"Depth Estimation from Single Image"

I. INTRODUCTION

The process of estimating a dense depth map for a single RGB picture is known as depth estimation from a single image (SIDE, short for Single Image Depth Estimation).To be more precise, a metric depth value has to be estimated, for each pixel in the provided RGB picture. Figure 1 displays a sample input picture and the accompanying depth map.The depth of each pixel is indicated by the colour of that pixel in the depth map: blueish pixels indicate that the pixel is closer to us, while reddish pixels indicate that the pixel is farther away.

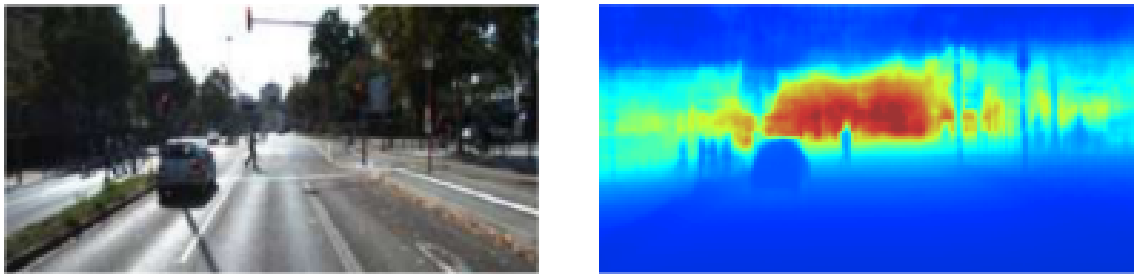

Fig. 1. Input RGB image and the depth map estimated by the neural network of Fu et al. [1].

The inherent ambiguity of the SIDE issue is what makes it intriguing and difficult. A single 2D image may be created from an infinite variety of various 3D scenes.This implies that depth maps are generated from RGB pictures in a one to many mapping. If so, how do people assess depth from monocular pictures, despite the fact that their visual systems much outperform those made by humans in terms of quality and generalisation? The cues that people use to accomplish SIDE are where the solution to this question may be found.

Lack of perspective shifts in the scenes and frames of the incoming picture data presents the main obstacle to outdoor monocular depth estimation [2] [3]. This occurs because, in contrast to when the same camera is used inside, when the subject of the image is significantly larger than the camera's focal view size, there is a much smaller overall change between the elements taken in a sequence of frames. Due to the absence of dynamic nature, traditional approaches that rely on motion and texture clues for depth fail [4] [5][6]. A timeline of the development of depth estimate methods is shown in Fig. 2.

Along with this problem, there aren't any publicly accessible outdoor datasets for model training and assessment [8]. Research in the domains of robotics and 3D reconstruction, both of which are mostly indoor activities in their current forms and industrial applications, heavily use depth estimation datasets. Because of this, there are extremely few outdoor datasets, and those that target very long-range data are even more scarce.

The requirement to record or artificially create scenes and their depth data, create deep learning-based models that incorporate classical methods where appropriate, and apply various learning approaches to improve on them is what motivates the present study in depth estimation [7].The depth estimation study is the main focus of this work, which also surveys recent breakthroughs and trends in deep learning-based methodologies. We also highlight potential future research avenues and take a look at the limits of recent studies

Objective:

- We look at various approaches to the depth estimation issue.
- Our primary concern is the single-image depth estimation problem.
- Due to its characteristics, machine learning techniques are currently the most effective way to solve the single image depth estimation problem,most successfully with convolutional neural networks.
- Additionally, we look into multitask strategies that combine the depth estimation problem with related work like augmented reality,semantic segmentation and surface normal estimation.

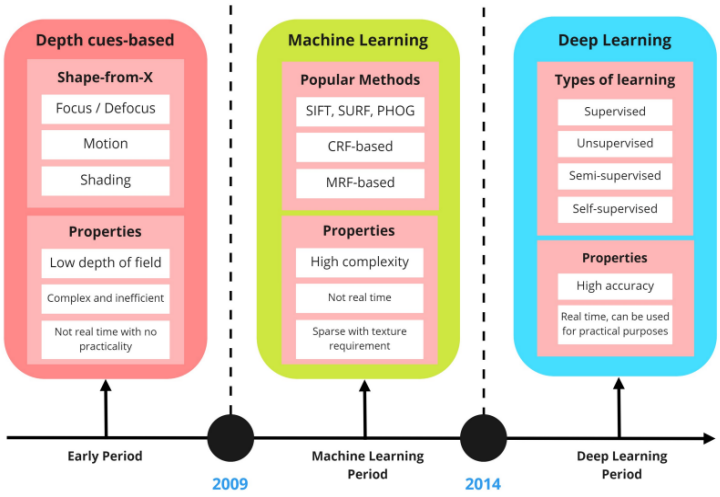

Fig. 2. Techniques for estimating depth have changed over time.. Inspired from[7]

II. RELATED WORKS

The SIDE problem was not immediately solved in the early days of the field. The authors of the historical book Hoeim et al[9] aimed to automatically create a 3D scene from the supplied RGB image to create a virtual environment. Their method evokes the idea that outdoor spaces are primarily composed of the ground plane, the sky, and vertical objects protruding from the ground. Using manually generated stimuli, they categorize superpixels into one of three types. Then, by placing the objects on the base plane vertically, they automatically generate a virtual environment using the three classes and the assumption mentioned above. Some features are omitted because the components of the derived scene are so basic that they resemble a picture that emerges from a child's In this work, we pursue the integration of semantic segmentation and the separation of indoor and outdoor data. The work of Michels et al[10] is another early work that uses SIDE to solve a specific problem. A high-speed remote control car must navigate obstacles in an uncontrolled outdoor environment. The framework created is divided into two parts: a vision part that simulates a 2D laser scanner and calculates the distance to the nearest obstacle in each direction, and a reinforcement learning part that navigates the vehicle around obstacles so we have binocular vision in it. The vision system is trained using linear regression with hand-crafted features under supervision. The input image is divided into vertical strips. Each strip is marked with the distance of the nearest obstacle in the log space. The distance between the estimated depth and the actual terrain depth in the log space is one of them. Additionally, the relative depth error is used as an error metric, which subtracts the mean from the true and estimated depth values in log space. In addition, synthetic data with varying degrees of realism are used to increase the effectiveness of the system. The vision problem is formulated significantly differently because the goal is to avoid obstacles. Further work by Saxena et al. from 2008 [11] had a significant impact on the field because it brought a very important assumption. Scenes consist of small planar surfaces, and the 3D location and orientation of each surface can be used to determine the depth of each pixel that belongs to it. This essentially means that 3D positioning and orientation of small surfaces can be used to display even the most complex 3D scenarios. The truth of this premise is demonstrated by graphic engines that make it possible to create several complicated models from straight triangular surfaces. They create these tiny surfaces by super pixelating an RGB image with the expectation that neighboring pixels will have a similar appearance and belong to the same surface. Again, in this effort, they use the MRF model and direct its training. Since additional constraints can be applied to depth estimation depending on these factors, it is crucial to capture these attributes. Fractional (relative) depth error is used when applying these constraints. It is expressed as (dˆ−d)/d, where the predicted depth is dˆ.and d represents the actual depth value. They further develop their work by detecting objects and using previous knowledge to more accurately estimate the depth of detected objects. For example, they can detect a person and predict that they are connected to the ground, or they can detect two people and use the pixels that make up each person's head to more accurately estimate the depth of each person. We follow a key assumption in this work that facilitates the simplification of the SIDE problem. Increasing the number of pixels of the input image and assuming coplanarity. Many researchers superpixel the input image and estimate the superpixel depth based on the assumption that a 3D scene can be described by small planar surfaces. The advantage of this assumption is the minimization of the number of

estimation points, which reduces the calculation cost. Superpixels with similar RGB values are more likely to be found together. Typically, this assumption is used to smooth the model estimate.

III. PROPOSED METHOD

We will look at efforts that are primarily related to improving the results of the metrics. Naturally, some works simply apply new extensions and developments to existing approaches while closely imitating earlier works. However, some works use human knowledge to improve performance. We will also look at methods that combine multitasking learning to achieve better results.

DEEP LEARNING APPROACHES:-

Convolutional Neural Network:-

A convolutional neural network [12] is a deep learning algorithm that can take an input image, assign importance (learnable weights and biases) to different aspects/objects in the image, and be able to distinguish one from the other. The preprocessing required in ConvNet is much lower compared to other classification algorithms. While in primitive methods the filters are created manually, with enough training ConvNets have the ability to learn these filters/characteristics. Compared to earlier approaches, Eigen et al. [13] to introduce CNN to the SIDE problem in 2014 and obtain reasonably high performance. At that time, CNN-based solutions already produced results that were largely suitable for a variety of vision problems. In order to solve the SIDE problem, Eigen et al. draw on their previous expertise with CNN and combine it with an understanding of the specific challenge. They formulate the problem as a supervised regression problem, which they then solve using their framework, which consists of two networks – a coarse network and a fine network. Convolutional layers and fully connected layers are located near the conclusion of the coarse mesh. The Fine network only considers local regions of the image and is a fully convolutional network.

$$loss = \frac{1}{2n}\sum_{i=1}^{n}(\log \hat{d}_i - \log d_i + \frac{1}{n}\sum_{i}(\log d_i - \log \hat{d}_i))^2 \quad (3)$$

$$= \frac{1}{2n^2}\sum_{i,j}((\log \hat{d}_i - \log \hat{d}_j) - (\log d_i - \log d_j))^2 \quad (4)$$

$$= \frac{1}{n}\sum_{i}(\log \hat{d}_i - \log d_i)^2 - \frac{1}{n^2}\sum_{i,j}((\log \hat{d}_i - \log d_i)(\log \hat{d}_j - \log d_j)) \quad (5)$$

Convolutional neural networks are primarily used on images with principal component convolutional layers, pooling layers (max pooling and average pooling) and activation functions, which together enable these networks to learn 2D spatial properties of input images. CNNs are used to extract depth features from images, reduce the size of these extractions using pooling layers, and reconstruct depth maps using their activation functions and FC layers.

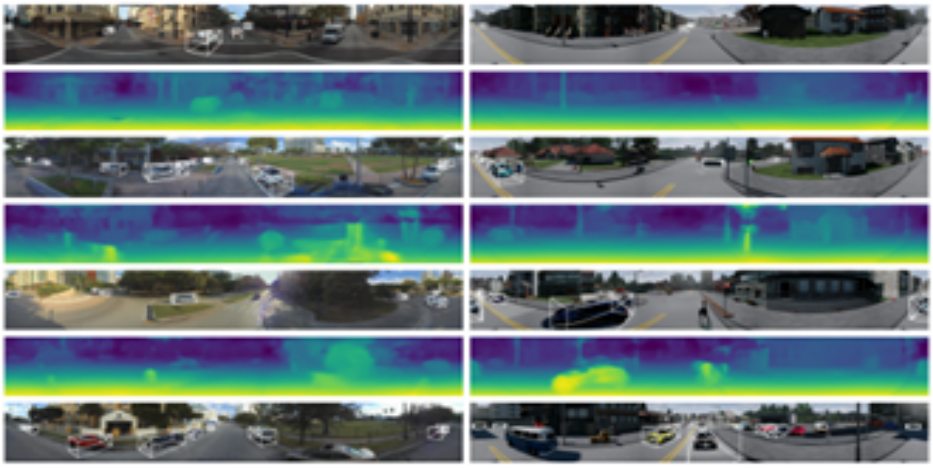

Fig. 3. Depth recovery from panoramic imagery using the approach described in [14]

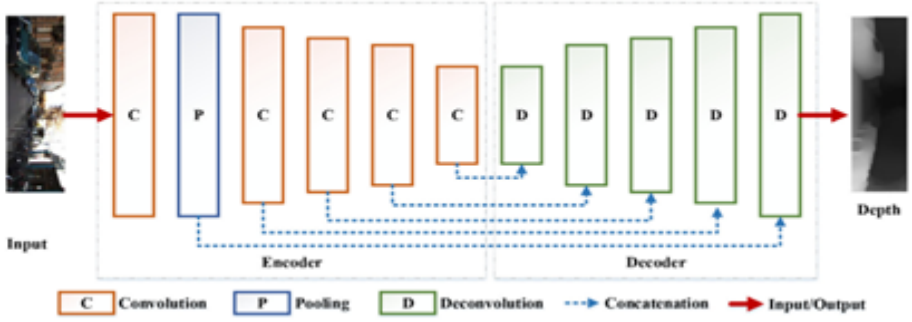

Fig. 4. The general pipeline of deep learning for monocular depth estimation using CNNs. Source: [15]

Recurrent Neural Network:-

Monocular depth estimation uses RNNs, memory-capable inter-sequence models, to extract temporal information from video sequences.

An RNN consists of three parts: an input unit, an output unit, and a hidden unit. The outputs of both the current input unit and the previously hidden unit form the input of the hidden unit. Feedback allows an LSTM, a particular kind of RNN, to learn the temporal dependencies between data points. This can be exploited by using video-based datasets, where LSTMs can be fed progressive frames and depth maps can then be derived.

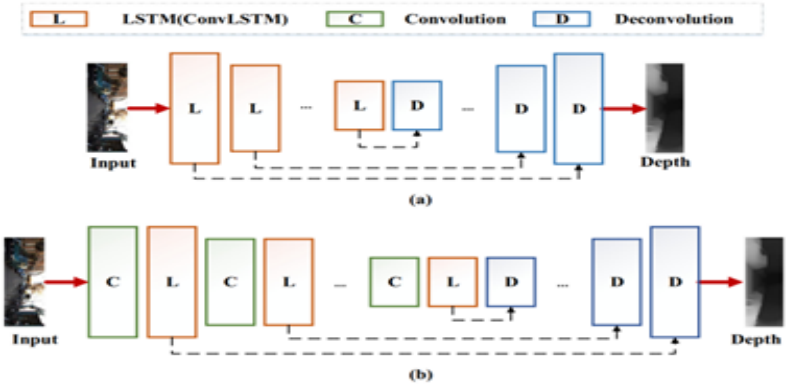

Fig.The two major architectures of RNN based networks for predicting monocular depth maps. (a) shows the architecture containing only LSTMs in the encoder while (b) uses convolutions in addition to the LSTMs.

DEEP LEARNING TRAINING:-

Supervised Learning:-

Using Ground Truth depth maps, supervised learning networks are developed for monocular depth estimation. The learning objective is to penalize the mismatch between the ground truth depth map and the prediction using the inverse Huber function and depth-based logarithmic loss (Berhu)[16]. The predicted depth value should be as close to GT as possible to allow the depth model to converge. Supervised learning networks for monocular depth estimation are built using Ground Truth depth maps. An inverse Huber function and a log depth-based loss function are used to penalize the difference between the prediction and the truth depth map (Berhu). In order for the depth model to converge, the projected depth value should be as close as possible to the GT.

Unsupervised Learning:-

Publicly available high-resolution datasets still require a lot of work and money. Therefore, for monocular depth estimation without using GT depth maps, researchers are looking for unsupervised deep learning techniques. Unsupervised monocular image depth estimation is often tested on monocular images after being trained on paired stereo images or sequences of monocular images.

Stereo matching and the use of monocular sequences are the two most used techniques for training unsupervised MDE models.

A well-known example of the former is [17], which builds a depth estimation system using a conventional belief propagation method. [18] propose a CNN-based design in which the model captures depth information from right and left viewpoints. Using monocular sequences is a different method for training networks based on unsupervised learning. Since monocular depth estimation data sets are more readily available and easier to obtain, this is an extremely attractive study problem. It also remains free of the projection and left/right source mapping issues that stereo pairing causes.

IV. METHODOLOGY

There are several works that incorporate segmentation into the overall depth estimation framework for supervised and self-supervised learning, however these two are considered as discrete modules rather than a framework that is dependent on one another. As a result, training these models now requires more compute, based on the use case, and creating/tuning them accordingly.A promising area is in the direction of integrating these models of work.

This study examines the publicly accessible datasets, machine learning algorithm techniques, deep learning, and describes our training procedures. Additionally, this paper covers the limitations of prominent methodologies (deep learning, machine learning) this semester as well as how well they perform in various settings. In the end, we choose the method that will accurately estimate the depth of the image.

References:

[1] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011.

[2] Zhou, Keyang, Kaiwei Wang, and Kailun Yang. ”PADENet: An efficient and robust panoramic monocular depth estimation network for outdoor scenes.” 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020

[3] Reza, Md Alimoor, Jana Kosecka, and Philip David. ”FarSight: LongRange Depth Estimation from Outdoor Images.” 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.

[4] Tsai, Yi-Min, Yu-Lin Chang, and Liang-Gee Chen. ”Block-based vanishing line and vanishing point detection for 3D scene reconstruction.” 2006 international symposium on intelligent signal processing and communications. IEEE, 2006.

[5] Tang, Chang, Chunping Hou, and Zhanjie Song. ”Depth recovery and refinement from a single image using defocus cues.” Journal of Modern Optics 62.6 (2015): 441-448.

[6]] Zhang, Ping, et al. ”Stereoscopic video saliency detection based on spatiotemporal correlation and depth confidence optimization.” Neurocomputing 377 (2020): 256-268.

[7] Ming, Yue, et al. ”Deep learning for monocular depth estimation: A review.” Neurocomputing 438 (2021): 14-33.

[8] Zhou, Keyang, Kaiwei Wang, and Kailun Yang. ”PADENet: An efficient and robust panoramic monocular depth estimation network for outdoor scenes.” 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020.

[9] D. Hoiem, A. A. Efros, and M. Hebert, “Automatic photo pop-up,” in ACM transactions on graphics (TOG), vol. 24. ACM, 2005, pp. 577–584.

[10] J. Michels, A. Saxena, and A. Y. Ng, “High-speed obstacle avoidance using monocular vision and reinforcement learning,” in Proceedings of the 22nd international conference on Machine learning. ACM, 2005, pp. 593–600.

[11] A. Saxena, S. H. Chung, and A. Y. Ng, “Learning depth from single monocular images,” in Advances in neural information processing systems, 2006, pp. 1161–1168.

[12] S. Saha, “A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way,” *Medium*, Dec. 2018, https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[13] 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, 2014, pp. 2366–2374.

[14] de La Garanderie, Greire Payen, Amir Atapour Abarghouei, and Toby P.Breckon. ”Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery.” Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[15] Ming, Yue, et al. ”Deep learning for monocular depth estimation: A review.” Neurocomputing 438 (2021): 14-33.11.

[16] Eigen, David, and Rob Fergus. ”Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture.”Proceedings of the IEEE international conference on computer vision. 2015.

[17] Sun, Jian, Nan-Ning Zheng, and Heung-Yeung Shum. ”Stereo matching using belief propagation.” IEEE Transactions on pattern analysis and machine intelligence 25.7 (2003): 787-800.

[18] Garg, Ravi, et al. ”Unsupervised cnn for single view depth estimation: Geometry to the rescue.” European conference on computer vision. Springer, Cham, 2016.