Categories	Relative Metric	Description
Waymo	Outdoor	Yes	Real	Object Detection, Depth Estimation, Semantic Segmentation, Trajectory Prediction, SLAM	RGB, LiDAR, GPS, IMU, 3D BBoxes, Labels	Metric	A large-scale dataset collected in both urban and highway environments, featuring multi-sensor data for autonomous driving tasks.
DrivingStereo	Outdoor	No	Real	Stereo Matching, Depth Estimation	High-Resolution Stereo Images, Depth Maps	Metric	A stereo dataset with high-resolution binocular images and ground-truth depth maps, designed for stereo vision and depth estimation.
Cityscapes	Outdoor	No	Real	Semantic Segmentation, Instance Segmentation, Depth Estimation	RGB Images, Semantic Labels	Metric	A widely used benchmark of urban street scenes that supports segmentation tasks and provides derived depth information.
BDD100K	Outdoor	Yes	Real	Object Detection, Semantic Segmentation, Driving Behavior Prediction, Depth Estimation	RGB, Semantic Labels, Videos, GPS, IMU	Relative	A large-scale driving dataset that covers diverse scenes and supports multiple tasks including object detection and driving behavior analysis.
Mapillary Vistas	Outdoor	No	Real	Semantic Segmentation, Instance Segmentation, Depth Estimation	RGB Images, Semantic Labels, Depth Maps	Relative	A large-scale street-level dataset that provides diverse scenes with rich semantic annotations for segmentation and depth estimation.
A2D2	Indoor Outdoor	Yes	Real	Semantic Segmentation, Object Detection, Depth Estimation, SLAM	RGB, Semantic Labels, LiDAR, IMU, GPS, 3D BBoxes	Metric	A multi-sensor dataset that covers both indoor and outdoor scenes, offering detailed annotations for driving-related tasks.
ScanNet	Indoor	No	Real	3D Reconstruction, Semantic Segmentation, Depth Estimation	RGB-D, Point Clouds, Semantic Labels	Metric	A large-scale dataset of indoor environments that provides RGB-D imagery and 3D point clouds for reconstruction and semantic understanding.
Taskonomy	Indoor Outdoor	No	Real	Multi-Task Learning, Depth Estimation, Semantic Segmentation	RGB, Depth Maps, Normals, Point Clouds	Metric	A dataset designed for multi-task learning, offering a broad set of ground-truth annotations across different vision tasks.
SUN-RGBD	Indoor	No	Real	3D Reconstruction, Semantic Segmentation, Depth Estimation	RGB-D, Point Clouds, Semantic Labels	Metric	A large-scale indoor dataset that provides RGB-D imagery and semantic labels for depth estimation and 3D reconstruction.
Diode Indoor	Indoor	No	Real	Depth Estimation, 3D Reconstruction	RGB, LiDAR Depth Maps, Point Clouds	Metric	A high-precision indoor dataset that combines LiDAR depth maps with RGB data for accurate metric depth estimation.
IBims-1	Indoor	No	Real	3D Reconstruction, Depth Estimation	RGB-D, 3D Reconstruction	Metric	A benchmark dataset that provides high-quality RGB-D data and 3D reconstructions of indoor building environments.
VOID	Indoor	No	Real	3D Reconstruction, Depth Estimation, SLAM	RGB-D, Point Clouds	Metric	A dataset of indoor RGB-D scenes with a focus on occlusion handling for reconstruction and SLAM.
HAMMER	Indoor Outdoor	No	Real	Depth Estimation, 3D Reconstruction, Semantic Segmentation, SLAM	RGB-D, Point Clouds, Semantic Labels	Metric	A dataset covering both indoor and outdoor environments, designed for depth estimation and geometric reconstruction with high accuracy.
ETH-3D	Indoor Outdoor	No	Real	Multi-View Stereo, Depth Estimation, 3D Reconstruction	High-Res RGB, Depth Maps, Point Clouds	Metric	A benchmark for multi-view stereo and 3D reconstruction that provides high-resolution RGB images and precise depth ground truth.
nuScenes	Outdoor	Yes	Real	Object Detection, Trajectory Prediction, Depth Estimation, SLAM	RGB, LiDAR, Radar, GPS, IMU, 3D BBoxes, Labels	Metric	A large-scale driving dataset that integrates RGB, LiDAR, and radar sensors to support autonomous navigation tasks.
DDAD	Outdoor	Yes	Real	Depth Estimation, Object Detection, 3D Reconstruction, SLAM	RGB-D, Point Clouds	Metric	A driving dataset with dense annotations and multi-sensor inputs, focused on high-quality depth estimation and 3D reconstruction.
BlendedMVS	Indoor Outdoor	No	Synthetic Real	Multi-View Stereo, Depth Estimation, 3D Reconstruction	RGB, Depth Maps, 3D Reconstruction	Metric	A hybrid dataset that blends real and synthetic imagery to support depth estimation and multi-view stereo benchmarks.
DIML	Indoor	No	Real	Stereo Matching, Depth Estimation	RGB Images, Depth Maps	Metric	A multi-view indoor dataset designed for stereo matching and monocular depth estimation.
HRWSI	Outdoor	No	Real	Stereo Matching, Depth Estimation	High-Resolution RGB, Depth Maps	Metric	A high-resolution dataset that provides RGB imagery and depth maps of outdoor scenes for stereo and depth tasks.
IRS	Indoor	No	Real	3D Reconstruction, Depth Estimation, SLAM	RGB-D, Point Clouds	Metric	An indoor RGB-D dataset created for reconstruction and depth estimation with emphasis on SLAM evaluation.
MegaDepth	Outdoor	No	Real	3D Reconstruction, Depth Estimation	High-Res RGB, Depth Maps	Relative	A large-scale dataset of outdoor scenes that provides high-resolution imagery with relative depth annotations.

Table A1. Cont.

Name	Indoor Outdoor	Driving Data	Synthetic Real	Tasks	Data Categories	Relative Metric	Description
TartanAir	Indoor Outdoor	No	Synthetic	SLAM, Depth Estimation, 3D Reconstruction, Semantic Segmentation	RGB, Depth Maps, Point Clouds, Labels	Metric	A synthetic dataset that offers diverse indoor and outdoor environments for SLAM, reconstruction, and depth learning.
Hypersim	Indoor Outdoor	No	Synthetic	3D Reconstruction, Scene Understanding, Semantic Segmentation	RGB, Depth Maps, Point Clouds, Labels	Metric	A photorealistic synthetic dataset designed to support 3D reconstruction and semantic scene understanding tasks.
vKITTI	Outdoor	Yes	Synthetic	Object Detection, Semantic Segmentation, Depth Estimation, SLAM	Synthetic RGB, Labels, Depth Maps, 3D BBoxes	Metric	A synthetic driving dataset that provides ground-truth labels for object detection, depth estimation, and segmentation tasks.
KITTI	Outdoor	Yes	Real	Object Detection, Stereo Matching, Depth Estimation, SLAM	RGB, LiDAR, Depth Maps, Semantic Labels	Metric	A foundational autonomous driving dataset with multi-sensor data, widely used for vision and robotics benchmarks.
NYU-D	Indoor	No	Real	Semantic Segmentation, Depth Estimation, 3D Reconstruction	RGB-D, Semantic Labels, Point Clouds	Metric	A benchmark indoor RGB-D dataset that provides semantic labels and depth information for depth estimation and segmentation tasks.
Sintel	Outdoor	No	Synthetic	Optical Flow, Depth Estimation, Semantic Segmentation	RGB, Depth Maps, Labels, Optical Flow	Metric	A synthetic dataset that resembles movie scenes, supporting depth estimation, segmentation, and optical flow tasks.
ReDWeb	Outdoor	No	Real	Monocular Depth Estimation	RGB Images	Relative	A dataset created from diverse web videos, designed to supervise monocular depth estimation with relative annotations.
Movies	Indoor Outdoor	No	Synthetic Real	Monocular Depth Estimation, Zero-Shot Learning	RGB Images	Relative	A blended dataset sourced from various movies, used primarily for zero-shot monocular depth estimation.
ApolloScape	Outdoor	Yes	Real	Object Detection, Semantic Segmentation, Depth Estimation	RGB, LiDAR	Metric	A driving dataset with dense annotations that supports object detection, semantic segmentation, and depth estimation.
WSVD	Indoor Outdoor	No	Real	Monocular Depth Estimation	RGB Images	Metric	A dataset collected from web videos that emphasizes dynamic scenes and moving objects for depth learning.
DIW	Outdoor	No	Real	Monocular Depth Estimation	RGB Images	Relative	A dataset that provides relative depth annotations for diverse outdoor scenes.
ETH3D	Indoor Outdoor	No	Real	Multi-View Stereo, Depth Estimation	High-Res RGB, Videos	Metric	A benchmark dataset for multi-view stereo and depth estimation, containing high-resolution imagery.
TUM	Indoor	No	Real	RGB-D SLAM	RGB-D	Metric	An indoor RGB-D dataset widely used for evaluating SLAM algorithms and monocular depth estimation.
3D Ken Burns	Indoor	No	Synthetic	Depth Estimation, Image Animation	Static Images, Depth Maps	Metric	A dataset designed for producing image-based animations with depth information, such as the "Ken Burns" effect.
Objaverse	Indoor Outdoor	No	Synthetic Real	3D Object Detection, Classification	3D Models	Relative	A large-scale dataset of 3D objects designed for recognition, classification, and generative modeling.
OmniObject3D	Indoor	No	Synthetic	3D Object Detection, Recognition, Reconstruction	3D Models, RGB-D	Metric	A synthetic dataset that provides multi-view images and aligned depth information for 3D object understanding.

References

1. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [[CrossRef](#)]
2. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **2023**, *42*, 1–14. [[CrossRef](#)]
3. Ye, C.; Nie, Y.; Chang, J.; Chen, Y.; Zhi, Y.; Han, X. Gaustudio: A modular framework for 3D gaussian splatting and beyond. *arXiv* **2024**, arXiv:2403.19632. [[CrossRef](#)]
4. Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Nature: Cham, Switzerland, 2022.
5. Zheng, J.; Lin, C.; Sun, J.; Zhao, Z.; Li, Q.; Shen, C. Physical 3D adversarial attacks against monocular depth estimation in autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 24452–24461.
6. Leduc, A.; Cioppa, A.; Giancola, S.; Ghanem, B.; Van Droogenbroeck, M. SoccerNet-Depth: A scalable dataset for monocular depth estimation in sports videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 3280–3292.

7. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 3836–3847.
8. Khan, N.; Xiao, L.; Lanman, D. Tiled multiplane images for practical 3D photography. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 10454–10464.
9. Liew, J.H.; Yan, H.; Zhang, J.; Xu, Z.; Feng, J. Magicedit: High-fidelity and temporally coherent video editing. *arXiv* **2023**, arXiv:2308.14749.
10. Xu, D.; Jiang, Y.; Wang, P.; Fan, Z.; Wang, Y.; Wang, Z. Neurallift-360: Lifting an in-the-wild 2D photo to a 3D object with 360deg views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 4479–4489.
11. Shahbazi, M.; Claessens, L.; Niemeyer, M.; Collins, E.; Tonioni, A.; Van Gool, L.; Tombari, F. Inerf: Text-driven generative object insertion in neural 3D scenes. *arXiv* **2024**, arXiv:2401.05335.
12. Shriram, J.; Trevithick, A.; Liu, L.; Ramamoorthi, R. Realmdreamer: Text-driven 3D scene generation with inpainting and depth diffusion. *arXiv* **2024**, arXiv:2404.07199.
13. Deng, J.; Yin, W.; Guo, X.; Zhang, Q.; Hu, X.; Ren, W.; Long, X.X.; Tan, P. Boost 3D reconstruction using diffusion-based monocular camera calibration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 7110–7121.
14. Guo, J.; Ding, Y.; Chen, X.; Chen, S.; Li, B.; Zou, Y.; Lyu, X.; Tan, F.; Qi, X.; Li, Z.; et al. Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4D driving scene generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 27231–27241.
15. Daxberger, E.; Wenzel, N.; Griffiths, D.; Gang, H.; Lazarow, J.; Kohavi, G.; Kang, K.; Eichner, M.; Yang, Y.; Dehghan, A.; et al. Mm-spatial: Exploring 3D spatial understanding in multimodal llms. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 7395–7408.
16. Yu, Y.; Liu, S.; Pautrat, R.; Pollefeys, M.; Larsson, V. Relative pose estimation through affine corrections of monocular depth priors. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 16706–16716.
17. Kim, H.; Baik, S.; Joo, H. DAViD: Modeling dynamic affordance of 3D objects using pre-trained video diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 10330–10341.
18. Bhat, S.F.; Birkl, R.; Wofk, D.; Wonka, P.; Müller, M. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv* **2023**, arXiv:2302.12288.
19. Bochkovskiy, A.; Delaunoy, A.; Germain, H.; Santos, M.; Zhou, Y.; Richter, S.; Koltun, V. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. In Proceedings of the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
20. Saxena, S.; Hur, J.; Herrmann, C.; Sun, D.; Fleet, D.J. Zero-shot metric depth with a field-of-view conditioned diffusion model. *arXiv* **2023**, arXiv:2312.13252.
21. Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; Zhao, H. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 10371–10381.
22. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth anything v2. In *Advances in Neural Information Processing Systems 37*; Curran Associates, Inc.: New York, NY, USA, 2024; pp. 21875–21911.
23. Guo, Y.; Garg, S.; Miangoleh, S.M.H.; Huang, X.; Ren, L. Depth any camera: Zero-shot metric depth estimation from any camera. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10–17 June 2025; pp. 26996–27006.
24. Bhoi, A. Monocular depth estimation: A survey. *arXiv* **2019**, arXiv:1901.09402. [[CrossRef](#)]
25. Khan, F.; Salahuddin, S.; Javidnia, H. Deep learning-based monocular depth estimation methods—A state-of-the-art review. *Sensors* **2020**, *20*, 2272. [[CrossRef](#)]
26. Zhao, C.; Sun, Q.; Zhang, C.; Tang, Y.; Qian, F. Monocular depth estimation based on deep learning: An overview. *Sci. China Technol. Sci.* **2020**, *63*, 1612–1627. [[CrossRef](#)]
27. Ruan, X.; Yan, W.; Huang, J.; Guo, P.; Guo, W. Monocular depth estimation based on deep learning: A survey. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; IEEE: New York, NY, USA, 2021; pp. 2436–2440.
28. Lahiri, S.; Ren, J.; Lin, X. Deep learning-based stereopsis and monocular depth estimation techniques: A review. *Vehicles* **2024**, *6*, 305–351. [[CrossRef](#)]

29. Tosi, F.; Ramirez, P.Z.; Poggi, M. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer Nature: Cham, Switzerland, 2024; pp. 236–257.
30. Vyas, P.; Saxena, C.; Badapanda, A.; Goswami, A. Outdoor monocular depth estimation: A research review. *arXiv* **2022**, arXiv:2205.01399. [[CrossRef](#)]
31. Dong, X.; Garratt, M.A.; Anavatti, S.G.; Abbass, H.A. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 16940–16961. [[CrossRef](#)]
32. Masoumian, A.; Rashwan, H.A.; Cristiano, J.; Asif, M.S.; Puig, D. Monocular depth estimation using deep learning: A review. *Sensors* **2022**, *22*, 5353. [[CrossRef](#)]
33. Arampatzakis, V.; Pavlidis, G.; Mitianoudis, N.; Papamarkos, N. Monocular depth estimation: A thorough review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 2396–2414. [[CrossRef](#)]
34. Rajapaksha, U.; Sohel, F.; Laga, H.; Diepeveen, D.; Bennamoun, M. Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. *ACM Comput. Surv.* **2024**, *56*, 1–51. [[CrossRef](#)]
35. Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R.C.; Schindler, K. Repurposing diffusion-based image generators for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 9492–9502.
36. Bhat, S.F.; Alhashim, I.; Wonka, P. Adabins: Depth estimation using adaptive bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4009–4018.
37. Miangoleh, S.M.H.; Dille, S.; Mai, L.; Paris, S.; Aksoy, Y. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9685–9694.
38. Jampani, V.; Chang, H.; Sargent, K.; Kar, A.; Tucker, R.; Krainin, M.; Liu, C. Slide: Single image 3D photography with soft layering and depth-aware inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12518–12527.
39. Izadi, S.; Kim, D.; Hilliges, O.; Molyneaux, D.; Newcombe, R.; Kohli, P.; Shotton, J.; Hodges, S.; Freeman, D.; Davison, A.; et al. Kinectfusion: Real-time 3D reconstruction and interaction using a moving depth camera. In Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, 16–19 October 2011; pp. 559–568.
40. Foix, S.; Alenya, G.; Torras, C. Lock-in time-of-flight (ToF) cameras: A survey. *IEEE Sens. J.* **2011**, *11*, 1917–1926. [[CrossRef](#)]
41. Sarbol, H.; Lefloch, D.; Kolb, A. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Comput. Vis. Image Underst.* **2015**, *139*, 1–20. [[CrossRef](#)]
42. He, Y.; Liang, B.; Zou, Y.; He, J.; Yang, J. Depth errors analysis and correction for time-of-flight (ToF) cameras. *Sensors* **2017**, *17*, 92. [[CrossRef](#)]
43. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [[CrossRef](#)]
44. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 328–341. [[CrossRef](#)]
45. Kok, K.Y.; Rajendran, P. A review on stereo vision algorithm: Challenges and solutions. *ECTI Trans. Comput. Inf. Technol. (ECTI-CIT)* **2019**, *13*, 112–128. [[CrossRef](#)]
46. Wofk, D.; Ranftl, R.; Müller, M.; Koltun, V. Monocular visual-inertial depth estimation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; IEEE: New York, NY, USA, 2023; pp. 6095–6101.
47. Singh, A.D.; Ba, Y.; Sarker, A.; Zhang, H.; Kadambi, A.; Soatto, S.; Srivastava, M.; Wong, A. Depth estimation from camera image and mmwave radar point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 9275–9285.
48. Huang, J.; Zhou, Q.; Rabeti, H.; Korovko, A.; Ling, H.; Ren, X.; Shen, T.; Gao, J.; Slepichev, D.; Lin, C.H.; et al. Vipe: Video pose engine for 3D geometric perception. *arXiv* **2025**, arXiv:2508.10934.
49. Garg, R.; Bg, V.K.; Carneiro, G.; Reid, I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VIII 14; Springer International Publishing: Cham, Switzerland, 2016; pp. 740–756.
50. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12179–12188.
51. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
52. Eigen, D.; Puhersch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems 27*; Curran Associates, Inc.: New York, NY, USA, 2014.

53. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
54. Birkl, R.; Wofk, D.; Müller, M. Midas v3. 1—a model zoo for robust monocular relative depth estimation. *arXiv* **2023**, arXiv:2307.14460.
55. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [[CrossRef](#)]
56. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2002–2011.
57. Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Shen, C. Metric3d: Towards zero-shot metric 3D prediction from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 9043–9053.
58. Guizilini, V.; Vasiljevic, I.; Chen, D.; Ambrus, R.; Gaidon, A. Towards zero-shot scale-aware monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 9233–9243.
59. Spencer, J.; Russell, C.; Hadfield, S.; Bowden, R. Kick back & relax++: Scaling beyond ground-truth depth with slowtv & cribstv. *arXiv* **2024**, arXiv:2403.01569.
60. Bhat, S.F.; Alhashim, I.; Wonka, P. Localbins: Improving depth estimation by learning local distributions. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 480–496.
61. Li, Z.; Wang, X.; Liu, X.; Jiang, J. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Trans. Image Process.* **2024**, *33*, 3964–3976. [[CrossRef](#)]
62. Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; Tan, P. NeW CRFs: Neural window fully-connected CRFs for monocular depth estimation. *arXiv* **2022**, arXiv:2203.01502.
63. Spencer, J.; Tosi, F.; Poggi, M.; Arora, R.S.; Russell, C.; Hadfield, S.; Elder, J.H. The third monocular depth estimation challenge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 1–14.
64. Marsal, R.; Chabot, F.; Loesch, A.; Grolleau, W.; Sahbi, H. MonoProb: Self-supervised monocular depth estimation with interpretable uncertainty. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 3637–3646.
65. Haji-Esmaili, M.M.; Montazer, G. Large-scale monocular depth estimation in the wild. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107189. [[CrossRef](#)]
66. Shao, S.; Pei, Z.; Chen, W.; Sun, D.; Chen, P.C.; Li, Z. MonoDiffusion: Self-supervised monocular depth estimation using diffusion model. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *35*, 3664–3678. [[CrossRef](#)]
67. Wang, Y.; Liang, Y.; Xu, H.; Jiao, S.; Yu, H. Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, Stanford, CA, USA, 25–27 March 2024; Volume 38, pp. 5713–5721.
68. Piccinelli, L.; Yang, Y.H.; Sakaridis, C.; Segu, M.; Li, S.; Van Gool, L.; Yu, F. UniDepth: Universal monocular metric depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 10106–10116.
69. Sun, B.; Jin, M.; Yin, B.; Hou, Q. Depth Anything at Any Condition. *arXiv* **2025**, arXiv:2507.01634. [[CrossRef](#)]
70. Lin, H.; Peng, S.; Chen, J.; Peng, S.; Sun, J.; Liu, M.; Bao, H.; Feng, J.; Zhou, X.; Kang, B. Prompting depth anything for 4k resolution accurate metric depth estimation. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 17070–17080.
71. Wang, R.; Xu, S.; Dai, C.; Xiang, J.; Deng, Y.; Tong, X.; Yang, J. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 5261–5271.
72. Zhang, Z.; Yang, L.; Yang, T.; Yu, C.; Guo, X.; Lao, Y.; Zhao, H. StableDepth: Scene-Consistent and Scale-Invariant Monocular Depth. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 7069–7078.
73. Wang, Z.; Chen, S.; Yang, L.; Wang, J.; Zhang, Z.; Zhao, H.; Zhao, Z. Depth Anything with Any Prior. *arXiv* **2025**, arXiv:2505.10565. [[CrossRef](#)]
74. Wang, Y.; Li, J.; Hong, C.; Li, R.; Sun, L.; Song, X.; Wang, Z.; Cao, Z.; Lin, G. TacoDepth: Towards Efficient Radar-Camera Depth Estimation with One-stage Fusion. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 10523–10533.

75. Li, Z.; Bhat, S.F.; Wonka, P. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 10016–10025.
76. Li, Z.; Bhat, S.F.; Wonka, P. PatchRefiner: Leveraging Synthetic Data for Real-Domain High-Resolution Monocular Metric Depth Estimation. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer Nature: Cham, Switzerland, 2024; pp. 250–267.
77. Duan, Y.; Guo, X.; Zhu, Z. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer Nature: Cham, Switzerland, 2024; pp. 432–449.
78. Zavadski, D.; Kalšan, D.; Rother, C. Primedepth: Efficient monocular depth estimation with a stable diffusion preimage. In Proceedings of the Asian Conference on Computer Vision, Hanoi, Vietnam, 8–12 December 2024; pp. 922–940.
79. Patni, S.; Agarwal, A.; Arora, C. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 28285–28295.
80. Fu, X.; Yin, W.; Hu, M.; Wang, K.; Ma, Y.; Tan, P.; Long, X. Geowizard: Unleashing the diffusion priors for 3D geometry estimation from a single image. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer Nature: Cham, Switzerland, 2024; pp. 241–258.
81. Pham, D.H.; Do, T.; Nguyen, P.; Hua, B.S.; Nguyen, K.; Nguyen, R. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 17060–17069.
82. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1623–1637. [[CrossRef](#)] [[PubMed](#)]
83. Hu, M.; Yin, W.; Zhang, C.; Cai, Z.; Long, X.; Chen, H.; Wang, K.; Yu, G.; Shen, C.; Shen, S. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10579–10596. [[CrossRef](#)] [[PubMed](#)]
84. Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupperecht, C.; Novotny, D. Vgggt: Visual geometry grounded transformer. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 5294–5306.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.