Applied Computer Systems
ISSN 2255-8691 (online)
ISSN 2255-8683 (print)
2025, vol. 30, no. 1, pp. 21–33
https://doi.org/10.2478/acss-2025-0003
https://content.sciendo.com

Monocular Depth Estimation: A Review on Hybrid Architectures, Transformers and Addressing Adverse Weather Conditions

Lakindu Kumara[1*], Nipuna Senanayake[2], Guhanathan Poravi[3]
[1–3]*Informatics Institute of Technology, Colombo, Sri Lanka*

***Abstract* – Monocular depth estimation is one of the essential tasks in computer vision as it can provide depth information from 2D images and is extremely beneficial for applications such as autonomous driving, robot navigation, etc. Monocular depth estimation has significantly improved over the past couple of years and deep learning-based methods have surpassed traditional and machine learning-based methods. Deep learning-based methods have further been enhanced using transformer and hybrid approaches. This paper first discusses the sensors used for depth estimation and their limitations. Then, we briefly discuss the evolution of depth estimation. Then we dive into the deep learning methods including transformer and CNN-transformer hybrid methods and their limitations. Later, we discuss several methods addressing challenging weather conditions. Finally, we discuss the current trends, challenges and future directions of the transformer and hybrid methods.**

***Keywords* – Addressing weather conditions, attention, CNN-transformer hybrid methods, monocular depth estimation.**

I. Introduction

Depth estimation from 2D images has been one of the most important tasks in computer vision as it can provide a better understanding of the surroundings, which can be extremely useful for applications such as autonomous driving, simultaneous localisation and mapping (SLAM), indoor localisation, robotic navigation, architectural modelling, virtual reality, etc. [1]–[4]. Numerous studies have been performed in this field, and many methods have been proposed and achieved outstanding results. However, it is quite a challenging task due to its inherent scale ambiguity, making it an ill-posed problem. This means there can be numerous valid depth maps for an image [2].

A. Sensors Used in Autonomous Driving

Autonomous vehicles and vehicles with advanced driver-assistant systems are becoming extremely popular on our roads. These vehicles use advanced sensors to gather information about their surroundings and act accordingly without the need for human interaction. The sensors collect real-time information about the surroundings and feed it into a computer, which processes it and makes critical decisions on how the vehicle must be controlled. The decisions are heavily impacted by the quality of the data fed into the computer; therefore, high-quality data is crucial for accurate decision-making. Many sensors are used in self-driving vehicles; the most common are LIDAR, RADAR and Camera [5].

B. Light Detection and Ranging (LIDAR)

These sensors use a laser light to measure the distances. Laser pulses are emitted by the LIDAR, which bounces off the object and returns to the LIDAR, allowing the system to create a 3D map of the vehicle surroundings.

LIDAR sensors are extremely accurate and they are also effective in several tricky conditions such as low light, rain, snow and fog as well, which makes them ideal for autonomous driving vehicles [6].

C. Radio Detection and Ranging (RADAR)

RADAR sensors are similar to LIDAR sensors but the former use radio waves instead of laser beams to detect objects. A radio signal that bounces off objects in its path is emitted by the RADAR sensor and it can determine the speed, direction and location of the object allowing the computer to calculate the distance.

RADAR sensors also have a larger range when compared to LIDAR, which makes them extremely useful for detecting objects that are further away; they can also work in more complex weather conditions and they are relatively less expensive when compared to LIDAR sensors [7].

D. Why are many shifting to monocular depth estimation now?

Sensors such as LIDAR and RADAR offer a lot of advantages, such as effectiveness in rough weather conditions, detection over a larger range, accurate distance measurements etc. However, they also come with several disadvantages. While LIDAR sensors can offer accurate distance measurements, they only generate a sparse depth map [8] and their functionality can be affected by smoke, dust and other environmental conditions. Another major downside of LIDAR is its cost as it is relatively expensive compared to other sensors [6].

*Corresponding author's e-mail: lakindu.20200987@iit.ac.lk
Article received 2024-08-22; accepted 2024-12-18

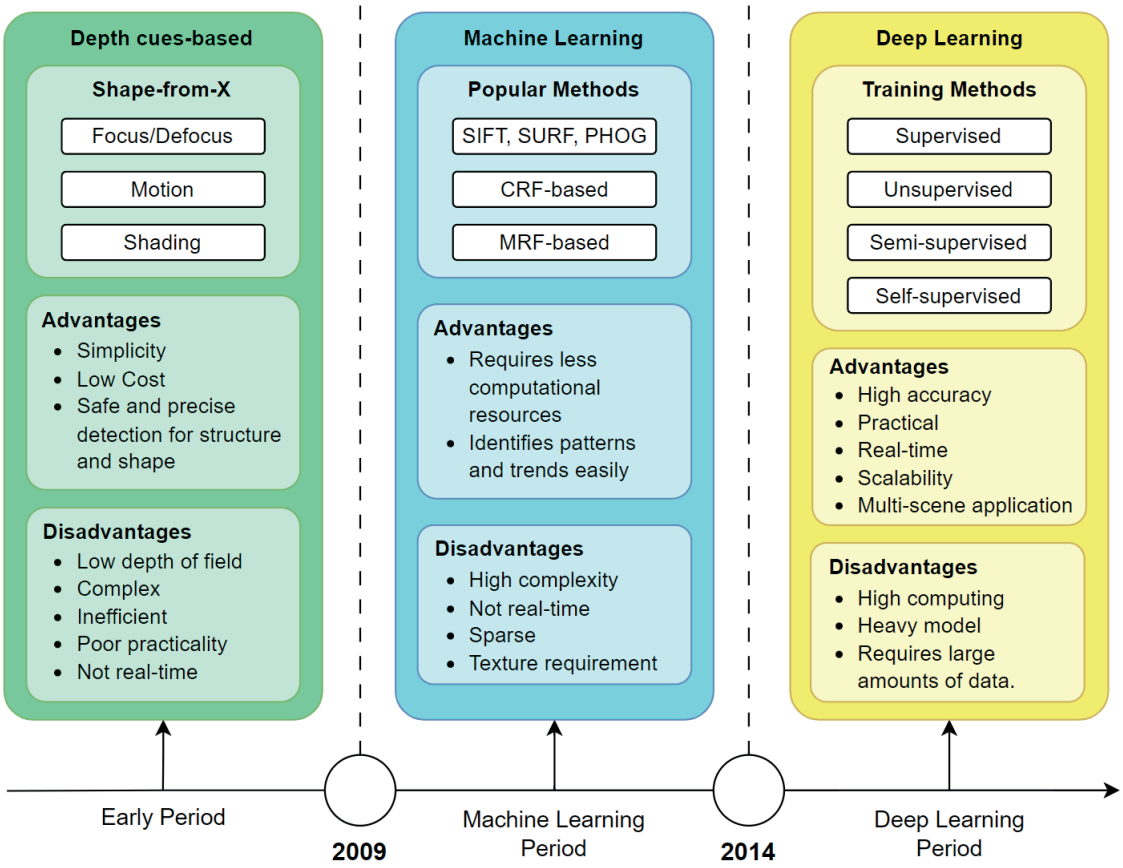

Fig. 1. Evolution of depth estimation [1], [3], [4].

RADAR sensors are a more cost-effective solution and perform well in various weather conditions. However, RADAR sensors have a lower resolution, which makes it difficult to identify smaller objects and RADAR is also vulnerable to interference generated from other RADAR sources which might affect its accuracy. Another major downside of RADAR is that it has a limited field of view, a flexible field of view like a smaller field of view for the highway and a wider field of view for the detection of pedestrians must be needed to detect the surroundings more effectively [7].

II. Camera-Based Depth Estimation

2D images are the most commonly used data format in computer vision. One of the major disadvantages of using 2D images is that it does not contain depth information. The depth information is extremely important for several applications in computer vision, such as autonomous driving, navigating complex environments, augmented reality and robotics.

Depth estimation tackles this challenge by reconstructing the 3D structure from a 2D image. This can be done by using images from stereo cameras, known as stereo depth estimation, or by using images from a single camera, known as monocular depth estimation (MDE). With recent advancements in deep learning, depth estimation has come a long way starting from traditional methods such as depth cues to machine learning statistical methods to deep learning-based methods such as CNNs, and this evolution has significantly improved the accuracy of depth estimation.

A. Early Methods of Depth Estimation

At the early stages of estimating the depth from 2D images, researchers utilised depth cues or visual features. There are two types of visual cues for depth estimation: monocular cues and stereo cues. Monocular cues include methods like focus/defocus, texture variations and gradients, occlusion, known object sizes, etc., which are used to provide cues to estimate the depth. On the other hand, stereo cues combine the views from stereo cameras to estimate the depth by calculating the difference in the position of the corresponding points in the two views, which is known as disparity.

Most of the monocular cues depend on contextual information, which is global information about the image, and therefore cannot be deduced from small image segments. Some depth information could be given by the colour and texture of a segment; however, it is insufficient for accurate depth estimation.

Stereo cues provide better depth estimation compared to monocular cues; however, since the disparity is inversely proportional to the object's distance, it is not a reliable cue for minor depth variations over large distances [9].

B. Machine Learning-Based Methods

Several techniques, which use probabilistic graphical models and handcrafted features for monocular depth estimation, were introduced with the rise of machine learning. These methods aimed to extract informative cues from single images to infer depth. Scale-invariant feature transform (SIFT) and speeded up robust features (SURF) both identify key points and stable image features across various scales and orientations with SURF offering computational efficiency [4].

Additionally, the pyramid histogram of oriented gradients (PHOG) captured the distribution of gradient orientations within an image. Machine learning techniques, including parameter and non-parameter learning, were employed to exploit these features and estimate depth maps [4]. Furthermore, Conditional Random Fields (CRF) and Markov Random Fields (MRF) were introduced as probabilistic graphical models to model the relationships between image features and depth values [4].

C. Deep Learning-Based Methods

Monocular depth estimation (MDE) has made significant progress in recent years due to advancements in deep learning. Most existing architectures consist of encoder-decoder structures using either the U-Net or autoencoder with skip connections as the base. Convolutional neural networks (CNNs) have become an essential method for extracting depth information from 2D images and they have achieved good results as well, with the use of transfer-learning the results have improved even further. However, recently Li et al. [2] have pointed out potential limitations in MDE using CNNs, specifically due to their limited local receptive field.

III. Deep Learning-Based Monocular Depth Estimation Methods

A. Training Techniques

Supervised Learning: Supervised learning in MDE is a method that requires ground truth (GT) depth maps and images to train a network. The goal of this method is to minimise the error between the image and the GT depth map using various loss functions for accurate depth prediction. In MDE, GT depth is mostly collected by utilising LIDAR sensors or stereo cameras [10]–[12]. Eigen et al. [13] were one of the first to explore depth estimation by using a CNN, which contains two networks: a global course-scale network and a local fine-scale network. Later, multiple methods were proposed such as using residual learning for modelling the mapping between the depth maps and images [14], and using conditional random fields (CRF) and regression [15].

Due to the limited access to large amounts of depth data [16], researchers have explored transfer learning techniques to tackle this challenge of limited labelled data by using the knowledge from pre-trained models such as DenseNet [17] and ResNet [16], [18], [19].

Unsupervised Learning: Unsupervised learning was also another method used for MDE and was preferable because it is difficult to find large amounts of accurate ground truth data [20]. Garg et al. [21] proposed a method where the depth is predicted using an inverse warp of the target image that is created using the predicted depth and known inter-view displacement. Godard et al. [20] proposed a method that predicts depth by deducing the disparities that warp the left image to align with the right one. There have also been several methods proposed that address the slow inference speed of existing unsupervised learning methods [22], [23].

Self-Supervised Learning: Self-supervised learning methods have become very popular recently, addressing the challenge of obtaining per-pixel ground truth depth data [24]. Most of the recent work is based on the architecture proposed by Godard et al. [24], which contains a DepthNet to predict the depth and a PoseNet, which predicts the pose between a pair of frames. The performance of self-supervised MDE has increased significantly by using various loss functions and complex model architectures, and most of the recent architectures utilising this method have achieved state-of-the-art results [25]–[27]. These architectures contain two main parts: the depth network and the pose network.

***Depth Network*:** This network extracts depth information from the image and predicts the depth, and most of the existing models use the U-Net as the base of the depth network.

***Pose Network*:** This network focuses on analysing the camera motion (also known as ego-motion) across a pair of frames to estimate the target's movement and the ResNet-18 network is mostly used as the pose network in the existing research.

B. CNNs vs. Transformers for Monocular Depth Estimation

Convolutional Neural Networks (CNNs): Many CNN-based methods have been proposed for MDE using various architectures and pre-trained models, and they have achieved commendable results [16], [19]. These methods are based on encoder-decoder networks where various models such as DenseNet and ResNet are used as the encoder of the network for better feature extraction. However, one of the major limitations of CNNs is its limited receptive field, which may lead to poor performance because it hinders the ability of the network to model long-range relationships among the pixels [28]–[31]. Also, sometimes the CNN-based methods fail to predict the background and foreground structures mainly due to the lack of long-range relationships and global context [28].

Transformer-based methods: Transformer-based methods have been gaining more interest in computer vision and they have shown outstanding results due to their ability to capture long-range relationships [28], [32].

However, there are a few limitations in transformer-based methods as well. The lack of spatial inductive bias in transformers can hinder their ability to effectively model local information, potentially affecting the performance of depth estimation. The fixed token scale of ViTs might limit their ability to model multi-scale features as well [31].

CNN-Transformer Hybrid Methods: Recently, CNN-Transformer hybrid methods have become very popular in monocular depth estimation, and this is mainly because these networks are able to extract both local information and long-range relationships [28], [30], [31]. CNNs excel in extracting local features, and transformers excel in capturing long-range

relationships; therefore, combining these helps overcome the limitations of both CNNs and transformers resulting in better depth estimation performance [28], [30], [31], [33].

IV. Transformer-Based Methods and Hybrid Architectures

A. Transformer-Based and Hybrid Methods

Zhao et al. [28] propose a novel CNN-Transformer hybrid encoder-decoder network called MonoViT to capture both the local and global features. The network consists of a depth network and a pose network. The encoder of the depth network contains a series of combined convolution layers and transformer blocks. The convolution layers focus on extracting the local information, and the transformer blocks focus on capturing the long-range relationships. The authors have used view reconstruction loss and smoothness loss, and they have also used the ResNet18 architecture for the pose network making the architecture lightweight.

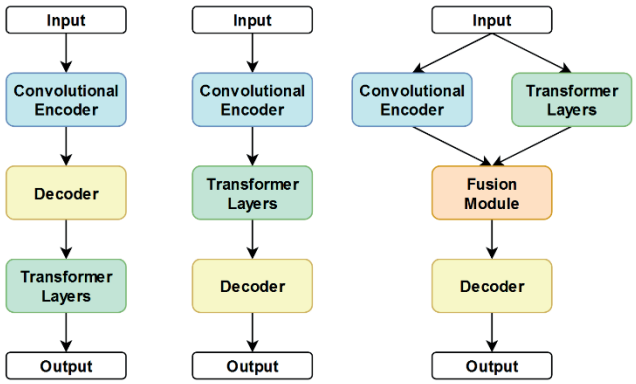

Fig. 2. Popular hybrid architectures in MDE [26], [28], [31].

Shim and Kim [34] introduce another method called Swin-Depth, which contains a convolution-free transformer as the encoder and a densely cascaded network as the decoder. The convolution-free transformer is used to capture both local and global features, and it reduces the computational expenses by utilising an attention window instead of the standard multi-head self-attention and patch merging strategy.

Rahman and Fattah [29] also propose a transformer-based framework called DwinFormer for MDE and the proposed architecture contains two key components: the dual window self-attention transformer (Dwin-SAT) and the dual window cross-attention transformer (Dwin-CAT). The Dwin-SAT module is responsible for extracting the local features and global context, and the Dwin-CAT module is responsible for combining the features from the encoder and decoder. The Dwin-SAT module uses the Dwin-SA module to incorporate both local and global contexts by alternating between the Lwin-SA (local window self-attention) and Gwin-SA (global window self-attention) modules. On the other hand, the Dwin-CAT module utilises Dwin-CA to combine the encoded and decoded features by utilising local window cross attention (Lwin-CA) and global window cross attention (Gwin-CA), which combines both the local and global features. The authors have utilised the scale-invariant loss function, and they have achieved state-of-the-art results. The proposed method outperforms the existing transformer-based methods as well.

Manimaran and Swaminathan [30] introduce another architecture called Focal-WNet that uses two encoders and a single decoder. The authors also reduce the computational requirements by utilising focal self-attention. The DenseNet-161 has been chosen as the first encoder because of its efficient utilisation of feature sharing and its capability to mitigate the issue of vanishing gradients. The authors have used the data loss for the KITTI dataset and a combination of data loss and gradient loss for the NYU Depth-V2 dataset. The authors also use focal transformers as the second encoder in their proposed network because of their capacity to establish these relationships with precise details in nearby surroundings and broader details with tokens that are distant from the query. After the image has been encoded by two distinct encoders at varying input resolutions, the latent variables from each encoding stage are fed into the up-sampling layer. During the final decoding stage, the grayscale version of the input is processed through a separate convolution block before being fed into the final up-sampling layer. This approach allows the two encoders to extract important depth information, while preserving precise object boundaries in the resulting depth map. The authors also mention that this method will be extended to other problems with dense predictions.

The limited receptive field of the CNN problem has also been addressed by Li et al. [31]. They proposed a novel architecture that consists of a parallel encoder containing a transformer branch and a convolution branch in parallel for effective feature extraction. The transformer branch of the proposed architecture effectively extracts the global context by using an attention mechanism, and the local information is effectively preserved by the convolution branch. Due to the excessive memory requirements of using global attention, the authors have implemented a deformable scheme to reduce the memory requirements. The authors also introduce a HAHI (hierarchical aggregation and heterogeneous interaction) module to improve the transformer features using the deformable self-attention module. The HAHI module is also introduced to better model the relationship between transformer and CNN features using a set-to-set translation method using the deformable cross-attention module. The decoder of the proposed method is built of a series of UpConv layers (that contain convolutional and upsampling layers) with skip connections. The authors mention that the explainability and transparency of the proposed method could be explored in the future.

Ning and Gan [35] address the quadratic complexity of multi-head attention by introducing a novel attention mechanism called Trap Attention. This attention mechanism places traps around each pixel in an extended space, and this attention mechanism is formed based on the feature retention ratio of the convolution window, effectively converting the quadratic computational complexity into a linear form. The authors also built an encoder-decoder architecture, which consists of a vision transformer as the encoder and utilises the trap attention mechanism in the decoder to capture the spatial relationships

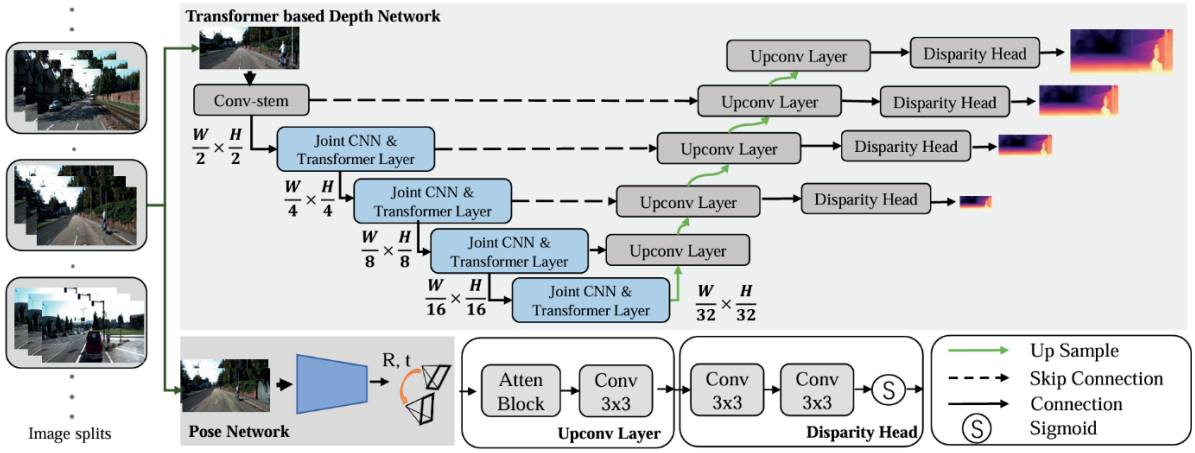

Fig. 3. The proposed architecture of Zhao et al. [28].

effectively. The authors also utilise the scale-invariant logarithmic loss when training. The trap attention mechanism relies on manual traps to identify the feature importance, then it utilises a depth-wise separable convolution layer to condense the relevant information and attention. The authors use a reverse pixel shuffle operation for pixel rearrangement in order to maintain consistent resolution between the input and output. Then, the trap interpolation method (the operation where manual traps are set) is applied for feature classification and then a depth-wise separable convolution layer is used to extract the feature relevance and attention information. A block selection unit is introduced in the vision transformer (ViT) encoder to reduce attention on the background by comparing feature maps of *n* consecutive candidate blocks pixel-wise and selecting the pixel with the maximum value.

Astudillo et al. [36] propose an efficient encoder-decoder architecture called DAttNet. The input image first goes through an encoder with a ResNet backbone where feature extraction will take place. Then these feature maps will be passed through a self-aware attention module (SAAM), which will extract the most important details from the feature maps. The features are upsampled using nearest-neighbour interpolation and the channel counts are aligned with intermediate backbone outputs through convolutional layers followed by an exponential linear unit (ELU), ensuring multi-scale information integration for the final output. The proposed SAAM module consists of two parallel sub-modules: the context attention layer and the self-attention layer. The context attention layer extracts global context via global average pooling and refines feature maps for depth estimation. Meanwhile, the self-attention layer learns spatial relationships within the feature map to re-weight it according to relevance, combining both layers results in a more comprehensive feature map capturing both global and local context information. The authors use a combination of scale-invariant logarithmic loss and multi-scale L1 loss, and the robustness of this method in adverse weather conditions could also be explored in the future.

Agarwal and Arora [37] introduce a novel approach called PixelFormer that applies an encoder and decoder feature fusion method utilising attention. The authors used the scaled version of the scale-invariant loss function. In the proposed method, a Swin Transformer encoder initially extracts feature maps with multiple scales from input images, and these feature maps contain a global receptive field due to the transformer encoder. The feature map with the coarsest resolution is passed as inputs to the Pixel Query Initializer (PQI) module and this PQI module initialises pixel queries by combining the scene information using global average pooling. These pixel queries are then passed as input into the Bin Centre Predictor (BCP) module where the bin widths are produced. Then the Skip Attention Module (SAM) is used to refine the initial pixels to a higher resolution. Finally, the probability distribution of each pixel across the bin centres is taken by a convolution and softmax operation. This method could be extended to other dense prediction methods in the future.

Zhao et al. [38] proposed an architecture called SAU-NET, which contains a convolution-free transformer for feature extraction. The network also contains a PoseNet, which is a standard CNN containing a ResNet34 [18] as the encoder and the AU-Net as the decoder, and the gradient loss function is used. The proposed Stratified Transformer is based on the Swin Transformer [39]. The authors optimised the segmentation map by using a single Swin Block and reclassifying each pixel to avoid the problem of missing segmentation during the first time of feature mapping. Window multi-head self-attention was also used to reduce the computational requirements. The authors also mention that the standard skip connection of the U-Net [40] is too simple to improve the connectivity of the features in the decoder. Therefore, they use an attention mechanism in the skip connection to extract the most important image features, and these features are joined using the method of adding elements.

Xia et al. [33] also address the limited receptive field of CNNs by proposing a parallel CNN and Transformer architecture called PCTDepth. Both local and global features (long-range dependencies) are extracted by using a ResNet and Swin Transformer. Additionally, the authors also introduce a Hierarchical Fusion Module (HFM) for the effective fusion of features, and they use the scale-invariant loss function.

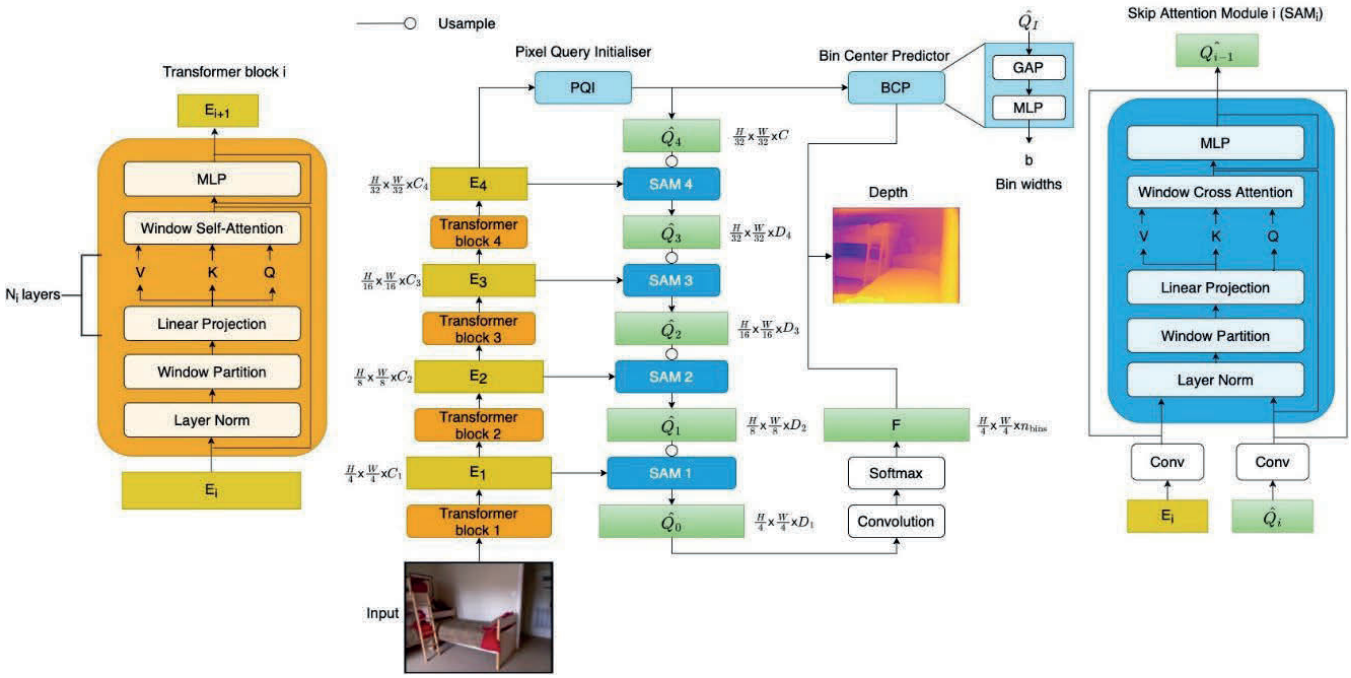

Fig. 4. Proposed architecture of Agarwal and Arora [37].

Then the accuracy of the model is improved by focusing on the spatial locations and by improving the inter-channel correlations of the fused features. First, the features are extracted by the ResNet and Transformer blocks, and these features are then combined using the HFM, which combines the features through adaptive feature alignment. Then, the features are passed through a series of convolutions, upsampling operations and with the help of the proposed Dual Attention Module (DAM) module to obtain the original resolution. The DAM attention module consists of a Channel Attention block (CA) and a Spatial Attention block (SA). The CA module is responsible for capturing global context, and the SA module is responsible for capturing local features. The authors mention that model pruning techniques and lightweight architectures will be explored in the future to make the model lightweight and practical.

Xing et al. [41] propose a novel method where they fuse the semantic and depth information. The authors also present a novel local adaptive attention approach to enhance geometrically aware representations. Specifically, the authors use geometric clues from semantic data to train local adaptive bounding boxes, directing the unsupervised aggregation of features. The transformer attention mechanism is built to extract global dependencies, enabling the creation of spatial interactions over expanded regions. The authors also utilise a multi-head structure to produce multiple proposals, each offering distinct ROIs for local attention.

Yan et al. [42] proposed a lightweight hybrid CNN-transformer network called EMTNet for MDE. The authors also introduced a mobile transformer block (MTB), which reduces the number of parameters by reusing them using self-attention. The encoder of the proposed method uses a Linear Block (LB) to extract local features and the proposed MTB block, which consists of MoSA and MoFFN modules to extract global context effectively. A fusion module is also introduced to prevent the loss of information during the decoding process by fusing the features from the corresponding encoder stage with the previous fusion module features. The MoSA module in the MTB reuses weights in the Query Key and Value calculations by using a branch-sharing scheme, and the MoFFN module uses the Ghost [43] module instead of the standard linear layer of the self-attention mechanism. The authors have achieved good results. However, there are a few limitations as well. The researchers highlight a major concern regarding the parameter count and computational complexity of their model compared to other hybrid models. The authors also mention that this could be due to the limited coordination between the components of the CNN and transformer. Therefore, the authors will explore alternate modelling techniques to overcome these limitations.

B. Lightweight Model Architectures

Complex model architectures achieve commendable performance; however, they are not suitable for real-time inference due to the computational requirements and high inference times. Therefore, methods to decrease the complexity and lightweight model architectures must be explored to achieve real-time accurate depth estimation and to improve the practicality of the models [33].

Existing lightweight MDE methods frequently face challenges in terms of their representation capacity and can require higher computational resources for image reconstruction [44], showing a trade-off between complexity and accuracy.

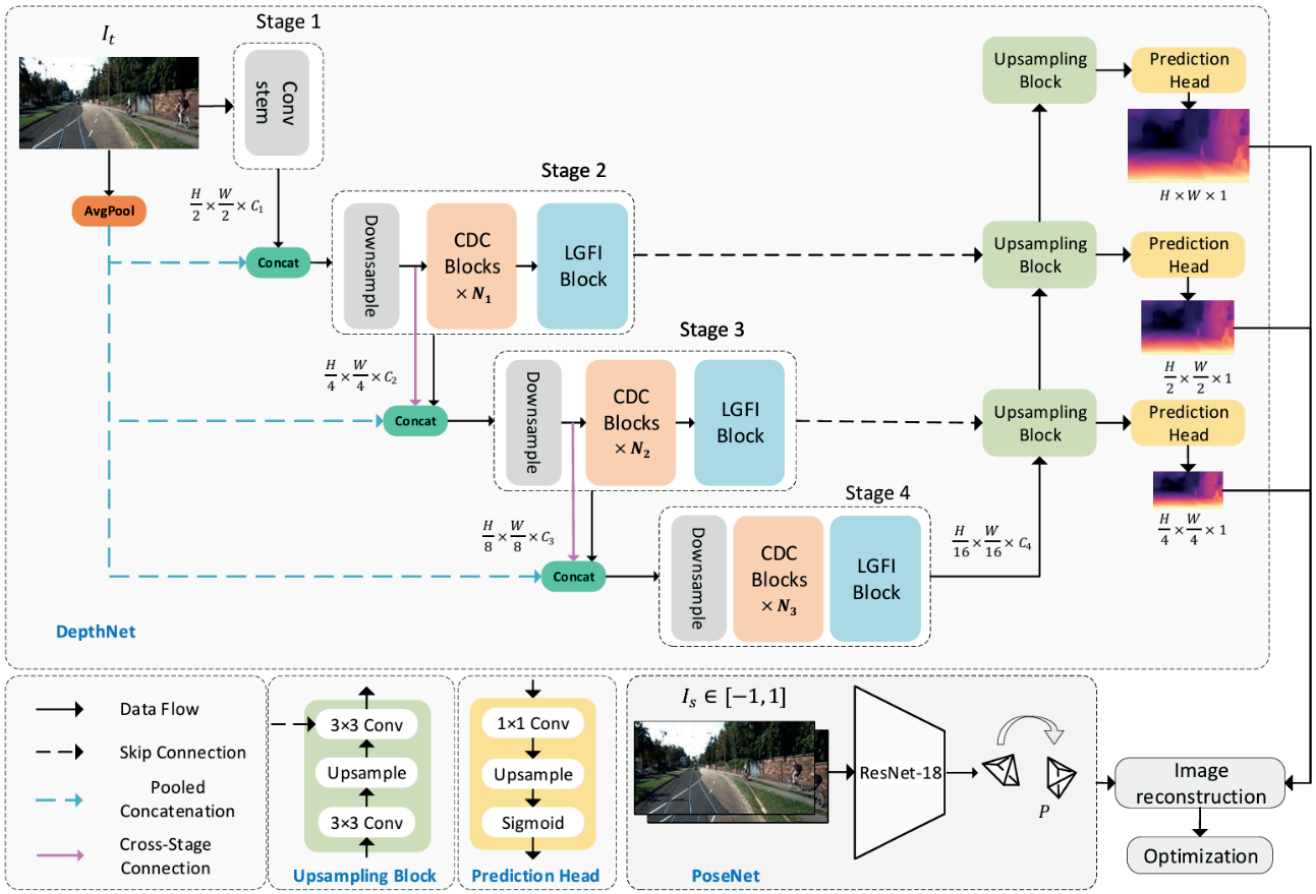

Fig. 5. The proposed architecture of Zhang et al. [26].

Recently, Papa et al. [45] addressed the challenge of monocular depth estimation on resource-constrained devices by introducing a novel lightweight framework called METER. The proposed method is an encoder-decoder architecture that combines convolutional and transformer operations. The encoder contains several METER blocks that extract both local and global features from the image. Each METER block contains a convolutional block that is followed by a transformer block. The decoder contains a fully convolutional network that generates the depth map by upsampling the encoded features. The proposed method uses a loss function, which combines several losses, depth loss, gradient loss, normalized gradient loss and SSIM loss, to ensure a great balance between the details of the edges, the accuracy of the depth and the overall depth map quality. There are three configurations of the METER and all three of them achieve state-of-the-art results on the NYU Depth v2 and the KITTI datasets.

Jin et al. [25] propose a network architecture for MDE that integrates two key modules, the Detail Highlight Module (DHM) and the Dense Geometric Constraints Module (DGC). The DHM module combines the information across various scales, emphasising the important details and enhancing the depth estimation. The loss function of the proposed method is a combination of image reconstruction, photometric reprojection and edge-aware smoothing losses. The DGC module retrieves precise scale factors, particularly valuable in autonomous driving scenarios where additional sensors may not be available. The proposed network architecture also contains a DepthNet and PoseNet where the multi-level features of the image are extracted by the DepthNet encoder by utilising multiple dilated convolution and feature interaction modules, and the spatial dimensions of the encoded features are increased using bilinear upsampling. These features are then integrated with the newly generated features through convolution layers and the DHM module. The PoseNet of the proposed architecture uses a ResNet18 backbone.

Recently, Liu and Zhou [46] have introduced LightDepthNet, which is another lightweight architecture for quick and accurate depth estimation on edge devices. The authors used a MobileNetV2-based encoder and Residual Merging Upsampling Modules (RMU) in the decoder to extract important depth information by combining feature maps from the encoder and the upsampled features. The authors have also reduced the computational cost by minimising the number of channels using a unified connected channel construction scheme. The proposed architecture is based on an approach that further reduces the computational requirements by using cosine recovery adaptive channel pruning. The authors have achieved a lower RMSE and better speed compared to other state-of-the-art methods.

Fig. 6. Fog augmentation by Tang et al. [47], (a) Original Image, (b) Visibility at 50 m, (c) Visibility at 100 m and (d) Visibility at 200 m.

Zhang et al. [26] introduce a lightweight self-supervised method that uses a DepthNet and PoseNet and a hybrid CNN and transformer network to extract both global and local features from an image. The CNN focuses on extracting the local features from various scales, and the transformer focuses on modelling the global long-range context. The authors also propose a Consecutive Dilated Convolutions (CDC) module, which allows the network to view larger areas of the image without any additional parameters, and a Local-Global Features Interaction (LGFI) module, which extracts global information by using cross-covariance attention in the feature channels, which significantly decreases the amount of memory required and the time complexity. Due to the efficiency of this method, it is highly suitable for fast inference on edge devices in applications like autonomous driving where inference speed is crucial.

Song et al. [44] propose a method, which uses a Spatial-Aware Dynamic Encoder (SAD-Encoder) and a Multi-Scale Sub Pixel Decoder (SUB-Decoder). The proposed SAD-Encoder contains the spatial-aware dynamic depth-wise separable convolution (SAD-DS) module, which is used to generate input-adaptive dynamic depth-wise convolution kernels based on the spatial information of the input to improve the depth-wise separable convolution representation capacity. The model size is reduced by the multi-scale sub-pixel lightweight decoder while still preserving detailed high-resolution image reconstruction by using sub-pixel convolution.

V. Addressing Adverse Weather Conditions

Monocular depth estimation has come a long way and has improved a lot, many existing MDE methods achieve state-of-the-art results and high-quality depth maps. However, most of the existing methods struggle to obtain good results in rough weather conditions mainly due to the lack of ground truth data and because they rely on simple depth cues [48].

To address this issue, Saunders et al. [48] introduced Robust-Depth. This method combines generative models and computer graphics to simulate realistic weather conditions (such as rain and fog) on sunny-weather images. The authors also introduced a pseudo-supervised loss function, which exploits the relationship between augmented and unaugmented data to reinforce the robustness of depth estimation through self-supervision. This loss function is combined with the photometric loss for model training.

The authors used physics-based renderer (PBR) [49] for the fog and rain augmentations, CoMoGAN [50] to create realistic dawn, dusk and night-time augmentations and Automold [51] to create ground snow and motion blur augmentations.

One limitation is that the method requires large amounts of training data so that the depth estimation network can learn to extract meaningful features from both unaugmented and augmented images. Also, this can lead to computational expenses during training due to its requirement of training the network on both augmented and unaugmented data.

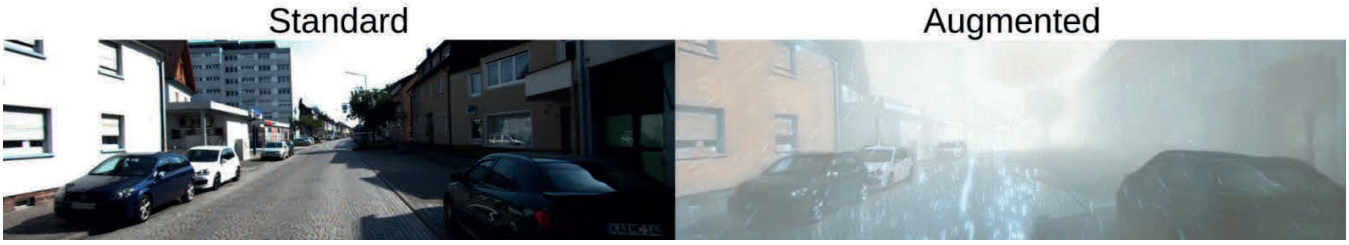

Fig. 7. Weather data augmentation [48].

Additionally, the effectiveness of this method might vary with different datasets, and it might struggle with data that is significantly different from its training set.

Additionally, Wang et al. [27] proposed a self-supervised method called WeatherDepth. It addresses this problem by using curriculum contrastive learning. The proposed method uses three simple to complex curricula, starting from sunny scenes with small changes in contrast, saturation and brightness, then moving on to more complex weather scenes with ground snow, groundwater reflections and droplets, and then moving on to more adverse weather conditions like raindrops, streaks of snow, veiling effect and rain.

Fig. 8. Depth map from KITTI dataset Geiger et al. [11].

Fig. 9. Dense disparity map from ApolloScape dataset Huang et al. [52].

The authors use contrastive learning to train the depth estimation model to differentiate between negative pairs and positive pairs of data where the images of various weather conditions of the same scene are the positive pairs of data and the images of different scenes are the negative pairs of data.

The authors also use an adaptive curriculum scheduler to change the difficulty of the training data, while the model is being trained. The scheduler uses a self-supervised loss, so if the model improves, the loss decreases rapidly, and if the model does not improve, the loss will decrease slowly. When the loss decreases rapidly, the scheduler introduces more complex images to the training data, and the complexity of the training data will remain the same if the loss decreases slowly. The proposed method achieves good results, it significantly increases the robustness of the model to complex weather conditions and outperforms other SoTA methods as well.

Recently, Tang et al. [47] have proposed a method for foggy weather simulation. The proposed approach leverages a self-supervised network to predict a depth map. Then the absolute depth information is derived using dense geometric constraints. A transmittance map is generated based on the simulated image visibility. A dark channel map is then utilised to identify sky regions and estimate atmospheric illumination. Finally, an atmospheric scattering model is utilised to generate simulations of fog in the image according to the specified visibility conditions. The authors have achieved good results; however, there are a few limitations in the depth estimation network such as the depth information on the edges of the depth map being less precise, unexpected variations of depth of areas that are supposed to be stable etc. The proposed method produces consistent results, which proves that this method is an effective simulation method and it can be used in other areas in the future.

VI. Datasets

Depth estimation has gained a lot of attention over the past few years and some new datasets have been made as well. The most widely used datasets are mentioned below.

A. KITTI

KITTI [11] is the most widely used benchmark dataset for evaluating the performance of depth estimation models. It contains 56 scenes and stereo images, LIDAR depth maps of resolution of 1242×375, and few data splits for this dataset have been proposed. One of the most common data splits is the one proposed by Eigen et al. [13] known as the Eigen Split where 28 scenes are used for training, and 28 scenes are used for testing.

B. NYU Depth-V2

NYU Depth-V2 dataset [53] is one of the most widely used indoor datasets for MDE. It contains 1449 RGB and RGB-D image pairs of size 640×480 and contains 26 scene classes and 464 various scenes across those classes.

C. DrivingStereo

The DrivingStereo dataset [12] contains over 180k images, which have been cropped to 1762×800, and it covers a wide range of driving scenes. The dataset contains stereo images along with disparity and depth maps, which have been captured using LIDAR sensors, and it covers various weather conditions such as sunny, cloudy, rainy, dusky and foggy weather.

D. Cityscapes

Cordts et al. [54] proposed a dataset, which contains a large number of stereo images that were collected from 50 cities; 5000 of the images contain pixel-level annotations and 20 000 course annotations.

E. Make3D

The Make3D dataset [9], [55] is another widely used outdoor dataset that contains 534 RGB-D image pairs with a 1704×2272 image resolution and a 55×305 depth map resolution. The training set contains 400 of these pairs, and the testing set contains 134 pairs.

F. nuScenes

The nuScenes dataset [10] contains night time images and rainy images as well, and these can be useful to evaluate the performance of MDE models in night and rainy images. This dataset contains 1000 scenes with about 1.4 million images and 390k LIDAR sweeps.

G. ApolloScape

The ApolloScape [52] is another dataset used for autonomous driving, and it also can be used for MDE as it provides 5165 total images and dense disparity map pairs where the training data contain 4156 of these pairs, and the testing data contains 1009 pairs. This dataset provides dense disparity maps, which were obtained by collecting 3D point clouds using LIDAR sensors and then fitting 3D CAD models to each moving car, which was taken from the 3D car instance dataset.

VII. Evaluation Metrics

The most widely used evaluation metrics for MDE are the metrics proposed by Eigen et al. [13], and these metrics include error metrics, which are the absolute relative difference (Abs-Rel), root mean square error (RMSE), square relative error (Sq-Rel) and RMSE-log. The proposed metrics also include accuracy metrics which are $\delta < 1.25^t$ and here the $t = 1, 2, 3$. The formulas for these metrics are given below.

$$\text{Abs-Rel} = \frac{1}{T}\sum_{i=T}\frac{d_i - d_i^{gt}}{d_i^{gt}} \quad (1)$$

$$\text{Sq-Rel} = \frac{1}{T}\sum_{i=T}\frac{||d_i - d_i^{gt}||^2}{d_i^{gt}} \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{i=T}||d_i - d_i^{gt}||^2} \quad (3)$$

$$\text{RMSE-log} = \sqrt{\frac{1}{T}\sum_{i=T}||\log(d_i) - \log(d_i^{gt})||^2} \quad (4)$$

$$\text{Accuracy} = \%\text{ of } d_i \text{ s.t. } \max\left(\frac{d_i}{d_i^{gt}}, \frac{d_i^{gt}}{d_i}\right) = \delta < \text{threshold} \quad (5)$$

The results of these metrics for the complex models are presented in Table I and the results for the lightweight models are presented in Table II.

VIII. Discussion

A. Current Trends

Monocular depth estimation has undergone significant improvements in recent years, and the current trends in MDE are discussed in this section.

Transformers have become extremely popular in monocular depth estimation mainly due to their ability to extract long-range relationships, and the network can capture both local information and global context when transformers are combined with CNNs [28].

Since the standard attention mechanisms require a lot of computation, several efficient attention mechanisms have been introduced, and these attention mechanisms aim to reduce computational complexity while maintaining or even improving performance [34]–[36].

Many existing architectures use the U-Net architecture as the baseline, and since the standard skip connections in the U-Net are too simple to effectively connect features, several methods such as Skip Attention and Feature Fusion modules have been introduced to overcome this limitation [37], [38]. Some methods also utilise semantic information to provide geometric guidance and improve the depth estimation accuracy.

B. Challenges in Transformer-Based and Hybrid Architectures

CNN-transformer hybrid methods and some transformer-based methods effectively capture both local features and global context and achieve outstanding results. However, most of these architectures are complex, which makes them unsuitable for real-time inference because of their high inference time and the amount of computational resources needed [33], [44].

Many lightweight model architectures have also been proposed. However, existing lightweight MDE methods frequently face challenges in terms of their representation capacity and can require higher computational resources for image reconstruction [44], highlighting a trade-off between computational efficiency and representation accuracy in lightweight MDE approaches.

Most of the existing MDE models work extremely well in sunny weather conditions. However, they perform poorly in challenging weather conditions, and this could be mainly due to the lack of data and the reliance on simple depth cues [27], [48].

Standard attention mechanisms require a lot of computation. Therefore, efficient attention mechanisms could also be explored [33], [35], [36].

IX. Future Work

Most of the existing models are complex and not suitable for fast inference. Therefore, more lightweight architectures and model pruning techniques could be explored [33].

Explainability and interpretability of MDE models can be extremely important as they can help identify the specific features influencing the depth predictions of the model. Existing transformer-based and hybrid architectures are difficult to interpret. Therefore, model explainability and transparency could be explored in the future to make the model predictions more understandable for the user [31].

TABLE I
EVALUATION METRICS OF THE COMPLEX MODELS TESTED ON THE KITTI BENCHMARK DATASET [11]

Model	Abs-Rel	Sq-Rel	RMSE	RMSE-log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
MonoViT [28]	0.099	0.708	4.372	0.175	0.900	0.967	0.984
SwinDepth [34]	0.106	0.739	4.510	0.182	0.890	0.964	0.984
ROIFormer [41]	0.096	0.616	4.148	0.169	0.905	0.969	0.986
DwinFormer [29]	0.047	-	1.959	-	0.980	-	-
Focal-WNet [30]	0.082	0.355	3.076	0.120	0.926	0.986	0.997
DepthFormer [31]	0.052	0.158	2.143	-	0.975	0.997	0.999
Trap-S [35]	0.055	0.177	2.278	0.085	0.967	0.996	0.999
DAttNet [36]	0.064	0.270	2.895	0.103	0.947	0.991	0.998
PixelFormer [37]	0.051	0.149	2.081	-	0.976	0.997	0.999
SAU-Net [38]	0.108	0.745	4.598	0.183	0.891	0.964	0.985
PCTDepth [33]	0.053	0.192	2.282	0.079	0.965	0.994	0.999
EMTNet [42]	0.082	0.324	2.946	0.075	0.928	0.988	0.997

TABLE II
EVALUATION METRICS OF THE LIGHTWEIGHT MODELS TESTED ON THE KITTI BENCHMARK DATASET [11]

Model	Abs-Rel	Sq-Rel	RMSE	RMSE-log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Lite-Mono [26]	0.107	0.765	4.561	0.183	0.886	0.963	0.983
METER S [45]	-	-	4.603	-	0.829	-	-
Jin et al. [25]	0.107	0.785	4.601	0.184	0.885	0.963	0.983
LightDepthNet [46]	-	-	4.729	-	0.871	-	-
SAD-Depth [44]	0.106	0.749	4.591	0.182	0.881	0.962	0.984

Most of the existing methods perform poorly in rough weather conditions [48] and various weather augmentations [27], [48] or data sets [10], [12] could be used for training. It is essential to combine the features of the encoder and decoder effectively for accurate depth estimation. Therefore, more methods for feature fusion could be further explored [31], [37], [38].

X. CONCLUSION

MDE has come a long way, and research has led to significant improvements in the accuracy and performance of MDE systems mainly by the advancements in deep learning, particularly with CNNs and the cutting-edge hybrid methods combining CNNs and transformers.

The sensors used for depth estimation offer a lot of advantages; however, these sensors can be affected by environmental factors and interference and are quite expensive as well, ultimately leading to MDE.

Deep learning-based MDE methods have surpassed traditional and machine learning-based depth estimation methods over the past couple of years, and they have achieved even better results with hybrid architectures as they have the ability to capture both local features and global context.

In this paper, we have thoroughly discussed the sensors used in MDE, the different methodologies and limitations of state-of-the-art transformer and hybrid architectures, as well as different techniques to deal with rough weather conditions. Finally, the current trends, challenges and future directions in MDE have also been addressed.

REFERENCES

[1] P. Vyas, C. Saxena, A. Badapanda, and A. Goswami, "Outdoor monocular depth estimation: A research review," *arXiv preprint arXiv:2205.01399*, May 2022. https://doi.org/10.48550/arXiv.2205.01399

[2] Q. Li *et al*., "Deep learning based monocular depth prediction: Datasets, methods and applications," *arXiv preprint arXiv:2011.04123*, 2020. https://arxiv.org/pdf/2011.04123

[3] A. Masoumian, H. A. Rashwan, J. Cristiano, M. S. Asif, and D. Puig, "Monocular depth estimation using deep learning: A review," *Sensors*, vol. 22, no. 14, Art. no. 5353, Jul. 2022. https://doi.org/10.3390/s22145353

[4] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, May 2021. https://doi.org/10.1016/j.neucom.2020.12.089

[5] Foresight, "An overview of autonomous sensors – LIDAR, RADAR, and cameras," 2023. [Online]. Available: https://www.foresightauto.com/an-overview-of-autonomous-sensors-lidar-radar-and-cameras/

[6] Y. Li and J. Ibanez-Guzman, "Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 50–61, Jul. 2020. https://doi.org/10.1109/MSP.2020.2973615

[7] J. Hasch, "Driving towards 2020: Automotive radar technology trends," in *2015 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, Heidelberg, Germany, Apr. 2015, pp. 1–4. https://doi.org/10.1109/ICMIM.2015.7117956

[8] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Science China Technological Sciences*, vol. 63, no. 9, pp. 1612–1627, June 2020. https://doi.org/10.1007/s11431-020-1582-8

[9] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," in *IJCAI-07*, 2007, pp. 2197–2203. [Online]. Available: https://www.ijcai.org/Proceedings/07/Papers/354.pdf

[10] H. Caesar *et al*., "nuScenes: A multimodal dataset for autonomous driving," *arXiv preprint arXiv:1903.11027*, Mar. 2019. https://doi.org/10.48550/arXiv.1903.11027
[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, June 2012, pp. 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074
[12] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019, pp. 899–908. https://doi.org/10.1109/CVPR.2019.00099
[13] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," Adv. Neural. Inf. Process. Syst., vol. 27, 2014. https://doi.org/10.48550/arXiv.1406.2283
[14] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth International Conference on 3D Vision (3DV)*, Stanford, CA, USA, Oct. 2016, pp. 239–248. https://doi.org/10.1109/3DV.2016.32
[15] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFS," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 1119–1127. https://doi.org/10.1109/CVPR.2015.7298715
[16] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, Dec. 2018. https://doi.org/10.48550/arXiv.1812.11941
[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017, pp. 4700–4708. https://doi.org/10.1109/CVPR.2017.243
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
[19] C.-H. Yeh, Y.-P. Huang, C.-Y. Lin, and C.-Y. Chang, "Transfer2Depth: Dual attention network with transfer learning for monocular depth estimation," *IEEE Access*, vol. 8, pp. 86081–86090, May 2020. https://doi.org/10.1109/ACCESS.2020.2992815
[20] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017, pp. 270–279. https://doi.org/10.1109/CVPR.2017.699
[21] R. Garg, V. Kumar B.G., G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, Part VIII 14, Oct. 2016, pp. 740–756. https://doi.org/10.1007/978-3-319-46484-8_45
[22] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "Towards real-time unsupervised monocular depth estimation on CPU," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, Oct. 2018, pp. 5848–5854. https://doi.org/10.1109/IROS.2018.8593814
[23] J. Liu, Q. Li, R. Cao, W. Tang, and G. Qiu, "MiniNet: An extremely lightweight convolutional neural network for real-time unsupervised monocular depth estimation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 255–267, Aug. 2020. https://doi.org/10.1016/j.isprsjprs.2020.06.004
[24] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), Oct. 2019, pp. 3828–3838. https://doi.org/10.1109/ICCV.2019.00393
[25] J. Jin, B. Tao, X. Qian, J. Hu, and G. Li, "Lightweight monocular absolute depth estimation based on attention mechanism," *Journal of Electronic Imaging*, vol. 33, no. 2, Mar. 2024, Art. no. 23010. https://doi.org/10.1117/1.JEI.33.2.023010
[26] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-Mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, June 2023, pp. 18537–18546. https://doi.org/10.1109/CVPR52729.2023.01778
[27] J. Wang *et al*., "WeatherDepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions," *arXiv preprint arXiv:2310.05556*, Oct. 2023. https://doi.org/10.48550/arXiv.2310.05556
[28] C. Zhao *et al*., "MonoViT: Self-supervised monocular depth estimation with a vision transformer," in *2022 International Conference on 3D Vision (3DV)*, Prague, Czech Republic, Sep. 2022, pp. 668–678. https://doi.org/10.1109/3DV57658.2022.00077
[29] M. A. Rahman and S. A. Fattah, "DwinFormer: Dual window transformers for end-to-end monocular depth estimation," *IEEE Sensors Journal*, vol. 23, no. 18, Aug. 2023. https://doi.org/10.1109/JSEN.2023.3299782
[30] G. Manimaran and J. Swaminathan, "Focal-WNet: An architecture unifying convolution and attention for depth estimation," in *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, Mumbai, India, Apr. 2022, pp. 1–7. https://doi.org/10.1109/I2CT54291.2022.9824488
[31] Z. Li, Z. Chen, X. Liu, and J. Jiang, "DepthFormer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *Machine Intelligence Research*, vol. 20, no. 6, pp. 837–854, Dec. 2023. https://doi.org/10.1007/s11633-023-1458-0
[32] A. Dosovitskiy *et al*., "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, Oct. 2020. https://doi.org/10.48550/arXiv.2010.11929
[33] C. Xia et al., "PCTDepth: Exploiting parallel CNNs and transformer via dual attention for monocular depth estimation," *Neural Processing Letters*, vol. 56, no. 2, Feb. 2024, Art. no. 73. https://doi.org/10.1007/s11063-024-11524-0
[34] D. Shim and H. J. Kim, "SwinDepth: Unsupervised depth estimation using monocular sequences via Swin transformer and densely cascaded network," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, United Kingdom, May 2023, pp. 4983–4990. https://doi.org/10.1109/ICRA48891.2023.10160657
[35] C. Ning and H. Gan, "Trap attention: Monocular depth estimation with manual traps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, June 2023, pp. 5033–5043. https://doi.org/10.1109/CVPR52729.2023.00487
[36] A. Astudillo, A. Barrera, C. Guindel, A. Al-Kaff, and F. García, "DAttNet: monocular depth estimation network based on attention mechanisms," *Neural Computing and Applications*, vol. 36, no. 7, pp. 3347–3356, Dec. 2023. https://doi.org/10.1007/s00521-023-09210-8
[37] A. Agarwal and C. Arora, "Attention attention everywhere: Monocular depth prediction with skip attention," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, HI, USA, Jan. 2023, pp. 5861–5870. https://doi.org/10.1109/WACV56688.2023.00581
[38] W. Zhao, Y. Song, and T. Wang, "SAU-Net: Monocular depth estimation combining multi-scale features and attention mechanisms," *IEEE Access*, vol. 11, Dec. 2023, pp. 137734–137746. https://doi.org/10.1109/ACCESS.2023.3339152
[39] Z. Liu *et al*., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, Oct. 2021, pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986
[40] 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: 18th International Conference*, part III 18, Munich, Germany, Oct. 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
[41] D. Xing, J. Shen, C. Ho, and A. Tzes, "ROIFormer: semantic-aware region of interest transformer for efficient self-supervised monocular depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2983–2991. https://doi.org/10.1609/aaai.v37i3.25401
[42] L. Yan, F. Yu, and C. Dong, "EMTNet: efficient mobile transformer network for real-time monocular depth estimation," *Pattern Analysis and Applications,* vol. 26, no. 4, pp. 1833–1846, Oct. 2023. https://doi.org/10.1007/s10044-023-01205-4
[43] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 2020, pp. 1580–1589. https://doi.org/10.1109/CVPR42600.2020.00165

[44] L. Song *et al*., "Spatial-aware dynamic lightweight self-supervised monocular depth estimation," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 883–890, Nov. 2023. https://doi.org/10.1109/LRA.2023.3337991
[45] L. Papa, P. Russo, and I. Amerini, "METER: a mobile vision transformer architecture for monocular depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5882–5893, Mar. 2023. https://doi.org/10.1109/TCSVT.2023.3260310
[46] Q. Liu and S. Zhou, "LightDepthNet: Lightweight CNN architecture for monocular depth estimation on edge devices," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 4, pp. 2389–2393, Nov. 2023. https://doi.org/10.1109/TCSII.2023.3337369
[47] M. Tang, Z. Zhao, and J. Qiu, "A foggy weather simulation algorithm for traffic image synthesis based on monocular depth estimation," *Sensors*, vol. 24, no. 6, Mar. 2024, Art. no. 1966. https://doi.org/10.3390/s24061966
[48] 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*, Paris, France, Oct. 2023, pp. 8907–8917. https://doi.org/10.1109/ICCV51070.2023.00818
[49] M. Tremblay, S. S. Halder, R. de Charette, and J. F. Lalonde, "Rain rendering for evaluating and improving robustness to bad weather," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 341–360, Feb. 2021. https://doi.org/10.1007/s11263-020-01366-3
[50] F. Pizzati and R. de Charette, "CoMoGAN: continuous model-guided image-to-image translation", [Online]. Available: https://github.com/cv-rits/CoMoGAN. Accessed on: Jul. 04, 2024.
[51] U. Saxena and R. Giriraj, "Automold--Road-Augmentation-Library," GitHub, Feb. 12, 2023. [Online]. Available: https://github.com/UjjwalSaxena/Automold--Road-Augmentation-Library
[52] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020. https://doi.org/10.1109/TPAMI.2019.2926463
[53] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision*, Part V 12, Florence, Italy, Oct. 2012, pp. 746–760. https://doi.org/10.1007/978-3-642-33715-4_54
[54] M. Cordts *et al*., "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016, pp. 3213–3223. https://doi.org/10.1109/CVPR.2016.350
[55] A. Saxena, S. Chung, and A. Ng, "Learning depth from single monocular images," *Neural Information Processing Systems (NIPS)*, vol. 18, pp. 1–8, Dec. 2005.

Lakindu Kumara received a B. Sc. (Hons) degree in Artificial Intelligence and Data Science from the Informatics Institute of Technology (IIT), Colombo Sri Lanka affiliated with the Robert Gordon University (RGU), Aberdeen, Scotland. He has worked as an AI Intern at Kingslake Engineering Systems. He was involved in several AI projects. His research on monocular depth estimation during his degree and his research interests include artificial intelligence, data science and autonomous driving.
E-mail: lakindukumara2003@gmail.com
ORCID iD: https://orcid.org/0009-0003-0422-1740

Nipuna Senanayake received his M. Sc. in Computer Science from Georgia State University, USA, in 2018 and his B. Sc. (Hons) is from the University of Kelaniya, Sri Lanka in 2013. He is currently a Senior Lecturer at the Informatics Institute of Technology (IIT), Colombo, Sri Lanka. He has published his works in several reputed conferences and journals, and his research interests include artificial intelligence, machine learning and computer security. Before joining IIT, Nipuna has worked as a Software Engineer at companies such as Turner Broadcasting Systems, Virtusa and hSenid Business Solutions.
E-mail: nipuna.s@iit.ac.lk

Guhanathan Poravi received a B. Sc. degree in Information Systems Management from the University of Madras, India in 2004, a PhD in Computer Science from the University of Peradeniya, Sri Lanka in 2006, an MBA in IT from the University of Moratuwa, Sri Lanka in 2008, and MBCS from BCS, the UK in 2009. From 2004 to 2006, he was a Lecturer and Assistant Manager in education delivery at NIIT, Sri Lanka. From 2007 to 2012, he worked as a Senior Software Engineer at Cambio Healthcare Systems, Sri Lanka. Since 2012, he has been working as a Senior Lecturer (Grade 1) at the Informatics Institute of Technology, Sri Lanka. His main research interests include machine learning, big data & data science, and software engineering. Mr. Poravi was a member of the Institute of Electrical and Electronics Engineers (IEEE), Institution of Engineering and Technology (IET), and British Computer Society (BCS).
E-mail: guhanathan.p@iit.ac.lk