Published in AJSE, Vol: 24, Issue:3
Received on 7th May 2026
Revised on 31st May 2026
Published on 31th July 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)
https://doi.org/10.53799/zsh31s95

The Evolution of Monocular Depth Estimation: From Spatial Regression to Generative Foundations and the Reliability Gaps

Hasan Mahmud Shanto, Mohammad Tofiqul Islam, Muhammad Ryan Hasan, Supprio Saha Pranto, Tanvir Ahmed*, Rifath Mahmud, Syeda Anika Tasnim

***Abstract*— Monocular Depth Estimation (MDE) is one of the most rigorously studied problems in modern computer vision, yet it is fundamentally ill-posed. Recovering absolute three-dimensional geometry from a single two-dimensional projection is mathematically impossible without strong inductive priors. This paper presents a structurally organized review of MDE's evolution spanning from 2005 to 2026. We trace the trajectory from handcrafted Markov random fields to brute-force pixel-wise regression with convolutional neural networks (CNNs), and then to photometric self-supervision, which liberated the field from expensive LiDAR sensor suites. We further analyzed the vision transformers (ViTs) and generative diffusion priors have achieved unprecedented zero-shot metric generalization - with models such as Metric3D v2 attaining an Absolute Relative Error (AbsRel) as low as 0.039 on the KITTI benchmark. We also show that the standard self supervised baseline Monodepth2 degrades catastrophically to an AbsRel of 1.185 under nighttime conditions in NuScenes-Night dataset, which is a massive performance collapse from its clear-weather baseline. Physical-prior models such as PhysDepth recover this to 0.118 AbsRel by embedding Rayleigh scattering theory directly into the network. At the efficiency frontier, architectures such as LEDepth achieve 0.101 AbsRel at 5.7 ms inference time with only 3.1M parameters. We conclude that the future of depth estimation lies in embedding rigid physical and spatiotemporal priors into foundational latent spaces to ensure unyielding reliability in the physical world.**

***Index Terms*— Monocular depth estimation, self-supervised learning, vision transformers, adverse weather, robustness, deep learning, autonomous vehicle.**

I. Introduction

The human visual system is remarkably adept at deciphering the three-dimensional layout of the physical world from seemingly flattened two-dimensional retinal projections. By employing an innate, subconscious synthesis of monocular cues, such as occlusion, relative size, texture gradients, and aerial perspective, humans can accurately gauge depth and navigate complex environments without strictly requiring binocular stereopsis [1]. For the better part of the last half-century, endowing robotic systems with the same intuitive geometric awareness has been considered a functionally impossible pursuit.

Being a projective transformation of 3d space, even an infinite number of object scales and distances can give rise to the same 2D pixel array provided that it is viewed at the right angle (and background/illumination conditions). Because of this fundamental scale ambiguity, most early machine perception systems successfully ignored monocular vision and instead relied on active, hardware-dominant solutions.

A. The Economic and Hardware Imperative for MDE

Historically self-driving cars and robotic platforms have achieved spatial awareness using Light Detection and Ranging (LiDAR) sensors or meticulously calibrated stereo camera rigs. But these active sensor suites come with prohibitive constraints, often rendering them unsuccessful as they significantly bottleneck their scalability. State of the art high-density LiDAR systems are extremely costly, mechanically delicate and power-hungry sensorial hardware. More critically, LiDAR pulses suffer from both backscattering and signal blocking in adverse weather, making them extremely unreliable during heavy rain, snow or fog [2]. Similar to this, stereo vision relies on rigid mechanical baselines; even a microscopic vibration between camera alignment smashes the epipolar geometry enabling useful error metrics with respect to matching disparity.

In contrast, a monocular camera is inexpensive, lightweight, passively silent, and universally deployable. If software architecture could reliably extract metric depth from a single lens, it would democratize 3D perception across domains where LiDAR or stereo rigs are physically impossible to mount. This includes weight-constrained consumer drones, augmented reality (AR) headsets, and minimally invasive robotic surgery, where navigating a monocular endoscope through the highly deformable, texture less cavities of the human gastrointestinal tract requires real-time geometric mapping to prevent tissue damage [3, 4]

Hasan Mahmud Shanto is a student of American International University Bangladesh (AIUB). (e-mail: 22-49453-3@student.aiub.edu)
Mohammad Tofiqul Islam is a student of American International University Bangladesh (AIUB). (e-mail: 22-49450-3@student.aiub.edu)
Muhammad Ryan Hasan is a student of American International University Bangladesh (AIUB). (e-mail: 22-49550-3@student.aiub.edu)
Supprio Saha Pranto is a student of American International University Bangladesh (AIUB). (e-mail: 22-49451-3@student.aiub.edu)
Tanvir Ahmed is with American International University-Bangladesh (AIUB). (e-mail: tanvir.ahmed@aiub.edu)
Rifath Mahmud is with American International University-Bangladesh (AIUB). (e-mail: rifath.mahmud@aiub.edu)
Syeda Anika Tasnim is with American International University-Bangladesh (AIUB). (e-mail: anika.tasnim@aiub.edu)

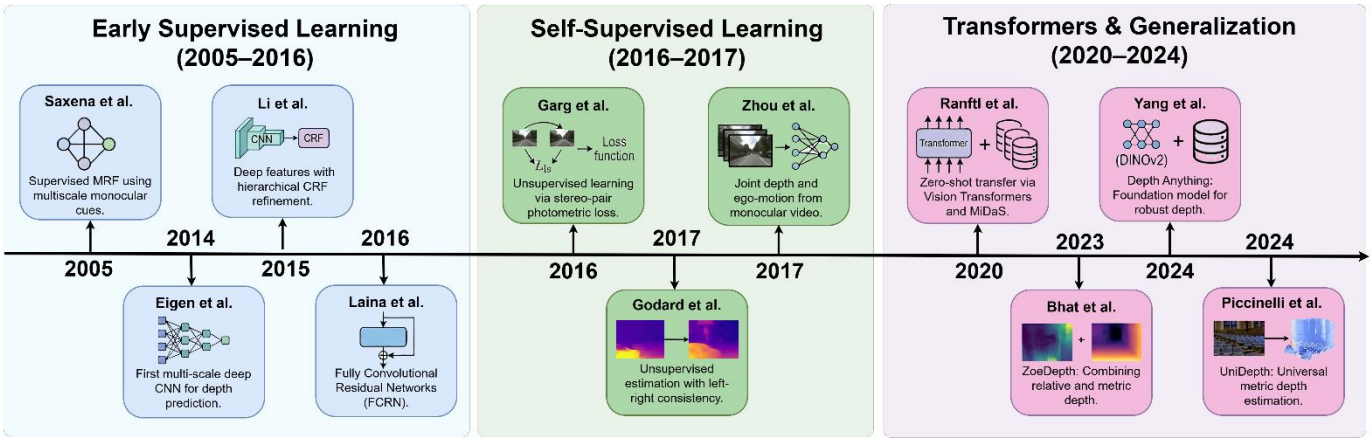


Fig. 1. Overall Paradigm of Monocular Depth Estimation Throughout the Years (Saxena et al. [6], Eigen et al. [7], Li et al. [21], Laina et al. [18], Garg et al. [10], Godard et al. [11], Zhou et al. [39], Ranftl et al. [29], Bhat et al. [62], Yang et al. [78], Piccinelli et al. [80]).

incremental progression, but as a series of distinct paradigm shifts, as mapped chronologically in Fig. 1.

The initial phase, based on early probabilistic modelling, tried to manually create the visual prompts that humans so naturally utilize. Researchers imposed low level geometric priors based on MRFs applied over superpixels, such as forcing regions that are identified to be roads to be horizontal and those identified to be buildings also vertical. But when it came to really chaotic environments this method fell apart. The following deep Convolutional Neural Networks (CNNs) also brought in the supervised regression era, where networks were trained to map RGB pixels directly to depth values generated by LiDAR.

But supervised learning soon fell scratching at a wall: it is prohibitively expensive to collect pixels-accurate ground truth data for all imaginable surroundings. This bottleneck made the second biggest change possible - the Self-Supervised Revolution. By designing depth estimation as a view-synthesis problem where a network learns geometry by generating a video frame from surrounding temporal frames, they completely removed the need for LiDAR and permitted models to learn in throughput on unlabeled video streams.

We are now in the era of its 3rd and overall largest epoch: The Foundation Era. Years being hampered with limited localized receptive fields resulted in CNNs having to infer depth via local textures ultimately lacking the global context of understanding a scene. Vision Transformers (ViTs) removed this restriction and enabled networks to study the spatial relation between each pixel concurrently. Modern foundation models have now demonstrated zero-shot metric universality in the sense that they can measure absolute distances on environments never seen explicitly during model training, coupling these global attention mechanisms with huge pseudo-labeling pipelines as well as generative diffusion priors.

C. Motivation and Scope of this Review

While multiple surveys catalog the architectural taxonomy of depth estimation, nearly all of it is limited to pure performance improvements on clean (sunlit) benchmarks. This narrow vision had hidden the crisis brewing in the field: the "Reality Gap".

Even a model that achieves state-of-the-art error rates on an easy, all-day clear-weather dataset such as KITTI will often hallucinate deadly geometric mistakes in the photon-starved blackness of a rural night, through the bright light fractalization of salty water around an underwater gas pipeline, or buried in the sensor noise footprint laid down by even moderate blizzards. Finally, running a billion-parameter ViT on an edge-compute vehicular processor under tight thermal restrictions is simply not feasible.

Consequently, this review is uniquely organised not to only chart its historical architectural progression but to profoundly challenge the deployment reality frontiers. We organize the literature in chronological clusters ending with a quantitative evaluation of state-of-the-art approaches against adversarial attacks.

The organization of this review is: Section II presents the early days and transition to supervised, multi-scale deep learning frameworks from probabilistic Markov random fields. Next, Section III, focuses on the supervised refinements and shows how progressive actors jump from regression to adaptive classification. Section IV investigates the self-supervised breakthrough and describes how networks learn geometry from pure photometric temporal motion only including monocular videos. Section V identifies the paradigm shift from sensor modalities toward global context via vision transformers and hybrid edge-efficient architectures. In Section VI we study the "reliability gap", and identify the problem with reliable physical priors required to be robust under weather, low-light, and underwater domains. From Foundation Era to Metric Zero-Shot Universality (Section VII). Finally, Sections VIII, IX and X are the formal definition of the benchmarking datasets, evaluation metrics and quantitative comparative analysis respectively. In Section XI, we discuss key research gaps that remain to be filled, such as Explainable AI (XAI) and deployment onto edge-devices. Section XII concludes the review.

D. Review Scope and Selection Process

The literature selection for this review was conducted using a defined search protocol, with pre-specified inclusion criteria developed prior to the execution of the review; a set of transparent screening stages and process; and a reproducible data-extraction template so that the resulting paper set is traceable and easily reconstructible. The protocol

was constructed in the spirit of commonly used reporting guidelines for literature reviews, such as PRISMA 2020 [5], without any claim to formal systematic review scope. The start of this publication window is 2005, which corresponds to the year of Saxena's Make3D up until 2026 so that the review should encompass work from all early depth models based on learning, to more recent foundation-era and robustness-focused works. For the searches we relied mainly on Google Scholar, complemented by IEEE Xplore, arXiv, ScienceDirect, SpringerLink and MDPI along with the publisher pages of the journals and conferences hosting the most-cited works. The reference lists of these core papers were also searched to recover earlier or neighbouring studies that keyword search alone often overlooks, with the keywords extracted from the natural vocabulary of the field and combined in chains including (but not limited to) monocular depth estimation, self-supervised depth, vision transformer depth, zero-shot metric depth, adverse weather but notably excluded nighttime and lightweight depth estimation.

A paper was retained when it addressed monocular depth estimation as a primary or substantial secondary contribution, reported quantitative results on a recognized benchmark such as KITTI, NYU Depth V2, Cityscapes, nuScenes, KITTI-C, or FLSea, and appeared in a peer-reviewed journal or major conference, or as a widely cited preprint that introduced a method or benchmark later taken up by peer-reviewed work. Papers were excluded if they were confined to stereo or multi-view setups with no monocular component, if they were workshop abstracts, posters, or theses without a full methodology, or if the full text was not retrievable. Where the same work appeared as both a preprint and a later peer-reviewed paper, the peer-reviewed version was kept.

Candidate papers were screened by their titles and abstracts adding to a second screening on the basis of full-text review using the above criteria. We extracted the following from each retained paper: publication year, learning paradigm, architectural family, training and evaluation datasets, evaluation metrics and performance, limitations explicitly mentioned by the authors of the study foremost study in focus. This homogenous exaction is what enables the comparative tables below, and why we can traverse very diverse generations of MDE on a common time bulk in the chronological synthesis.

II. Evolution of Monocular Depth Estimation: From Probabilistic Models to Self-Supervision (2005–2017)

Monocular depth estimation has undergone a fundamental transformation in approach over the past two decades. Historically, depth estimation was an ill-formed problem, as a single 2D image hypothetically corresponds to an infinite number of 3D scenes [6, 7]. Consequently, early approaches preferred motion parallax or stereo triangulation. However, the introduction of learning-based methods transformed the field from approaches based on crafted geometric priors [8] to data-driven deep learning paradigms. This section chronologically traces the evolution from Markov Random Fields (MRFs) to supervised CNNs [9], and ultimately to the emergence of self-supervised learning [10, 11].

A. The Probabilistic Foundation: Learning from Handcrafted Features

Prior to the emergence of deep learning, the field of depth estimation was largely dominated by shape-from-X techniques. Saxena et al. [6] proposed that depth cues could be learned from monocular images in a framework often referred to as Make3D, which formulated depth estimation as a discriminative modeling problem using Markov Random Fields (MRFs) [12]. Rather than treating images as a grid of independent pixels, the scene was decomposed into small homogeneous patches (superpixels) and the world was assumed to consist of piecewise planar structures [6]. Handcrafted features such as texture gradients and color filters [13] were extracted to train a model to predict the 3D position and orientation of these planes. The MRF imposed connectivity limitations between adjacent planes, restricting the predicted depth map from a randomly dispersed cloud of isolated polygons [14]. This approach worked for monocular depth recovery, but it produced “cardboard effect”, where objects look like flat cutouts with no volume [7].

In addition, handcrafted features rely limited the model's generalization to complicated non-planar geometries in real world scenes, such as landscapes from NYU Depth Dataset [15].

B. The Deep Learning Breakthrough: Multi-Scale Regression

In 2014, Eigen et al. [7] published a landmark work that essentially retired handcrafted features in favor of directly learned features from raw pixel data, radically shifting the paradigm. The main point was that one network architecture could not capture both the global context needed for understanding scene layout and the local resolution needed for the fine details. To solve this problem, Eigen et al. [12] proposed a multi-scale deep network with two different stacks, as illustrated in Fig. 2. The first stack is a coarse-scale network that takes the entire image as input to predict an approximate global depth map, and the second stack is a fine-scale network to improve the prediction by attending to local gradients. This architecture was further refined in subsequent work [16], which introduced a third scale for even finer resolution. It demonstrates that a single multiscale framework could unify the depth prediction, the surface normal estimation and the semantic labeling. A significant contribution of the 2014 work [7] is the Scale Invariant Error (SILog) loss function:

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N} d_i^2 - \frac{\lambda}{N^2}\left(\sum_{i=1}^{N} d_i\right)^2 \qquad (1)$$

where $d_i = \log y_i - \log \hat{y}_i$ where y_i is the ground truth depth, $\hat{y}_i$ is the predicted depth, N is the total number of valid pixels, and λ is a variance tuning parameter (typically set to 0.5 or 1). Since monocular cameras have no reference for absolute scale [17], this loss is therefore calculated on relative, rather than absolute differences between points, and consequently reduces error rates of over 30% compared with the state-of-the art MRF based approach by Saxena.

However, Eigen's architecture was overly dependent on fully connected layers, which either required a specific input size or destroyed spatial information. Laina et al. To alleviate this shortcoming, [18] proposed Fully Convolutional Residual Network (FCRN). They showed that rich semantic features are extracted by deeper networks using the ResNet-50 architecture [19]. More importantly, the standard up sampling layers are replaced with Up-Projection blocks. In contrast to simple bilinear interpolation [20], which may yield blurry results, these Up-Projection blocks learn up sampling weights and do this efficiently by balancing out the pooling operations performed by the encoder. Laina et al. also proposed a loss called the Reverse Huber (BerHu) loss that combines the advantages of L1 and L2 norms, and was therefore more robust to outliers in depth values than Euclidean distance.

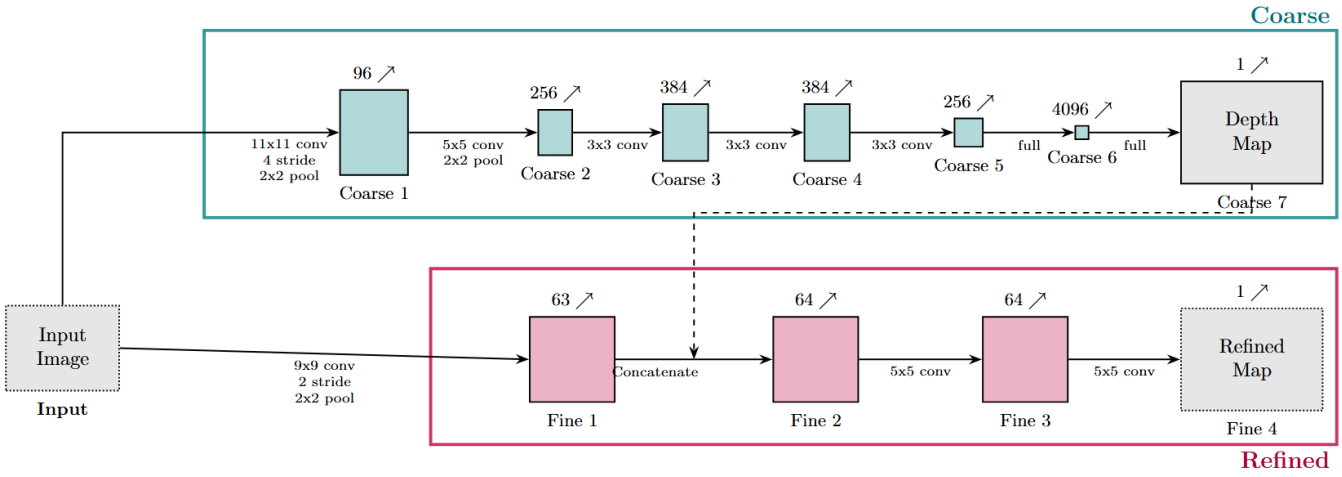


Fig. 2. Architectural diagram of the multi-scale deep network introduced by Eigen et al. [7]. A coarse network processes the full image; a fine network refines the prediction using local gradients.

C. Bridging CNNs and Probabilistic Graphical Models

CNNs are good at regression tasks; however, they often produce depth maps that are overly smooth with blurred object boundaries. To address this, the feature learning power of CNNs and the structure preserving properties of Conditional Random Fields (CRFs) were combined by the researchers [21, 22].

Li et al. [21] designed a hierarchical framework working at both the superpixel and pixel levels. Their model first used a deep CNN to regress depth values for superpixels, leveraging the semantic strength of deep features. This initial prediction was then refined using a hierarchical CRF that integrated data terms from CNN with smoothness terms respecting object boundaries defined by color and texture [20]. This hybrid approach effectively combined the best discriminative learning and probabilistic modeling.

Xu et al. [23] took this integration further with Multi-Scale Continuous CRFs (MCCRFs) implemented as a sequential deep network. Unlike Li et al., who treated the CRF as a post-processing step, Xu et al. integrated the CRF optimization directly into the neural network architecture with the help of a mean field update rule that enabled end to end training [22]. By integrating edge priors similar to those in holistically nested edge detection [24], depth discontinuities were sharpened and performance on complex scenes improved compared with pure CNN approaches.

D. The Shift to Self-Supervision: Learning without Labels

As such, by 2017, dependence on LiDAR [25] ground-truth depth data had become a major scaling bottleneck for MDE. While Garg et al. [10] have already considered the use of stereo pairs for supervision, The paradigm was disrupted by Godard et al. [11] by their work on Unsupervised Monocular Depth Estimation

Instead of formulating the problem as an error and minimizing that against some ground-truth depth map, Godard et al formulated it as image reconstruction. They proposed the Left-Right Consistency loss [11] based on stereo image pairs when training, and a single image during testing. Using ideas from view synthesis methods like Deep3D [26], the network represented the disparity maps required to warp a left image into alignment with a right image. Termed Monodepth, this method showed that self-supervised signals can outperform supervised approaches and laid the groundwork for large scale unsupervised training pipelines which characterizes the modern era of depth estimation [11].

E. Critical Summary of The Formulative Era

The years from 2005 to 2017 were truly pivotal for monocular depth estimation, but in retrospect each architectural advancement was packaged with shortcomings that were as impactful as the advancements themselves:

1) *Saxena et al. [6]:*

It was based on planar geometric assumptions of an organization that is effective in controlled settings but collapsed when faced with curves or unstructured arrangements. The system had no genuine way to process geometric complexity past simply what it's handcrafted highlights could represent, and that imperfection came through outwardly as the so-called "cardboard effect" items rendered without any real sense of mass or degree, as in the event that the world had been pressed down into a stage set.

2) *Eigen et al. [7]:*
SILog loss was meaningful, but the fully connected layers at the heart of the architecture it led to a totally different problem. Spatial detail was just not making it through the processing pipeline. These depth maps were locally implausible but globally plausible: they held together at a large scale, and became smudged across fine boundaries.

3) *Laina et al. [18]:*
Even if FCRNs and up-projection blocks mitigated much of the spatial destruction caused by fully connected layers, the model could only regress depth as a continuous variable. Standard continuous loss functions (e.g., L1, L2 or BerHu) naturally encouraged "mean-seeking" predictions. The network mathematically minimized the average penalty (loss) by smoothing over critical depth discontinuities, which yields structurally distorted 3D reconstructions.

4) *Li et al. [21] and Xu et al. [23]:*
Although, we were able to impose edge-priors and sharpen depth boundaries via the utilization of CRFs, this creates a debilitating computational trade-off. Iterative mean field updates were required for CRF convergence, and the resulting inference latency was prohibitively large, which made these hybrid models largely incompatible with real-time robot or vehicle edge deployment.

5) *Godard et al. [11]:*
Although Monodepth circumvented the LiDAR bottleneck, its foundational assumption of photometric consistency was brittle at best. The model failed miserably on occluded segments of the scene, on texture-absent surfaces predominant in indoor scenes and with non-Lambertian reflections. Moreover, the strict requirement for rigidly calibrated stereo camera rigs during training only moved the hardware dependency instead of achieving real monocular independence.

To sum up, the initial stage of monocular depth estimation was characterized by tighter couplings between spatial resolution, computational efficiency and generalization ability. The early architectures were essentially bottlenecks due to the continuous regression flaws of the design, spatial degradation by convolutional down sampling and computationally expensive probabilistic post-processing. But these initial models were still tethered to scarce LiDAR ground truth at best, and custom stereo-camera arrays, which limit how you can train it. Consequently, it rendered fundamentally unprepared for navigating the messy and open domain shifts of the physical world, hence required radical architectural pivots towards ordinal classification, adaptive binning and global attention mechanisms that characterize what became the state-of-the-art in depth estimation in subsequent years.

III. The Supervised Refinement: From Continuous Regression to Adaptive Classification (2018–2023)

The period from 2014 to 2017 consolidated the first proof-of-concept study of MDE with CNNs, and also recognized pixel-wise regression as a limitation of the technique. There was a dawning realization that regressing directly to depth as an instance of pure L_2 loss (Euclidean losses fix) is not the best. These losses often lead to slow convergence and mean-seeking prediction, where the smooth of object boundaries is fitted around as averaged error [27]. It is during this phase of research that we have seen three dominant paradigm shifts: from regression to ordinal classification, the re-emergence of geometric constraints and the use of frequency-domain features to offset against loss of spatial information.

A. The Discretization Shift: Ordinal Regression and Adaptive Bins

The major challenge for MDE is the wide range of depth measurements, which lead to high error rates both far and near. Fu et al. [27] address this by reformulating depth estimation from continuous regression to ordinal classification. More related, their Deep Ordinal Regression Network (DORN) claimed that it is easier for a network to classify a pixel into a depth interval (from 10m–12m) than regress an exact floating-point value. SID: A Spacing-Increasing Discretization (SID) strategy is proposed here as preliminarily SID quantizes the depth range into binned bins with narrow width for near regions and wide bins for far areas which corresponds to logarithmically human perception of space. They employed ordinal regression loss which only penalizes wrong orderings of depth labels rather than their absolute values. In semantic segmentation tasks, DORN achieves state-of-the-art results on the KITTI benchmark [25] using dilated convolutions inspired by Atrous Spatial Pyramid Pooling (ASPP) to maintain high-resolution features with full receptive field capacity.

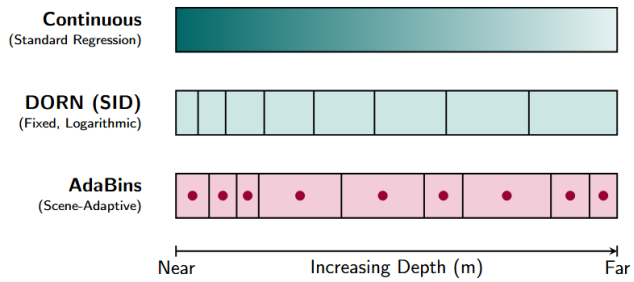


Fig. 3. Conceptual comparison of depth discretization strategies. Continuous regression maps to infinite floating points, DORN utilizes fixed spacing-increasing bins, and AdaBins dynamically clusters bin centers based on global scene context

While good, a "quantization error" ceiling was imposed by the fixed, pre-defined bins used to predict DORN. Bhat et al. AdaBins [28] vowed to solve this with a premise that depth bins ought to be continuously adapting rather than fixed. Exclusive to indoor and outdoor driving environments, the researchers illustrated that indoor scenes need different depth bin distributions to that of outdoor driving environments. AdaBins proposed a transformer-based "Mini-ViT" block that learned to use global content for images in order to find adaptive bin center locations per image. This hybrid "classification-regression" architecture enabled the model to retain the resilience of classification while recovering the

fine-grained accuracy of regression [28], motivated by dataset mixing techniques for strong priors [29].

Li et al. BinsFormer [30] built on adaptive paradigm. Motivated by the fact that previously proposed AdaBins regarded probability distribution regularization and bin value creation as distinct processes, BinsFormer used a transformer decoder to perform set-to-set reasoning for enriching joint interactions between each estimated bins with an arbitrary granularity and coarse depth information. This led to a finer granularity and achieved state-of-the-art performance on challenging indoor datasets aligned with the NYU Depth V2 database [30].

B. Geometric Guidance and Contextual Feature Fusion

With the progress in deep learning methods, researchers returned to geometric priors and incorporated easier planar assumptions inspired by those first proposed to the vision community with Make3D [6, 12]. Inspired by this idea, Lee et al. [31] proposed a new framework called From Big to Small (BTS) which brought forth Local Planar Guidance (LPG). This work demonstrated that many of the prevalent upsampling techniques, such as bilinear interpolation, often lead to distortions in surface geometry, due to not preserving the true 3D structure.

Rather than utilizing conventional up sampling techniques, the BTS network predicted local 3D plane equation coefficients per low-resolution pixel. These planes are then employed to guide the high-resolution up sampling, explicitly enforcing geometric consistency and alleviating common artefacts in flat surface, such as roads and walls [31]. To maximize feature reuse, the researchers adopt DenseNet [32] as their backbone, which enables geometric guidance signals to propagate from the bottleneck stage to the final output.

Abdulwahab et al. In a deep fusion network that learned contextual features, [33] showed the significance of incorporating contextual information. They demonstrated that loss of depth discontinuities often occurs because networks value the dominant "content" (texture and color) over "context" (spatial relations). Their model explicitly fuses multi-scale content features and context aware representations which helps structure depth maps better in low texture areas which is one of the traditional weaknesses of pure CNN architectures [34].

C. Frequency Domain Analysis: Mitigating Spatial Loss

CNN-based encoders such as ResNet and DenseNet require excessive downsampling to increase their receptive field, resulting in the irreversible loss of high-frequency spatial information. While dilated convolutions [35] address this, they require high computational power. Paul et al. [36] introduced a novel solution in their Nested Wavelet-Net (NDWTN), which integrates Discrete Wavelet Transform (DWT) directly with the CNN encoder system.

The DWT decomposes the feature map into four frequency sub-bands (LL, LH, HL, and HH), whereas standard pooling methods discard this information. Paul et al. implemented these coefficients in a nested architecture that enabled the network to downsample while maintaining high-frequency edge details. This “lossless” downsampling enabled the decoder to reconstruct fine details, such as tree branches and distant poles, with fidelity exceeding traditional U-Net architectural designs [36, 37]. The use of spectral analysis in deep learning has grown in recent years to address spectral bias in standard convolutional filters [38].

D. Critical Summary of The Supervised Refinement Era

The period from 2018 to 2023 successfully addressed the smoothing artifacts inherent to continuous regression, but the solutions introduced new limitations that constrained deployment capability:

1) *DORN (Fu et al.) [27]:*
   While Spacing-Increasing Discretization (SID) eliminated the mean-seeking blurring of earlier models, it introduced an insurmountable “quantization error.” The model was mathematically incapable of predicting depth values that existed between its predefined bins, establishing a rigid ceiling on metric precision.
2) *AdaBins [28] and BinsFormer [30]:*
   While scene adaptive binning helped resolve the issue of quantization in DORN, it did so by adding computational costs associated with using Transformer based global attention for grouping bins. Moreover, these mechanisms proved extremely brittle under domain shift and overfitted to the scale at which they were trained.
3) *BTS (Lee et al.) [31]:*
   Although Local Planar Guidance was capable of smoothing out flat areas such as walls and roads, it fell apart whenever there was complex geometry involved. The network often had problems with non-planar shapes and would force them to fit within a flat box shape.
4) *NDWTN (Paul et al.) [36]:*
   Despite Discrete Wavelet Transform managed to preserve the information related to edges, the mathematical strictness of Spectral Analysis meant that it was very sensitive to noise, rain and optical distortions and struggled to distinguish between real texture and camera artifacts.

In summary, the supervised refinement era traded continuous spatial blurring for discrete quantization and complex mathematical priors. While benchmark accuracy significantly increased, models grew exponentially more computationally expensive and remained fundamentally domain locked. They were incapable of universally generalizing metric scale without re-training, highlighting the necessity for the massive generative priors that would subsequently define the Foundation Era.

IV. The Self-Supervised Revolution: Learning Geometry from Motion (2017–2023)

But as light detection and ranging (LiDAR) data is sparse, expensive, and largely absent in unstructured environments [25], supervised learning reached a practical limit in this problem domain in 2017. While Godard et al. [11] used a stereo pair to show that face self-supervision works, scaling with synchronised dual-camera systems is limited. Thus, the research shifted to learning joint depth and ego-motion from monocular video sequences alone, effectively reframing MDE as a Structure-from-Motion (SfM) problem

in which the network was required to deduce 3D structures that explain pixel displacements between neighbouring video frames.

A. The Monocular Structure-from-Motion Framework

Zhou et al. made the first breakthrough in self-supervised learning for videos. They generalised to learn to visualise the associated features and predict depth images from [39] using their SfMLearner framework, as shown in Fig. 4. They propose a coupled two-way learning system involving DepthNet predicting per-pixel depth in the target frame and PoseNet estimating the 6-DoF camera motion between corresponding pixels in the target and source frames. The training signal was based solely on photometric consistency: if the predicted depth and pose are correct, warping a source frame to the target frame should produce an identical image. Zhou's first approach, while revolutionary in a literal sense, produced significant artefacts in non-Lambertian surfaces and low-texture regions [40]. In addition, the depth predicted from monocular video is correct up to a scale factor that often drifts over long sequences [41].

Godard et al. [44] suggested two mechanisms in Monodepth2 to tackle the common failures of the standard photometric loss. To avoid "double vision" artifacts around object boundaries due to occlusions, they first proposed a minimum reprojection loss defined as $L_p = min_{t'} pe\ (I_t, I_{t'\to t})$. This selects the best-matching source pixel instead of averaging. Second, they proposed an auto-masking method to remove static pixels between frames, which is useful to deal with objects moving at camera speed that generate artificial 'infinite depth' holes. These solutions set a good baseline for following self-supervised depth estimation frameworks.

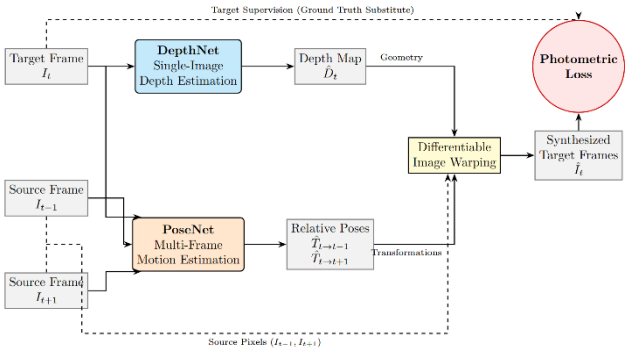


Fig. 4. The foundational Structure-from-Motion (SfM) framework for self-supervised monocular depth estimation. The network utilizes a sequence of frames (I_{t-1}, I_t, I_{t+1}). Depth and multi-frame ego-motion are predicted independently and coupled via differentiable view synthesis.

B. Architectural Innovations: Preserving Spatial Integrity

Improving loss functions alone did not resolve the fundamental limitation caused by standard downsampling operations (max-pooling and strided convolution), which eliminate spatial data irretrievably during upsampling. Guizilini et al. [42] determined that this information loss remains acceptable for classification but leads to disastrous effects in depth estimation tasks.

In PackNet, they introduced a novel 3D Packing and Unpacking block that uses "space-to-depth" operations to convert spatial dimensions into the channel dimension while maintaining complete pixel information. 3D convolution then processes packed features, enabling the extraction of three-dimensional features until they are "unpacked" at high resolution. This change enabled PackNet to generate significantly clearer images of thin objects, such as traffic signs and poles, than standard ResNet encoders without using more powerful backbones [42].

Masoumian et al. [43] studied Graph Convolutional Networks (GCNs) in GCNDepth, hypothesizing that traditional CNNs restrict their ability to model irregular objects and long-range dependencies through the use of predetermined square receptive fields. Their model constructed a graph representation of image features that enabled information flow between non-adjacent pixels sharing similar semantic attributes. Graph-based reasoning showed strong results for void-filling in textureless areas where standard CNNs create noisy artifacts [43, 44].

C. The Temporal and Dynamic Challenge

Early self-supervised methods by Zhou [39] and Godard [45] follow the static world assumption, which states that all scene elements remain fixed and that pixel movement results from camera motion. Dynamic objects including moving cars and pedestrians violate this assumption, creating substantial photometric loss gradients that mislead the network [46]. We achieve this constraint mathematically by applying a binary auto-mask $\mu = [min_{t'} pe\ (I_t, I_{t'\to t}) < min_{t'} pe\ (I_t, I_{t'})]$, to the photometric loss. This condition forces the network to learn only from static pixels, since it applies the loss if and only if the estimated camera motion yields a better photometric match than when the object is assumed static.

Zhou et al. [47] developed Manydepth2, which builds upon the multi-frame ManyDepth architecture from Watson et al. [48]. They added a motion-aware component using optical flow to detect moving regions, creating a pseudo-static reference frame that excludes high-motion regions from cost-volume calculations. This enables accurate depth estimation in fast-paced urban environments where standard self-supervised methods fail to handle moving vehicles [47].

Kopf et al. Reference [49] combined learning-based predictors with geometric optimization methods and developed the Robust Consistent Video Depth (RCVD) system. Instead of combining direct geometric optimization into the training pipeline, their efficient test-time optimization system circumvents the computational overhead of recurrent processing by using a separate network. By capitalizing on a method that collects per-frame predictions, including errors, and fuses them into a stable 3D form free of flickering, the authors aligned the low-frequency scale between frames using flexible deformation splines. At the same time, high-frequency details were filled in using geometry-aware filtering.

D. Semantic Guidance and Joint Learning: Masking Dynamic Domains

Although optical flow and heuristic auto-masking were handy solutions for the dynamic object problem, they fail on complex object trajectories or lead to cascaded failures when

the flow network degrades. To tap into this, an alternative paradigm that has achieved enormous success was key: explicit semantic segmentation priors as a guide for geometric learning.

Casser et al. This approach was first pioneered in the work Struct2Depth [46], which utilizes an off-the-shelf instance segmentation network (Mask R-CNN) to identify likely moving objects, such as vehicles, and transform their individual ego-motions, decoupled from the camera pose. Based on this, Klingner and colleagues proposed SGDepth (Semantic Guidance Depth) [50], which aimed to make the network semantically aware of moving objects. Field 1: SGDepth uses a semantic segmentation network to mark dynamic classes (vehicles, pedestrians) explicitly. These moving classes were masked during training to prevent the network from learning infinitely degenerate geometries.

Luo et al. EPC++ (Every Pixel Counts++) [51] stressed the importance of scene understanding holistically. Rather than considering depth, optical flow and moving object segmentation as independent problems, the work presented EPC++ as a joint learning task. By setting strict geometric constraints on the model, it was able to solve the problem jointly to predict the 3D camera pose, holistic depth and pixel-level motion segmentation.

In addition, Guizilini et al. "Semantics" is not just employed for the purpose of masking rather, it is also used for structural refinements [52]. By using adaptive convolution on the pixels based on the semantically guided features to make sure that depth discontinuities would always coincide with the objects boundaries, they merged the constraints of classical CRF models with the computational efficiencies of self-supervised learning models.

E. Critical Summary of The Self-Supervised Revolution

Self-supervision (2017-2023) gave birth to an entirely new framework where expensive LiDAR and stereo-rig ground truth was no longer required. The entire framework, however was flawed because of its underlying assumption that photometric consistency constraint was the holy grail a too extreme constraint:

1) *SfMLearner [39]:*
   Static World & Lambertian Reflection constraint had imposed mathematical constraints to the model. It struggled with reflective surface and texture less corridors, while suffering from irresolvable drift in scale.
2) *Monodepth2 [45]:*
   Although Minimum Reprojection Loss was a promising step towards handling occlusion, the automatic masking was a clumsy heuristic. While hiding away parallel movement it was a complete failure in the case of independent dynamic movement.
3) *PackNet [42] & GCNDepth [43]:*
   Maintaining spatial high frequency information caused severe computational bottleneck. The 3D-convolutions and generation of a dense graph caused memory overflow and extra latency which made the model far too complex to run on edge compute devices.
4) *Manydepth2 [47] & RCVD [49]:*
   Temporary fixes to handle static object movement and temporal flickering issues brought the system out of the real time boundary. Reliance on optical flow also formed a dangerous cascade failure path, while test time spline optimization consumed a lot of processing time and hence not useful for autonomous vehicles.
5) *Semantic Guidance (SGDepth, EPC++) [50, 51]:*
   The integration of semantic information made the structured understanding of dynamic objects very compelling, but the fundamental price for doing this was corrupting the theoretically unsupervised nature of the paradigm. These models depended on a pre-trained teacher network that could provide semantically annotated ground truth. They strongly embodied domain bias (Did not function in rural areas if trained on urban areas) and brought with them a dramatic parameter explosion causing extreme inference delays.

In summary, self-supervised learning did prove that geometry can be learned from motion, what it failed to acknowledge was the chaos of real life, a problem that the global semantic cutoff, using a sophisticated top-down learning methodology, overcame marking the arrival of the Visual Transformers and the advent of generative priors.

V. The Global Context Era: Vision Transformers and Hybrid Architectures (2021–2024)

In 2021, the theoretical constraints of CNNs for dense prediction tasks became apparent. Though CNNs capture very local features (textures and edges), their depth allows the receptive field to grow slowly. As such, deep CNNs [35] inherently lack the capacity to model long-range dependencies required to resolve scale ambiguity in monocular images. This need for a solution leads to an entirely new architecture, giving rise to Vision Transformers (ViTs) [50], which naturally can process information through a global receptive field at all layers of their operation. This is a delicate trade-off between global reasoning and computational efficiency for real-world applications.

A. The Transformer Breakthrough: Global Reasoning

Ranftl et al. [51] introduced the first application of Transformers to depth estimation in Vision Transformers for Dense Prediction (DPT). DPT maintains its encoder at a stable high resolution while U-Net architectures use extreme downsampling to increase their field of view. DPT applies self-attention mechanisms to image patches, which function as tokens analogous to words in NLP [52], establishing relationships between all pixels in the image.

The model uses global connectivity to deduce depth from both nearby textures and segment-based semantic knowledge such as "sky is always far" and "floors are continuous." DPT achieved a 28% performance boost compared with state-of-the-art CNNs, demonstrating that global coherence is an essential component for MDE [51]. However, standard self-attention incurs quadratic computational cost:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

with complexity $O(N^2)$ based on token count [53], making standard ViTs too slow for high-resolution deployment.

B. The Lightweight Counter-Movement: Efficient Edge Deployment

The operational expenses of ViTs created a deployment gap between research and practical implementation. Researchers have investigated Lightweight Hybrid Architectures designed to function on mobile devices and edge computing environments, as illustrated in Fig. 5. The architecture uses dual-path networks where a light CNN is used for efficient processing of local features and a simpler Transformer for global contexts. Li and Wei [58] tackled the speed limitation issue with MobileDepth using a Dilated Self-Attention Block (DSAB), where attention computation is restricted to a sparse grid structure in order to reduce token consumption but maintain global coverage. The network can run in real-time on mobile CPUs along with LGFE module, unlike DPT which requires higher parameter numbers [51].

LEDepth was invented by Zhang et al. [59], whose DCT module cuts down the conventional parameter requirements of 80% by including convolutional inductive bias to the transformer blocks. The cross-attention mechanism implemented by the CAFI module facilitates information exchange between the local CNN features and global Transformer features to avoid "feature drift" during hybrid model training [59].

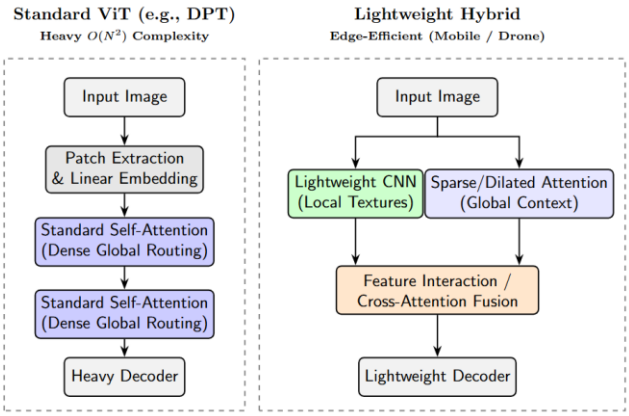


Fig. 5. Comparison between Vision Transformer architectures in regard to normal ViTs vs. edge-efficient hybrid networks. While normal ViTs have quadratic spatial complexity, edge-efficient hybrid networks take advantage of a dual-branch architecture that extracts local features using convolutions and global features using simplified attention.

Sui et al. [60] and Wang et al. [61] conducted studies on fusion-improved architectural designs. Sui et al. used "transposed attention" as a self-attention method that achieves lower complexity through channel-based rather than spatial based operations [60]. Wang et al. implemented a dual encoder system combining a CNN branch for texture analysis with a Transformer branch for structural analysis through a Feature Interaction Module (FIM) that dynamically weights local and global information [61].

C. Bridging the Metric Gap: Zero-Shot Transfer

A persistent problem in MDE is the division between "relative depth," which generalizes well but lacks scale information, and "metric depth," which provides correct distance measurements but fails to generalize. Indoor models trained on NYU Depth V2 fail to perform well in outdoor environments because KITTI testing requires handling entirely different value ranges.

Bhat et al. [62] solved this with ZoeDepth, the first zero-shot transfer framework, illustrated in Fig. 6. ZoeDepth uses a pre-trained relative depth backbone including MiDaS [29] to establish domain-specific "metric heads." The core innovation is a Metric Bins Module that automatically adjusts depth discretization bins according to different scene types. A latent classifier directs input images toward the appropriate metric head for either indoor or outdoor environments, enabling a single model to generate precise metric depth measurements for both settings without retraining - a major step toward universal depth estimation [62].

D. Structured Prediction and Panoramic Domains

Yuan et al. [63] proposed NeW CRFs (Neural Window Fully-connected CRFs). Unlike traditional CRFs, which require high computational power [23], NeW CRFs use local window optimization with a neural potential function. The network learns to minimize the CRF energy function within local windows:

$$E(d) = \sum_i \psi_u(d_i) + \sum_{i,j \in N} \psi_p(d_i, d_j) \qquad (3)$$

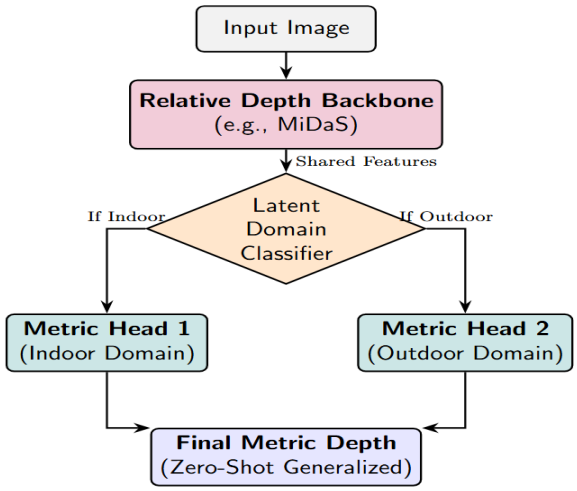


Fig. 6. The ZoeDepth zero-shot transfer architecture. A shared relative depth backbone extracts global features, which are evaluated by a latent domain classifier. The classifier dynamically routes features to the appropriate domain-specific metric head (indoor vs. outdoor), enabling universal metric scale recovery without retraining.

Here, ψ_u represents the unary potential (initial depth prediction for pixel i) and ψ_p represents the pairwise potential (neural penalty for depth differences between neighboring pixels i and j in window N). Integration of this optimization layer with a Swin Transformer backbone decoder resulted in improved depth discontinuities, restoring probabilistic graphical models to the current state-of-the-art loop [63].

Bai et al. [64] tackled 360-degree vision problems using GLPanoDepth. Equirectangular projection causes severe distortion that prevents standard CNNs from processing panoramic images. Their Cubemap Vision Transformer (CViT) transforms spherical images into cube format, enabling the Transformer to analyze each face while establishing cross-face connections, producing distortion-free depth estimation for all viewing angles [64].

E. Critical Summary of The Global Context Era

The integration of Vision Transformers (2021-2024) decisively conquered the spatial limitations of convolutional receptive fields, but the era was defined by a desperate struggle against computational reality:

1) *DPT [51]:*
   While pure Transformers achieved unprecedented geometric coherence, their $O(N^2)$ quadratic complexity rendered them practically undeployable for high-resolution, real-time edge processing.
2) *Lightweight Hybrids [58], [59]:*
   Attempts to circumvent the quadratic bottleneck via sparse grids, dilated attention, or heavy cross-attention compression successfully achieved real-time inference. However, these models inherently risk missing high-frequency, fine-grained geometric details that fall between their sparse sampling patterns.
3) *ZoeDepth [62]:*
   The zero-shot metric transfer framework elegantly bypassed domain-locking, but its core mechanism a hard-routing domain classifier which acts as a single point of failure. It lacks the continuous flexibility to interpret ambiguous or hybridized environments.
4) *NeW CRFs [63] and GLPanoDepth [64]:*
   Specialized adaptations for structured prediction and panoramas improved edge cases but introduced localized artifacts, such as window-boundary discontinuities and cubemap projection seams, respectively.

Overall, while Transformers proved that global semantic context is required to resolve depth scale accurately, manipulating the architecture to fit edge hardware exposed its brittleness. Achieving true, robust universality without these compromises would require the massive scale and unyielding priors of the subsequent Generative Foundation Era.

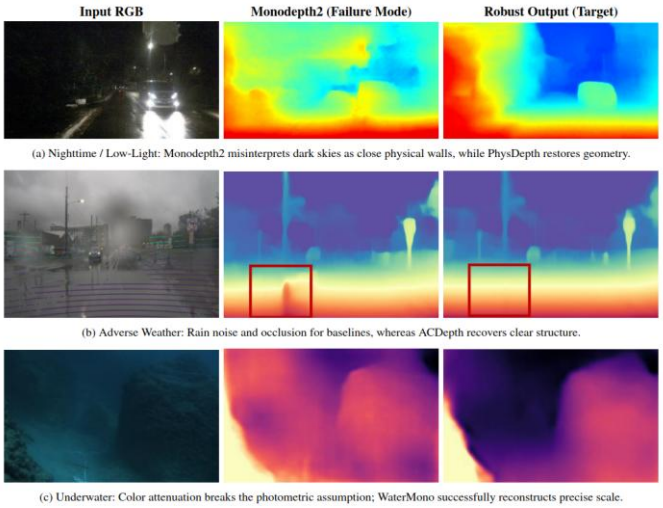


Fig. 7. The Reliability Gap in Extreme Environments. Standard architectures like Monodepth2 [45] trained on clear daytime datasets experience catastrophic failure modes when confronted with domain shifts. Robust architecture utilizes physical priors and generative augmentation to restore geometric accuracy. Nighttime examples reproduced from [67], adverse weather examples from [68], and underwater examples from [66].

VI. The Reliability Gap: Robustness in Adverse and Specialized Environments (2021–2026)

The modeling of the global context properties by Transformer-based architectures led to some very good performance benchmarks. However, there was a serious flaw left: reliance on the property of photometric consistency. Photometric self-supervision is based on the assumption that the appearance of a point does not change from the viewpoint variation. The photometric consistency assumption fails for several real-life situations such as low-light conditions resulting in maximum sensor noise, rainy/foggy conditions blocking geometric visibility, or underwater conditions where light propagates differently than in air [65, 66]. As shown in Fig. 7, the trend of current research since 2021 focuses on moving away from creating universal systems architecture to creating specialized frameworks with improved performance using physical phenomena and generative data augmentation that restores lost visual information.

A. Illuminating the Dark: Nighttime and Low-Light Perception

The problems that depth estimation faces at night time are primarily the signal-to-noise ratio reduction of the dark region and erratic lighting generated by streetlights and vehicle headlights. The reason why the conventional model that is trained based on daylight information does not work in this case is due to the misinterpretation of sky regions as physical planes [69].

In their study titled "Regularizing Nighttime Weirdness," Wang et al. [69] have shown that lack of visibility was not the main cause for failure. They used Priors-Based Regularization technique that utilizes the teacher network trained based on daytime data to stop student networks being trained on night-time data from learning geometrically wrong depths caused by photometric gradient noise.

The Mapping-Consistent Image Enhancement technique by Huang et al. [70] was developed as the answer to static regularization. They demonstrated that spatial texture disappears during nighttime while temporal motion cues maintain their strength. An adversarial branch in DASP implements a Spatiotemporal Priors Learning Block (SPLB) that employs orthogonal differencing to obtain motion-related time-dependent changes, requiring the encoder to depend on temporal consistency by separating motion features from illumination artifacts.

Peng et al. [67] developed PhysDepth, which introduces atmospheric scattering models as embedded components within its network. They showed that prediction error depends on the level of atmospheric haze and darkness. PhysDepth uses a Physical Prior Module (PPM) implementing Rayleigh Scattering theory [71] to enable the network to reverse the degradation process. This is grounded in the classic atmospheric scattering model:

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (4)$$

where J(x) is the clear scene radiance, t(x) is the medium transmission map (heavily dependent on depth), and A is the global atmospheric light. By explicitly modeling t(x), PhysDepth forces the network to learn the underlying geometry rather than fitting to noise, achieving state-of-the-art results by treating darkness as a physical light transformation rather than a data deficiency.

B. Weathering the Storm: Generative Augmentation for Rain and Fog

Rain and fog produce high-frequency noise from raindrops and low-frequency occlusion through fog, confusing depth estimation systems. Saunders et al. [65] argued in Let's Talk About The Weather that collecting real-world ground truth for every weather condition is unfeasible. They used Generative Adversarial Networks (GANs) to create synthetic weather conditions applied to sunny KITTI images, and created a Pseudo-Supervised Loss that requires the model to generate identical depth predictions for both clean and augmented images:

$$\mathcal{L}_{\text{pseudo}} = \left\| D(I_{\text{clear}}) - D(I_{\text{aug}}) \right\|_1 \quad (5)$$

This trains the model to treat weather artifacts as temporary disturbances. However, GAN-based generation lacks authentic physical appearance. Jiang et al. [68] advanced this paradigm in Always Clear Depth (ACDepth) by utilizing Denoising Diffusion Probabilistic Models (DDPMs) [72]. They developed a one-step diffusion process using LoRA adapters to create high-fidelity weather degradation effects. ACDepth uses Multi-Granularity Knowledge Distillation to transfer both depth map information and internal feature correlations from a clear-weather teacher to an adverse-weather student, enabling the student to extract structure from degraded inputs rather than memorizing the teacher's output.

C. Beneath the Surface: Underwater and Construction Domains

The "Jaffe-McGlamery" optical model shows that underwater environments create the most severe domain shift because red light is absorbed within a few meters and particles produce strong backscatter effects [73]. Ding et al. [66] proposed WaterMono, which showed that marine snow particles and dynamic lighting caustics function as anomalies. Their Teacher-Guided Anomaly Masking method uses a pre-trained terrestrial model to detect high-uncertainty areas that are masked from photometric loss assessment.

Wang and Ye [74] studied the spectral domain in Integrating Underwater Optical Imaging Priors. Their Underwater Light Priors Processing Module (ULPM) models attenuation coefficients across color channels, training with physical priors through a Transformer backbone to correct depth distortion that affects standard RGB-based models.

Shen et al. [75] addressed the unstructured chaos of industrial environments in LLD-DEPTH. Construction sites combine low light conditions, high dynamic range, and fast-moving machinery. The researchers used an enhanced version of ForkGAN for unsupervised day-to-night translation, jointly optimizing depth estimation, ego-motion tracking, and optical flow measurement while using flow consistency to identify dynamic objects that would otherwise create geometric holes.

D. Critical Summary of Robustness Era

The Robustness Era (2021-2026) dealt with the catastrophic failure modes of standard MDE models in specialized domains, demonstrating that networks need to take physical optics into account. However, the use of mathematical priors and synthetic degradation created distinct operational and theoretical bottlenecks:

1. *Physical Prior Rigidity (PhysDepth [67] & ULPM [74]):*
   The key is to embed formulas such as the Rayleigh scattering or the Jaffe-McGlamery model which gives strong geometric constraints. The problem is that these equations essentially assume homogeneous mediums. They break down spectacularly under localized, heterogeneous perturbations, such as high-beam headlights penetrating fog, or highly stratified underwater thermoclines.
2. *The Synthetic Domain Gap (ACDepth [68] & Saunders et al. [65]):*
   Mathematically, weather synthesis with DDPMs or GANs bridges the data deficiency, but synthetic degradation seldom replicates the chaotic physics of the real world. In many cases these models do not include secondary physical effects such as wind shield wiper occlusion, wet-surface specular reflections or active sensor blinding.
3. *Teacher-Student Cascade Failures (RNW [69] & WaterMono [66]):*
   The use of daytime or terrestrial "teacher" models to guide nighttime or underwater "students" inherently upper-bounds the capability of the student. When the terrestrial teacher misunderstands a complex structure, the robust model propagates the permanent geometric error, creating a dangerous cascade of blind spots.
4. *Unsupervised Translation Limits (LLD-DEPTH [75] & DASP [70]):*
   Day-to-night GAN translation or temporal differencing for domain adaptation is highly prone to hallucination artifacts. Moreover, masking moving objects with motion cues may cause the network to permanently ignore slow-moving hazards, treating them as infinite-depth "holes" in the environment.

The quest for robustness finally proved that physical reality cannot be ignored by data-driven networks. These

specialized frameworks were successful in adapting MDE to extreme environments but they did so at the cost of fragmenting the field into siloed environment-specific solutions. The other big missing piece from this era is the lack of unification: we gained environmental robustness, but at the expense of universality, requiring developers to deploy different networks for weather, lighting, or medium, instead of using one network that is domain-agnostic.

VII. The Foundation Era: Generative Priors and Metric Universality (2024–2026)

MDE has developed through three stages - handcrafted features, supervised regression, and self-supervised learning - before reaching its fourth and probable final paradigm shift when Geometric Foundation Models emerged in 2024. Inspired by the success of Large Language Models (LLMs), researchers began investigating depth estimation solutions that benefit from processing larger datasets with generative priors. The current era advances from "estimating" depth to "generating" geometry, leveraging world knowledge that resides in extensive pre-trained models to address MDE's scale ambiguity and generalization challenges that have persisted for two decades.

A. From Regression to Generative Diffusion

The MDE function was defined as a pixel-by-pixel regression problem for over a decade. Ke et al. [76] challenged this formulation with Marigold, arguing that regression models inherently create blurry depth maps for uncertain areas.

Marigold uses the Stable Diffusion latent space to create depth maps, training a denoising U-Net to remove Gaussian noise from depth maps conditioned on input images. The model established object shape knowledge and understanding of occlusion boundaries by training on billions of images. Marigold proved that depth should be treated as a generative process, enabling detailed reconstruction of fine elements such as hair or fur that regression models cannot reproduce, although multiple denoising steps increase processing time [77].

Mathematically, the diffusion model is trained to predict the noise ε added to the latent depth map at timestep t:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2] \qquad (6)$$

where ε_θ is the denoising U-Net predicting the noise conditioned on the input RGB image c.

B. The Data-Centric Paradigm: Scaling to General Intelligence

Yang et al. [78] investigated the data aspect of MDE in Depth Anything V2. They proved that real-world labels created noise issues restricting MDE development. They recommended developing student models trained mainly on synthetic data combined with extensive pseudo-labeled image data, rather than noisy real-world datasets such as KITTI. A teacher model with nearly one billion parameters produced pseudo-labels from 62 million unlabeled real images using five high-fidelity photorealistic synthetic datasets including Hypersim and Virtual KITTI 2 [79].

The extensive Teacher-Student loop enabled the model to develop strong representations for zero-shot learning across diverse environments. Depth Anything V2 achieved superior performance compared with models trained only on NYUv2 data because it had access to all datasets during testing, introducing the first "general purpose" depth estimation technology.

C. Solving the Metric Ambiguity

Relative depth models including MiDaS and Depth Anything cannot provide depth measurements corresponding to actual metric distances. Piccinelli et al. [80] developed UniDepth, which introduces a universal metric estimator.

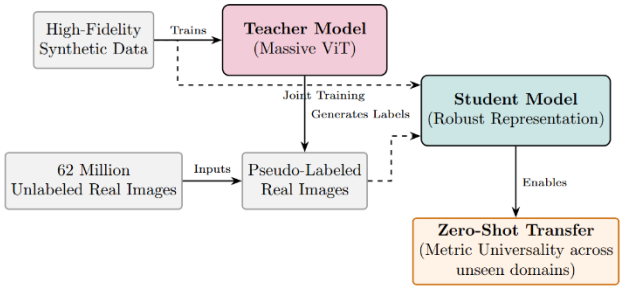


Fig. 8. The data-centric Foundation Era pipeline (Depth Anything V2). A massive Teacher model is trained purely on five high-fidelity synthetic datasets including Hypersim [79] and Virtual KITTI 2 to avoid real world sensor noise. It then generates pseudo-labels for millions of real images. A Student model is jointly trained on both data sources, achieving zero-shot metric universality [78].

They discovered that metric ambiguity occurs because actual camera intrinsic parameters remain unidentified. UniDepth uses a self-promptable camera system to create a pseudo-spherical camera parameter model, achieving zero-shot metric precision across ten distinct datasets by separating depth information from projection scale [80].

Hu et al. [81] developed Metric3D v2 as a solution to the "focal length gap." They discovered that different camera zooms levels produce different appearances for the same object, creating confusion for standard models. Metric3D v2 transforms all input images into a Canonical Camera Space, normalizing the focal length before estimation:

$$d_{\text{cano}} = d_{\text{pred}} \frac{f_{\text{cano}}}{f_{\text{in}}} \qquad (7)$$

The Canonical Transformation Module ensures objects of the same physical size project to the same pixel dimensions regardless of camera. A joint depth-normal optimization objective demonstrated that surface normal prediction helps to improve metric depth prediction results by reducing geometric inconsistencies [81, 82].

D. Test-Time Adaptation and Panoramic Expansion

Re-Depth Anything is a self-supervised refinement framework developed by Bhattarai and Rhodin [83] that

operates during inference. A classic example to illustrate this is that if there exists a domain gap between texture and shape, by using foundation models it's difficult to correct it. This model utilizes a re-lighting objective that modifies predicted depth maps until the 3D surface recreates input image shading, giving it no ground truth as a dependency as it relies exclusively on light transport physics [84].

Cao et al. [85] introduced PanDA (Panoramic Depth Anything) to address 360-degree domain challenges. Standard foundation cannot perform well in panoramas due to polar distortions. PanDA uses Möbius Spatial Augmentation to simulate various viewing angles and spherical distortions. The student model learns features that remain invariant to equirectangular projection warping, enabling foundation models to function in Virtual Reality (VR) applications without labeled panoramic datasets [85, 86].

Monocular depth estimation has evolved from geometric puzzles to semantic understanding tasks between the 2005 probabilistic MRFs and the 2026 billion-parameter generative foundation models. The field has effectively solved the relative depth problem, and current research approaches its final frontiers: absolute metric universality, real-time generative inference, and physical reliability in the most extreme environments.

E. Critical Summary of The Foundation Era

The Foundation Era (2024-2026) represents the most significant leap in MDE capability to date, functionally solving the relative generalization problem. However, the paradigm is heavily burdened by the brute-force nature of its solutions:

1) *Marigold (Ke et al.) [76]:*
   While latent diffusion perfectly resolves structural blurring and hallucinates fine details, the multi-step iterative denoising process is extraordinarily slow. It trades real-time computational viability for geometric aesthetics, disqualifying it from closed-loop autonomous control systems.
2) *Depth Anything V2 (Yang et al.) [78]:*
   The successful data-centric filter of physical sensor noise via synthetic pseudo-labeling is enforced but creates immense challenges in terms of the scale and barrier to entry on the training compute and environment footprint.
3) *Metric3D v2 [81] & UniDepth [80]:*
   Canonical space transformations sidestep the issue of linking structure to scale, but are still brittle to out-of-distribution camera intrinsics, non-standard aspect ratios and post-capture image cropping which damage focal length ratios.
4) *Re-Depth Anything [83]:*
   The test-time shading refinement approach manages to get rid of the need for ground truth yet inherently needs a form of physical light transport which leads to the failure of depth estimation when encountering reflective surfaces such as mirrors or glass.

In addition to these architectural weaknesses, the practical costs incurred in the deployment of these foundation models pose an obvious bottleneck for real-world robotics. Frameworks with massive Vision Transformer (ViT-Large or ViT-Giant) backbones such as Depth Anything V2 [78] and Metric3D v2 [81] achieve unprecedented zero-shot accuracy but are fundamentally not viable for edge deployment. Often their inference needs require high-end desktop GPUs with large amounts of VRAM, making them unsuitable for the tight power, thermal and memory constraints of mobile robotics, drones or real-time embedded systems. The field is now in a painful split as researchers have to choose between the high-fidelity generalization of large foundation models and the real-time operational efficiency of heavily quantized specialized architectures.

Finally, Monocular Depth Estimation has developed from geometric puzzles to semantic understanding tasks between the 2005 probabilistic MRFs and the 2026 billion-parameter generative foundation models. The field has effectively solved the "relative" depth problem because current research now approaches its final frontiers, including absolute metric universality, real-time generative inference and physical reliability in the most extreme environments.

VIII. Datasets and Benchmarks for Monocular Depth Estimation

The historical progression of MDE is inextricably linked to the availability, scale, and complexity of empirical datasets. Because depth estimation relies heavily on extracting structural priors from 2D pixel intensities, the inductive bias learned by any given model is almost entirely dictated by the environment it was trained on. Over the past two decades, the community has evolved from relying on small, highly constrained synthetic captures to leveraging massive, multi-modal, and spatiotemporally diverse benchmarks.

1) Structured Driving and Outdoor Perception: KITTI and Cityscapes:

Since the debut of KITTI Vision Benchmark Suite [26], MDE has fundamentally entered a new age with deep learning. It comprises more than 93,000 dense depth maps captured with a Velodyne HDL-64E LiDAR sensor and walking stereo cameras in small to medium-sized German cities. KITTI became the final benchmark as its stereo pairs facilitated photometric self-supervision [11]. However, KITTI, which has a strong domain bias: it consists almost exclusively of static, well-lighted clear-weather images of highly structured roads with predictable vanishing points. Networks trained only on KITTI can overfit a "bottom-heavy" depth prior that causes catastrophic failures when deployed in non-driving scenarios.

2) Indoor Clutter and Structural Ambiguity: NYU Depth V2:

For the previously mentioned absence of indoor diversity, NYU Depth V2 dataset [16] is still a golden benchmark. The dataset features challenges that do not exist when driving outdoors: extreme occlusion of objects, absence of horizon lines, large surface areas that are textureless (such as blank walls or ceilings), and the specular reflections from glass or mirrors. It was captured with a Microsoft Kinect sensor in 464 diverse commercial and residential environments. The estimated depth range may be artificially bounded at 10 meters or so, as the problem of estimating

infinite skies is traded for resolving razor-thin boundaries between foreground furniture and background walls.

3) The Frontier of Reliability: Adverse and Specialized Domains:

As standard benchmarks were saturated, the literature revealed a glaring gap in reliability. Many highly optimized models for KITTI provide graceful degradation or collapse completely under optical duress. The nuScenes [83] and Oxford RobotCar datasets specifically feature night-time scenes and extreme light drops respectively. KITTI-C provides synthetically corrupted weather states (rain, fog, snow) mapped to ground truth for real-world weather anomalies. The FLSea dataset [87] addresses issues like excessive color loss, marine snow, and light refraction, which is common in a deep underwater setting. This makes it difficult for the MDE models to utilize traditional physics-based priors.

IX. Evaluation Metrics: Quantifying Geometric Accuracy

It is crucial to have a standardized model evaluation process (which is mathematically precise) to compare the results obtained in such models. In the field of MDE, model performance is evaluated on a global scale using two different types of error metrics: continuous and discrete, where the latter punishes outliers heavily. The standard evaluation suite consists of the following, given a ground-truth depth d_i, the network-predicted depth $\hat{d}_i$ and also N, which is the count of number of valid pixels that were evaluated:

Absolute Relative Error (Abs Rel):

$\text{AbsRel} = \frac{1}{N}\sum_i \frac{|\hat{d}_i - d_i|}{d_i}$ This measures the average relative error, scaling the penalty based on the object's actual distance. An absolute error of 1 meter is catastrophic for an object 2 meters away, but negligible for a building 50 meters away.

Squared Relative Error (Sq Rel):

$\text{SqRel} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|\left|\hat{d}_i - d_i\right|\right|^2}{d_i}$. By squaring the numerator, this metric severely amplifies the penalty for catastrophic localized estimation failures.

Root Mean Squared Error (RMSE):

$\text{RMSE} = \sqrt{\frac{1}{N}\sum_i \left(\hat{d}_i - d_i\right)^2}$. Evaluated in absolute meters, RMSE heavily penalizes large-scale deviations and is particularly sensitive to outliers at long ranges.

Threshold Accuracy (δ_k): Defines the percentage of predicted pixels $\max\left(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\right) = \delta < 1.25^k$ for $k \in \{1, 2, 3\}$. Rather than measuring average penalty, threshold accuracy measures the strict reliability of the prediction.

X. Quantitative Evaluation and Comparative Discussion

Analyzing the literature purely through qualitative architectural diagrams leaves a fragmented narrative. To map the trajectory of MDE, we interrogate quantitative benchmarks across distinct operational envelopes. We present five exhaustive matrices detailing the shift from early supervised regression to the current era of generative foundation models, highlighting the diverging branches of edge-efficiency and extreme-weather robustness.

A. The Baseline Progression on Structured Outdoor Environments

Table I shows the quantitative performance evolution of monocular depth estimation architectures on the KITTI dataset, mainly following the standard Eigen split. We split the benchmark according to the supervised and self-supervised learning paradigms to show the diverging developmental pathways of both approaches.

In the supervised case, results reveal a gradual improvement from early convolutional baselines, e.g. Eigen et al. [7], to very specific depth classification methods. The state-of-the-art performance during this period is achieved with the introduction of adaptive depth binning in 2021, via the AdaBins [28] architecture, bringing the Absolute Relative error (Abs Rel) to an impressive 0.058.

On the other hand, the self-supervised part tells the story of the fast closing of the performance gap without the reliance on expensive ground-truth LiDAR data. Key foundational models such as SfMLearner [39] and Monodepth [11], further stabilize this via Monodepth2 [45] which sets the baseline viability of the photometric consistency assumption. The boundaries of self-supervision have again been pushed even further by the researchers in 2020; there they followed up with novel design using new packing structures and semantic priors. For example, the semantically guided architectures of Guizilini et al. By using global and local depth maps, [50] and SGDepth [52] were able to push the state-of-the-art for the paradigm. This stride resulted in an Abs Rel of 0.100 in the PackNet-Sem [50] variant, which has kept a record low (visually) and shows that mature self-supervised models can be strongly competitive against early supervised benchmarks.

TABLE I
PERFORMANCE EVOLUTION ON THE KITTI DATASET. EVALUATED PRIMARILY ON THE EIGEN SPLIT, ILLUSTRATING THE QUANTITATIVE TRANSITION FROM EARLY CONVOLUTIONAL NEURAL NETWORKS TO TRANSFORMER-BASED HYBRID ARCHITECTURES.

Method	Year	Supervision	Backbone / Variant	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ1 ↑	δ2 ↑	δ3 ↑
Eigen et al. [6]	2014	Supervised	VGG-based (Fine Net)	0.203	1.548	6.307	0.282	0.702	0.890	0.958
DORN [26]	2018		ResNet-101	0.072	0.307	2.727	0.120	0.932	0.984	0.994
BTS [30]	2019		ResNeXt-101	0.059	0.245	2.756	0.096	0.956	0.993	0.998

AdaBins [27]	2021		EfficientNet-B5	**0.058**	**0.190**	**2.360**	**0.088**	**0.964**	**0.995**	**0.999**
DPT [51]	2021		DPT-Hybrid	0.062	-	2.573	0.092	0.959	0.995	0.999
Monodepth [10]	2017	Self-Supervised	ResNet-50 pp (CS+K)	0.148	1.344	5.927	0.247	0.803	0.922	0.964
SfMLearner [38]	2017		VGG-based (Mono)	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Struct2Depth [45]	2018		ResNet (M+R)	0.109	0.825	4.750	0.187	0.874	0.958	**0.983**
Monodepth2 [44]	2019		ResNet-18 (Mono M)	0.115	0.903	4.863	0.193	0.877	0.959	0.981
EPC++ [51]	2020		ResNet-50 (Mono)	0.141	1.029	5.350	0.216	0.816	0.941	0.976
PackNet [41]	2020		PackNet (3D Packing)	0.111	0.785	4.601	0.189	0.878	0.960	0.982
Guizilini et al. (semdepth)[50]	2020		PackNet-Sem	**0.100**	**0.761**	**4.270**	**0.175**	**0.902**	**0.965**	0.982
SGDepth [52]	2020		ResNet-18 (CS+K)	0.107	0.768	4.468	0.186	0.891	0.963	0.982

B. Navigating Indoor Clutter and Structural Ambiguity

The quantitative evaluation for the indoor NYUv2 dataset in Table II shows the development of architectural solutions necessary for sustaining spatial structure despite drastic occlusion and texture-less indoor clutter. This supervised learning block develops over time starting from early convolutional baselines, such as Eigen et al. [7], FCRN [18] to more complex metrics. The development of adaptive depth binning with vision transformers was definitely a breakthrough of 2021 and especially for 2022. Architectures such as AdaBins [28], NeW CRFs [63] made great steps to get closer to the edges. But BinsFormer [30], which uses a Swin-Large backbone, broke the 0.100 barrier and received the best-ever Abs Rel = 0.094.

Also, one cannot overlook the self-supervised indoor setting, which has some specific features. Photometric consistency assumptions tend to break when applied to blank walls and irregular camera rotations observed in indoor datasets. Recently, there have been great advances in this field, including P2Net [88] and MonoIndoor [89]. Hence, reaching the absolute minimum milestones for establishing the minimum feasibility of indoor self-supervision with an Abs Rel equal to 0.134 could be viewed as a significant step toward closing the historical gap with supervised networks based on conventional ResNet-18 backbones.

TABLE II
PERFORMANCE EVOLUTION ON THE INDOOR NYUV2 DATASET, TRACKING THE ABILITY OF MODELS TO PRESERVE SPATIAL INTEGRITY AMIDST SEVERE OCCLUSION AND TEXTURELESS INDOOR CLUTTER.

Method	Year	Supervision	Architecture / Variant	Abs Rel ↓	RMSE ↓	log10 ↓	δ1 ↑	δ2 ↑	δ3 ↑
Eigen et al. [7]	2014	Supervised	VGG-based (Fine Net)	0.214	0.877	0.087	0.611	0.887	0.971
FCRN (Laina) [18]	2016		ResNet-50 (Up-Proj)	0.127	0.573	0.055	0.811	0.953	0.988
BTS [30]	2019		DenseNet-161	0.110	0.392	0.047	0.885	0.978	0.994
AdaBins [28]	2021		EfficientNet-B5	0.103	0.364	0.044	0.903	0.984	0.997
DPT [51]	2021		DPT-Hybrid	0.110	0.357	0.045	0.904	0.988	0.998
BinsFormer [30]	2022		Swin-Large	**0.094**	**0.330**	**0.040**	**0.925**	0.989	0.997
NeW CRFs [63]	2022		Swin-Large	0.095	0.334	0.041	0.922	**0.992**	**0.998**
Abdulwahab et al. [33]	2023		HRNet-v2 (Autoencoder)	0.121	0.523	0.053	0.852	0.974	0.994
P²Net [88]	2020	Self-Supervised	ResNet-18	0.147	0.553	0.062	0.801	0.951	0.987
MonoIndoor [89]	2021		ResNet-18	0.134	0.526	-	0.823	0.958	0.989

C. The Reliability Gap: Survival in Extreme Environments

The extreme cases of domain shift discussed in Section V are summarized in Table III. MonoDepth2 [45] provides a "Reliability Gap" case study where daytime baselines function quite satisfactorily in clear weather, yet suffer a dramatic decrease in performance during loss of structural visibility (with an Abs Rel error of 1.185 in the unconstrained low-light setting of nuScenes-Night). The adverse condition-based architectures succeed in bridging this gap impressively. PhysDepth [67] exploits physics priors of light and physical scattering within the architecture itself, leading to improvements of the Abs Rel to 0.109 in RobotCar-Night and 0.130 in nuScenes-Rain and becoming the preferable framework for both nighttime and rainy environments. Finally, Table III shows the extreme difficulty of light attenuation in FLSea. In FLSea, the self-supervised WaterMono [66] performs extremely well with an Abs Rel of 0.099, demonstrating that anomaly masking can help overcome problems associated with marine snow and caustics. On the other hand, physical priors of light together

with dense geometry supervision (such as used in Wang & Ye [74], DGS [90]) keep on setting boundaries. Currently, the SoTA boundary in this case is set by Wang & Ye at 0.097.

TABLE III
THE ADVERSE DOMAIN MATRIX. HIGHLIGHTING THE PERFORMANCE OF ARCHITECTURES SPECIFICALLY ENGINEERED TO ENDURE THE PROFOUND DOMAIN SHIFTS INDUCED BY LOW-LIGHT, ADVERSARIAL WEATHER, AND COMPLEX UNDERWATER LIGHT ATTENUATION.

Method	Year	Supervision	Domain Focus	Evaluation Dataset	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ1 ↑	δ2 ↑	δ3 ↑
MonoDepth2 (Baseline) [45]	2019	Self-Sup.	Nighttime	RobotCar-Night	0.399	7.451	6.642	0.744	0.892	0.928
RNW [69]	2021				0.121	**0.520**	**2.902**	**0.879**	**0.969**	**0.990**
ACDepth [68]	2025				0.121	0.690	3.432	0.845	-	-
DASP [70]	2025				0.164	1.442	6.315	0.777	0.930	0.981
PhysDepth [67]	2026				**0.109**	-	3.450	0.863	-	-
MonoDepth2 (Baseline) [45]	2019			nuScenes-Night	1.185	42.3059	21.613	0.184	0.360	0.504
RNW [69]	2021				0.315	3.793	9.641	0.508	0.778	0.896
PhysDepth [67]	2026				**0.118**	-	**6.349**	**0.853**	-	-
DASP [70]	2025				0.276	**3.072**	8.819	0.584	**0.809**	**0.916**
PhysDepth [67]	2026		Adverse Weather	nuScenes-Rain	**0.130**	-	**6.910**	**0.835**	-	-
ACDepth [68]	2025				0.138	-	6.970	0.813	-	-
Robust-Depth [65]	2023			KITTI-C (Bad w.)	**0.114**	**0.891**	**4.878**	**0.868**	**0.958**	**0.981**
WaterMono [66]	2024		Underwater	FLSea	**0.099**	**0.117**	**0.945**	**0.892**	**0.978**	**0.991**
Wang & Ye [74]	2025	Supervised	Underwater	FLSea	**0.097**	**0.041**	**0.233**	**0.941**	**0.995**	**0.998**
DGS [90]	2025				0.244	0.240	0.413	0.721	0.935	0.974

D. The Foundation Era and Cross-Domain Generalization

Table IV shows the paradigm shift of the Foundation Era, demonstrating the power of large Vision Transformers and diffusion processes to generalize across structurally diverse datasets. The table shows zero-shot and non-zero-shot metric transfer, and there is a huge jump in absolute depth estimation capabilities. For the non-zero-shot evaluations, Metric3D (ViT-G) [81] provides a strong state-of-the-art baseline with an unprecedented Absolute Relative error (Abs Rel) of 0.039 on KITTI and 0.045 on NYUv2. The Depth Anything V2 (L) [78] architecture is closely followed, which also passes the 0.050 barrier on KITTI. The table also shows the promise of generative methods for metric estimation, with the latent diffusion model Marigold v1.0 [76] obtaining highly competitive scores of 0.099 and 0.055 on the respective datasets. Finally, the explicit zero-shot block demonstrates the real potential of the foundation paradigm. UniDepth [80] demonstrates that foundation models can dynamically infer absolute physical scale without dataset-specific fine-tuning, achieving a 0.042 Abs Rel on KITTI that is directly competitive to explicitly supervised metric architectures.

TABLE IV
THE FOUNDATION ERA. DEMONSTRATING THE CAPACITY OF MASSIVE VISION TRANSFORMERS AND DIFFUSION PROCESSES TO ACHIEVE HIGH-FIDELITY ZERO-SHOT AND NON-ZERO SHOT METRIC TRANSFER ACROSS STRUCTURALLY DISTINCT DATASETS.

Method (Backbone)	Year	Eval. Dataset	Zero-Shot?	Abs Rel ↓	RMSE ↓	δ1 ↑	δ2 ↑	δ3 ↑
Depth Anything V2 (L) [78]	2024	KITTI	No	0.045	1.861	0.983	**0.998**	**1.000**
Metric3D (ViT-G) [81]	2023			**0.039**	1.766	**0.989**	**0.998**	**1.000**
DPT-Hybrid [55]	2021			0.062	2.573	0.959	0.995	0.999
Marigold v1.0 [76]	2025			0.099	-	0.916	-	-
Re-Depth Anything [83]	2025			0.283	6.710	0.593	0.818	0.917

Depth Anything V2 (L) [78]	2024	NYUv2		0.056	0.206	0.984	**0.998**	**1.000**
BinsFormer (Swin-L) [31]	2022			0.052	2.098	0.974	0.997	0.999
Metric3D (ViT-G) [81]	2023			0.045	**0.187**	0.987	0.997	0.999
ZoeDepth [62]	2023			0.075	0.270	0.955	0.995	0.999
DPT-Hybrid [55]	2021			0.110	0.357	0.904	0.988	0.998
Marigold v1.0 [76]	2025			0.055	-	0.964	-	-
UniDepth [80]	2024	KITTI	Yes	**0.042**	1.750	**0.986**	-	-
UniDepth [80]	2024	NYUv2		0.058	**0.201**	0.886	-	-

E. The Efficiency Imperative for Edge Deployment

Table V shifts the focus from raw predictive capability metrics to the harsh realities of engineering considerations for edge inference, detailing the complex relationship between latency, computation cost (in FLOPs) and the number of parameters. Lite-Mono [91] sets a good baseline for low-latency inference with only 4.5 ms on NVIDIA Titan Xp. The latest models in 2024 such as Sui et al. [60] and LEDepth [59] follow the same path of making aggressive design choices with their small architectural footprint of around 3 million parameters with under 5 GFLOPs. But this table clearly showcases the weakness of being too aggressive in quantization. MobileDepth [54, 58] makes the conscious decision of going with a bigger footprint of 17.6 million parameters and 15.22 GFLOPs to break the 0.100 Abs Rel limit. This sacrifice allows for the outstanding 0.094 Abs Rel on KITTI and 0.087 Abs Rel on NYUv2 while maintaining a very practical 9 ms latency on RTX 3090. This final part makes it clear that when it comes to self-contained control, the best architecture cannot be one that minimizes the error.

TABLE V
THE EFFICIENCY TRADE-OFF. EVALUATING INFERENCE LATENCY, FLOATING-POINT OPERATIONS (FLOPS), AND PARAMETER COUNTS FOR ARCHITECTURES SPECIFICALLY TARGETED AT REAL-TIME, RESOURCE-CONSTRAINED EDGE DEPLOYMENT.

Method	Year	Infer. (ms)	FLOPs (G)	Params (M)	Dataset	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ1 ↑	δ2 ↑	δ3 ↑	GPU	Batch Size
MobileDepth [58]	2024	9	15.22	17.6	KITTI	**0.094**	-	-	**0.969**	**0.993**	**0.996**	NVIDIA RTX 3090	12
Sui et al. [60]	2024	11.4	**4.6**	**3.0**		0.105	0.810	4.506	0.891	0.963	0.983	NVIDIA RTX 6000	12
Wang et al. [61]	2024	-	4.9	3.1		0.104	0.757	4.563	0.889	0.964	0.983	NVIDIA RTX 3060 Ti	**8**
Lite-Mono [91]	2023	**4.5**	5.1	3.1		0.107	0.765	4.561	0.886	0.963	0.983	NVIDIA Titan Xp	12
Lite-Mono-8M [91]	2023	6.5	11.2	8.7		0.101	**0.729**	4.454	0.897	0.965	0.983	NVIDIA Titan Xp	12
LEDepth [59]	2024	5.7	4.9	3.1		0.101	0.744	**4.425**	0.899	0.965	0.983	NVIDIA RTX 3090	12
MobileDepth [58]	2024	**9**	**15.22**	**17.6**	NYU	**0.087**	-	-	**0.935**	**0.991**	**0.998**	NVIDIA RTX 3090	**12**

XI. Open Challenges and Future Directions

While monocular depth estimation has crossed the initial hurdle of recovering dense geometric scales from monaural 2D images, deploying these algorithms into the hard scenario of autonomous deployment reveals critical weaknesses in performance against controlled benchmarks. The literature gives us three direct tasks for research into future.

1. Algorithmic Interpretability and Explainable AI (XAI): With cutting edge models now fully dependent on opaque vision transformers and multistep latent diffusion processes, the semantic transparency of depth estimation is largely gone. It appears that, it would offer no mechanism to audit the failure if a foundation model hallucinated a contiguous flat road over an abrupt drop-off or masked a pedestrian at night (there is no architecture-level were auditing available). The system of localized epistemic uncertainty estimation with interpretable attention weighting is no longer just a favorable feature, instead it becomes the mandatory prerequisite for regulatory certification of automated vehicular safety systems

ViT-based depth predictions is nontrivial, moreover, as standard gradient-based attribution tools like GradCAM were developed for CNN classifiers and do not directly

apply to multi-head self-attention architectures. Chefer et al. [92], proposed a relevancy propagation method similar to that of attention rollout that identifies, by how much each input token influences the final prediction across all transformer layers at once. When applied to dense prediction tasks (e.g. depth estimation), this allows us to see if a network is relying on semantically meaningful cues like object boundaries and horizon lines, or falling prey to spurious correlations like image compression artifacts or dataset specific sky colorimetry

Similar but equally important ongoing challenge which runs parallel to interpretability, is the quantification of predictive uncertainty. While deploying in safety-critical applications, a depth map done without additional confidence signal is by definition untrustworthy. Poggi et al. A study using quantifying uncertainty in deep neural networks for self-supervised monocular depth estimation [93] showed that predictive variance can be approximated through modern techniques such as Monte Carlo Dropout, learned log-likelihood output heads and wondered if ensembles of randomly sampled models (self-distillation). Importantly, the study demonstrated that self-supervised uncertainty estimates are accurately calibrated against true prediction error: high-uncertainty regions remain qualitatively correlated with geometrically ambiguous areas known to exist in the data (specular surfaces, object boundaries at distance and sky). This work provides a solid framework for associating per-pixel confidence maps to depth results, allowing downstream planners in fully autonomous systems to assign predictions weights or reject them based on these mappings.

The integration of these existing tools into modern foundation models like Depth Anything V2 and Metric3D v2 is largely unaddressed, creating a disconnect. These models generate high fidelity depth maps but lack any intrinsic mechanism by which a system operator may audit the cause of a particular geometric error. One of the most urgent and solvable research needs in the field is to close this gap by integrating uncertainty heads and attention visualization pipelines directly into the training and inference stack of large-scale depth models

2. The "Unstructured" Deployment Void: Current general models, while achieving robust zero-shot metrics, are implicitly biased by the hyper-structured lane-abiding environments of Western datasets like KITTI and Cityscapes. Profoundly chaotic and highly dynamic traffic ecologies, characterized by the absence of lane demarcations, extreme inter-vehicle proximity, and non-standard vehicle profiles, remain an uncharted frontier for probing the limits of self-supervision. Addressing the dynamic motion-masking problem in such environments requires radically adaptive optical flow pipelines that do not rigidly assume static physical world geometries.

3. Converging the Robustness and Generalization Branches: The field is now divided between massive "General Models" such as Depth Anything V2 and narrowly specialized "Robust Models" such as PhysDepth, but there is still no Robust Foundation Model available. Training on pristine synthetic data, or training only on photometric data, limits the usefulness of a general model in the immersed-in-underwater or pushed-into-thick-snow case. The most promising future research direction is to mathematically embed atmospheric scattering laws and spatiotemporal continuity directly into the latent training manifolds of diffusion models, practically forcing them to inherently respect the physics of light transport.

XII. Conclusion

This review offers a comprehensive map of the remarkable chronological and architectural evolution of monocular depth estimation. We have traced the path of the discipline from its probabilistic mathematical infancy with handcrafted Markov random fields, to the brute-force regression of early deep convolutional neural networks, to the freeing of geometric scale via photometric self-supervision.

The introduction of global attention mechanisms via Vision Transformers shattered the spatial limitations of local convolutions forever and directly paved the way for the current era of Foundation Models. These large-scale networks have successfully broken the classical dataset limit, demonstrating the possibility of a universal algorithmic understanding of absolute metric depth in indoor and outdoor scenes without heuristic post-processing.

However, optimal benchmark scores on clean datasets hide the grim reality of physical deployment. The real frontier of monocular depth estimation has moved. The future of the domain will not be driven by incremental drops in baseline error, but by the ability of architectures to cram billions of parameters onto edge-compute processors, and their absolute structural resilience when blinded by extreme darkness, crippled by adversarial weather, or submerged entirely below the surface.

XIII. Acknowledgment

The authors would like to thank the Department of Computer Science, Faculty of Science and Technology, American International University-Bangladesh (AIUB) for supporting this research. Claude (Anthropic, Claude Sonnet 4.6) was used in the preparation of this manuscript in the following capacities: (1) reformatting the manuscript from IEEE Access LaTeX template to AJSE submission format; (2) light revision of sentence structure and grammar throughout the manuscript to improve clarity and reduce repetitive phrasing; (3) correcting structural errors including section cross-references in the introduction roadmap; (4) typesetting mathematical equations as native Word equation objects; and (5) verifying citation accuracy, specifically identifying and correcting an incorrect reference for the Depth Anything V2 synthetic training pipeline. All technical content, experimental results, quantitative data, literature citations, scientific claims, and figures were reviewed, verified, and remain the sole responsibility of the authors. The authors reviewed all AI-assisted output and are fully accountable for the final manuscript.

XIV. References

[1] J. E. Cutting and P. M. Vishton, "Perceiving layout and knowing distances," in Perception of Space and Motion, Elsevier, 1995, pp.

69–117.
[2] M. Bijelic et al., “Seeing through fog without seeing fog,” in Proc. IEEE/CVF CVPR, 2020, pp. 11682–11692.
[3] S. Shao et al., “Self-supervised monocular depth and ego-motion estimation in endoscopy,” Medical Image Analysis, vol. 77, p. 102338, 2022.
[4] S. Nadeem and A. Kaufman, “Computer-aided detection of polyps in optical colonoscopy,” Machine Vision and Applications, vol. 27, no. 7, pp. 1037–1053, 2016.
[5] M. J. Page et al., “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. no. n71, 2021. doi: 10.1136/bmj.n71.
[6] A Saxena, S. H. Chung, and A. Y. Ng, “Learning depth from single monocular images,” in Proc. NIPS, 2005, pp. 1161–1168.
[7] D. Eigen, C. Puhrsch, and R. Fergus, “Depth Map Prediction from a Single Image using a Multi-Scale Deep Network,” in Advances in Neural Information Processing Systems, vol. 27, 2014
[8] D. Hoiem, A. A. Efros, and M. Hebert, “Geometric context from a single image,” in Proc. ICCV, 2005, pp. 654–661.
[9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, vol. abs/1409.1556, 2014. Available: https://api.semanticscholar.org/CorpusID:14124313.
[10] R. Garg, B. G. V. Kumar, G. Carneiro, and I. Reid, “Unsupervised CNN for single view depth estimation: Geometry to the rescue,” in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer, 2016, pp. 740–756.
[11] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 270–279.
[12] Saxena, M. Sun, and A. Y. Ng, “Make3D: Learning 3D scene structure from a single still image,” IEEE Trans. PAMI, vol. 31, no. 5, pp. 824–840, 2009.
[13] T. Lindeberg, “Scale-space theory: A basic tool for analysing structures at different scales,” J. Applied Statistics, vol. 21, pp. 224–270, 1994.
[14] J. Diebel and S. Thrun, “An application of Markov random fields to range sensing,” NIPS, pp. 291–298, 2005.
[15] N. Silberman et al., “Indoor segmentation and support inference from RGBD images,” in Proc. ECCV, 2012, pp. 746–760.
[16] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2650–2658
[17] K. Karsch, C. Liu, and S. B. Kang, “Depth transfer: Depth extraction from video using non-parametric sampling,” IEEE Trans. PAMI, vol. 36, no. 11, pp. 2144–2158, 2014.
[18] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in Proc. 4th Int. Conf. 3D Vis. (3DV), Stanford, CA, USA, 2016, pp. 239–248, doi: 10.1109/3DV.2016.32.
[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[20] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.
[21] B. Li et al., “Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs,” in Proc. IEEE CVPR, 2015, pp. 1119–1127.
[22] P. Wang et al., “Towards unified depth and semantic prediction from a single image,” in Proc. IEEE CVPR, 2015, pp. 2800–2809.
[23] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, “Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 161–169
[24] S. Xie and Z. Tu, “Holistically-Nested Edge Detection,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.
[25] Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in Proc. IEEE CVPR, 2012, pp. 3354–3361.
[26] JJ. Xie, R. Girshick, and A. Farhadi, “Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks,” in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer, 2016, pp. 842–857.
[27] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[28] S. F. Bhat, I. Alhashim, and P. Wonka, “AdaBins: Depth estimation using adaptive bins,” in Proc. IEEE/CVF CVPR, 2021, pp. 4008–4017.
[29] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler and V. Koltun, “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1623-1637, 1 March 2022, doi: 10.1109/TPAMI.2020.3019967.
[30] Z. Li, X. Wang, X. Liu and J. Jiang, “BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation,” in IEEE Transactions on Image Processing, vol. 33, pp. 3964-3976, 2024, doi: 10.1109/TIP.2024.3416065.
[31] J. H. Lee et al., “From big to small: Multi-scale local planar guidance for monocular depth estimation,” arXiv:1907.10326, 2021.
[32] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017.
[33] S. Abdulwahab et al., “Deep monocular depth estimation based on content and contextual features,” Sensors, vol. 23, no. 6, 2023.
[34] J. Hu, M. Ozay, Y. Zhang and T. Okatani, “Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps With Accurate Object Boundaries,” 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2019, pp. 1043-1051, doi: 10.1109/WACV.2019.00116
[35] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 1 April 2018, doi: 10.1109/TPAMI.2017.2699184.
[36] S. Paul, D. Mishra, and S. K. Marimuthu, “Nested DWT-based CNN architecture for monocular depth estimation,” Sensors, vol. 23, no. 6, 2023.
[37] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, N. Navab, J. Hornegger, W. M. Wells III, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, vol. 9351, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[38] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. H. Hamprecht, Y. Bengio, and A. C. Courville, “On the spectral bias of neural networks,” in Proc. 36th Int. Conf. Mach. Learn. (ICML), ser. Proc. Mach. Learn. Res., vol. 97, 2019, pp. 5301–5310.
[39] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017.
[40] R. Mahjourian, M. Wicke, and A. Angelova, “Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018.
[41] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, “Unsupervised scale-consistent depth and ego-motion learning from monocular video,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, Curran Associates, Inc.
[42] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, “3D packing for self-supervised monocular depth estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020.
[43] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif, and D. Puig, “GCNDepth: Self-supervised monocular depth estimation based on graph convolutional network,” Neurocomputing, vol. 517, pp. 81–92, 2023, doi: 10.1016/j.neucom.2022.10.073.
[44] Y. Lyu, L. Liu, X. Wang, D. Kong, Y. Liu, Y. Liu, and Y. Chen, “HR-Depth: High resolution self-supervised monocular depth estimation,” Proc. AAAI Conf. Artif. Intell., vol. 35, no. 3, pp. 2294–2301, 2021, doi: 10.1609/aaai.v35i3.16329.
[45] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019.

[46] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth Prediction without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos", AAAI, vol. 33, no. 01, pp. 8001–8008, Jul. 2019.

[47] K. Zhou et al., "Manydepth2: Motion-Aware Self-Supervised Monocular Depth Estimation in Dynamic Scenes," in IEEE Robotics and Automation Letters, vol. 10, no. 7, pp. 6704-6711, July 2025, doi: 10.1109/LRA.2025.3568337.

[48] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1164–1174.

[49] J. Kopf, X. Rong, and J.-B. Huang, "Robust consistent video depth estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1611–1621.

[50] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX. Berlin, Heidelberg: Springer-Verlag, 2020, p. 582–600. [Online]. Available: https://doi.org/10.1007/978-3-030-58565-5_35

[51] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every pixel counts ++: Joint learning of geometry and motion with 3d holistic understanding," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 10, pp. 2624–2641, 2020

[52] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," in International Conference on Learning Representations, 2020. [Online]. Available: https://openreview.net/forum?id=ByxT7TNFvH

[53] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018.

[54] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". ArXiv, abs/2010.11929.

[55] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 12179–12188.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, Curran Associates, Inc.

[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021, pp. 10012–10022.

[58] Y. Li and X. Wei, "MobileDepth: Monocular depth estimation based on lightweight vision transformer," Applied Artificial Intelligence, vol. 38, 2024.

[59] Zhang et al., "LEDepth: A lightweight self-supervised monocular depth estimation network combining CNN and transformer," Signal, Image and Video Processing, vol. 19, 2025.

[60] X. Sui et al., "Lightweight monocular depth estimation using a fusion-improved transformer," Scientific Reports, vol. 14, 2024.

[61] Z. Wang et al., "Lightweight self-supervised monocular depth estimation through CNN and transformer integration," IEEE Access, vol. 12, pp. 167934–167943, 2024.

[62] S. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Mueller, "ZoeDepth: Zero-shot transfer by combining relative and metric depth," arXiv preprint arXiv: 2302.12288, Feb. 2023. doi: 10.48550/arXiv.2302.12288.

[63] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "Neural window fully-connected CRFs for monocular depth estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, pp. 3916–3925.

[64] J. Bai, H. Qin, S. Lai, J. Guo and Y. Guo, "GLPanoDepth: Global-to-Local Panoramic Depth Estimation," in IEEE Transactions on Image Processing, vol. 33, pp. 2936-2949, 2024, doi: 10.1109/TIP.2024.3386403.

[65] K. Saunders, G. Vogiatzis, and L. J. Manso, "Self-supervised monocular depth estimation: Let's talk about the weather," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2023, pp. 8907–8917.

[66] Y. Ding, K. Li, H. Mei, S. Liu and G. Hou, "WaterMono: Teacher-Guided Anomaly Masking and Enhancement Boosting for Robust Underwater Self-Supervised Monocular Depth Estimation," in IEEE Transactions on Instrumentation and Measurement, vol. 74, pp. 1-14, 2025, Art no. 5022514, doi: 10.1109/TIM.2025.3553943.

[67] K. Peng, H. Li, Z. Qi, H. Chen, Z. Wang, W. Zhang, S. He, H. Yang, and Q. Guo, "PhysDepth: Plug-and-play physical refinement for monocular depth estimation in challenging environments," arXiv preprint arXiv: 2412.04666, 2026. Available: https://arxiv.org/abs/2412.04666.

[68] K. Jiang, J. Cao, Z. Yu, J. Jiang, and J. Zhou, "Always clear depth: Robust monocular depth estimation under adverse weather," in Proc. Int. Joint Conf. Artif. Intell., May 2025, pp. 1251–1259.

[69] K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, and J. Yang, "Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021, pp. 16055–16064.

[70] Y. Huang et al., "DASP: Self-supervised nighttime monocular depth estimation with domain adaptation of spatiotemporal priors," IEEE Robotics and Automation Letters, vol. 11, no. 2, pp. 2074–2081, 2026.

[71] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," Int. J. Comput. Vision, vol. 48, no. 3, pp. 233–254, 2002.

[72] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, pp. 10684–10695.

[73] D. Akkaynak and T. Treibitz, "Sea-Thru: A method for removing water from underwater images," in Proc. IEEE CVPR, 2019, pp. 1682–1691.

[74] J. Wang and X. Ye, "Advancing monocular depth estimation by integrating underwater optical imaging priors into transformer-based network," Optics Express, vol. 33, pp. 32631–32648, 2025.

[75] J. Shen, Z. Huang, and L. Jiao, "Self-supervised monocular depth estimation on construction sites in low-light conditions and dynamic scenes," Automation in Construction, vol. 168, p. 105848, 2024.

[76] B. Ke et al., "Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2025.3591076.

[77] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2020, pp. 6840–6851.

[78] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," in Advances in Neural Information Processing Systems, vol. 37, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2024, pp. 21875–21911, doi: 10.52202/079017-0688.

[79] Roberts et al., "Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding," in Proc. IEEE/CVF ICCV, 2021, pp. 10912–10922.

[80] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "UniDepth: Universal monocular metric depth estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2024, pp. 10106–10116.

[81] M. Hu et al., "Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," IEEE Trans. PAMI, vol. 46, no. 12, pp. 10579–10596, 2024.

[82] G. Bae, I. Budvytis, and R. Cipolla, "Estimating and exploiting the aleatoric uncertainty in surface normal estimation," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 13137–13146.

[83] A. R. Bhattarai and H. Rhodin, "ReDepth Anything: Test-time depth refinement via self-supervised re-lighting," arXiv preprint arXiv:2512.17908, 2026. Available: https://arxiv.org/abs/2512.17908

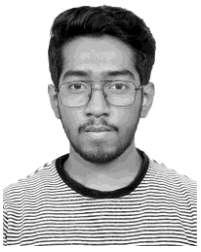
[84] R. Zhang and M. Shah, "Shape from intensity gradient", IEEE Trans.

Systems, Man, Cybernetics, vol. 29, no. 3, pp. 318–325, 1999.
[85] Z. Cao, J. Zhu, W. Zhang, H. Ai, H. Bai, H. Zhao, and L. Wang, "PanDA: Towards panoramic depth anything with unlabeled panoramas and Mobius spatial augmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2025, pp. 982–992.
[86] B. Coors, A. P. Condurache, and A. Geiger, "SphereNet: Learning spherical representations for detection and classification in omnidirectional images," in Proc. ECCV, 2018.
[87] Y. Randall et al., "FLSea: Underwater visual-inertial and stereo-vision forward-looking datasets," Ph.D. dissertation, Univ. of Haifa, Israel, 2023.
[88] Z. Yu, L. Jin, and S. Gao, "P2Net: Patch-match and plane-regularization for unsupervised indoor depth estimation," in Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, Aug. 2020, pp. 206–222. doi: 10.1007/978-3-030-58586-0_13.
[89] P. Ji, R. Li, B. Bhanu, and Y. Xu, "MonoIndoor: Towards good practice of self-supervised monocular depth estimation for indoor environments," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021, pp. 12767–12776. doi: 10.1109/ICCV48922.2021.01255.
[90] W. Gu and L. Qi, "Dense geometry supervision for underwater depth estimation," in Advanced Fiber Laser Conference (AFL 2024), vol. 13544, SPIE, 2025, p. 1354405. doi: 10.1117/12.305682.
[91] Zhang et al., "Lite-Mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation," in Proc. IEEE/CVF CVPR, 2023, pp. 18537.
[92] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, Online, 2021, pp. 782-791, doi: 10.1109/CVPR46437.2021.00084.
[93] M. Poggi, F. Aleotti, F. Tosi and S. Mattoccia, "On the Uncertainty of Self-Supervised Monocular Depth Estimation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3224-3234, doi: 10.1109/CVPR42600.2020.003.

Hasan Mahmud Shanto is an undergraduate student in Computer Science and Engineering at the American International University-Bangladesh. He has built a strong foundation in academic research through his roles as a Research Intern with the Applied Intelligence and Informatics Research Group in Nottingham, UK, as well as the Ubiquitous, Cloud, and Human-Computer Interaction (UCH) Research Group. His core research interests encompass a broad spectrum of artificial intelligence, specifically computer vision, deep learning, and explainable AI (XAI). Mr. Shanto's work focuses on robust learning engineering frameworks that are capable of operating reliably in challenging visual environments.

Mohammad Tofiqul Islam is pursuing a B.Sc. in Computer Science and Engineering at the American International University-Bangladesh (AIUB), Dhaka, Bangladesh. He is a Research Intern at the Applied Intelligence and Informatics Research Group, Nottingham, UK, and a Research Assistant in the Ubiquitous, Cloud and Human Computer Interaction (UCH) Research Group. His research focuses on computer vision, deep learning, and robust machine learning for autonomous systems. His recent work explores the security and reliability of large language model based recommender systems, particularly in the context of shilling attack detection and mitigation.

Muhammad Ryan Hasan is an undergraduate student studying Computer Science and Engineering in the Department of Computer Science at the American International University-Bangladesh (AIUB). His research interests include Computer Vision, Deep Learning, Machine Learning, Low-Resource Computation, Embedded Electronics, and IoT. He is committed to advancing fast, secure, and open-source technologies, with a focus on fostering transparency and innovation for society.



Supprio Saha Pranto is currently pursuing the B.Sc. degree in Computer Science and Engineering from American International University-Bangladesh (AIUB), Dhaka, Bangladesh. He is currently involved in research activities related to artificial intelligence and data-driven systems. His work focuses on applying machine learning techniques to solve real-world problems and developing efficient computational models. His interests include Machine Learning, TinyML, Natural Language Processing, Embedded System and IoT.

Tanvir Ahmed is an Assistant Professor of Computer Science under the Faculty of Science and Technology at the American International University-Bangladesh (AIUB). He completed his graduation from AIUB and was recognized for his notable academic performance. His research interests and passion are centered on Computer Vision, Pattern Recognition, Image Processing, Artificial Neural Networks, Machine Learning, Object Tracking and Detection, and Evolutionary Computing, with a focus on applications in intelligent visual systems. His current research focuses on developing deep learning and AI-based techniques for image analysis, object recognition, and advanced computer vision applications.

Rifath Mahmud is currently working as a Assistant Professor at the Department of Computer Science at American International University-Bangladesh (AIUB). Before that he worked as a Teaching Assistant at AIUB. Being a TA, his

main responsibilities were to assist teachers in their academic, official affairs and additional responsibilities assigned by the department. He also have completed his internship from Financial Solution and R&D department of BRAC IT Service Ltd. (biTS) which was mainly focused on Data Mining and Optical Character Recognition for remittance management software. Throughout the internship, he obtained knowledge of working in industry level projects and effective communication skills.

Syeda Anika Tasnim have completed her B.Sc. in Computer Science and Engineering and M.Sc. in Computer Science from American International University-Bangladesh and also completed her CCNA and MCSA (Windows Server 2016) certification. She have done her internship at National Bank IT Division. There she worked on different network monitoring tools like SolarWinds, Cacti, Nagios and Syslog. She is Good at Problem-solving, Mathematics, Data Mining, Relational Database (Oracle, MySQL).