Depth Estimation from image

ABSTRACT

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.

Method:

Multiple image methods, Single image methods, Supervised Learning based methods, Unsupervised Learning based methods

Keywords: Depth Estimation,Deep learning ,convolutional neural networks

I. INTRODUCTION

The process of estimating a dense depth map for a single RGB picture is known as depth estimation from a single imagee (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.

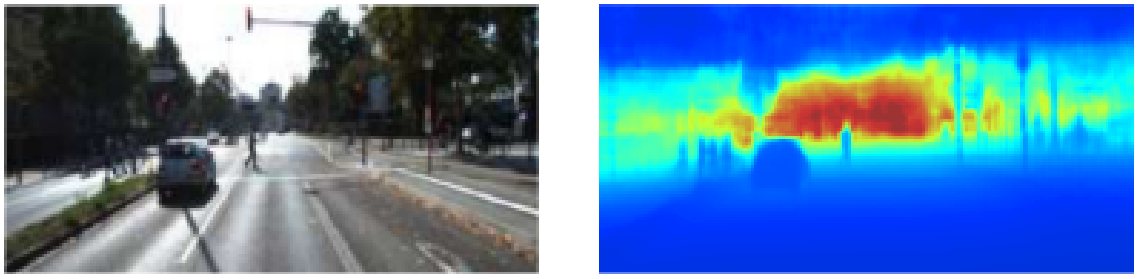

Fig. 1. Input RGB image and the depth map estimated by the neural network of Fu et al. [1].

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 [4] [5][6] fail. A timeline of the development of depth estimate methods is shown in Fig. 2.

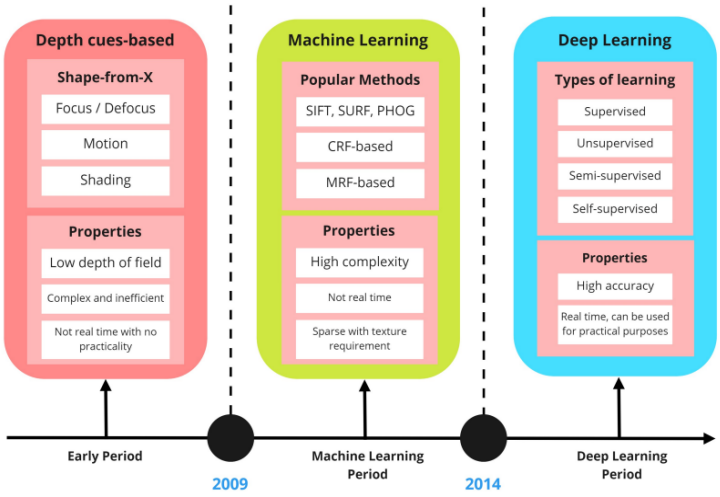

Fig. 2. Techniques for estimating depth have changed over time.. Inspired from[7]

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.

II. RELATED WORKS

The SIDE problem was not immediately addressed in the early days of the field. The writers of the historical book Hoeim et al[9] with the objective of automatically creating a 3D scene from a supplied RGB picture for building a virtual environment. Their method causes the notion that outdoor spaces primarily consist of the ground plane, the sky, and the vertical objects protruding from the ground.They categorise superpixels into one of the three types using hand-generated cues. Then, by positioning the objects on the ground plane vertically, they automatically generate the virtual environment utilising the three classes and the aforementioned supposition.There are certain features omitted since the inferred scene's components are so basic they resemble a picture that appears out of a kid's book.We observe the integration of semantic segmentation and the separation of inside and outdoor data in this work. The work of Michels et al[10] is another early piece that employs SIDE to resolve a specific issue. A high-speed remote control car must be driven through obstacles in an uncontrolled outdoor setting. The created framework is divided into two sections: a vision section that simulates a 2D laser scanner and calculates the distance to the closest obstacle in each direction, and a reinforcement learning section that navigate the vehicle around obstacles,so that we have binocular vision in it.The vision system is trained using linear regression with handcrafted features in a supervised manner.The input image is divided into vertical strips. Each strip is labeled with the nearest obstacle's distance in log space.The distance between the estimated depth and the ground truth depth in log space is one of them.Additionally, relative depth error is employed as an error metric, which subtracts the mean from actual and estimated depth values in log space. Additionally, to increase the effectiveness of the system, synthetic data with varying degrees of realism are used.The vision problem is formulated significantly differently because the aim at hand is to avoid obstacles. Another piece of work by Saxena et al. from 2008 [11] had a significant impact on the field because it made a very important assumption. Scenes are made up of tiny planar surfaces, and each surface's 3D location and orientation may be used to determine the depth of every pixel that belongs to it. This basically means that the 3D location and orientation of small surfaces can be used to depict even the most complicated 3D scenarios. The veracity of this premise is demonstrated by graphics engines, which enable the creation of several complicated models from

straightforward triangle surfaces. They produce these tiny surfaces by superpixelating the RGB image with the anticipation that neighbouring pixels will have a similar appearance and belong to the same surface.They once more utilize the MRF model in this endeavour and manage its training. As further restrictions can be applied for depth estimate depending on these factors, it is crucial to capture these attributes.When applying these limitations, fractional (relative) depth error is applied. It is expressed as (dˆ−d)/d,where the anticipated depth is dˆ.and the d represents the actual depth value.They further develop their work by detecting objects and using prior knowledge to more accurately estimate the depth of detected objects. For example, they might detect a human and anticipate that it is connected to the ground, or they might detect two humans and use the pixels that make up each of their heads to more accurately estimate the depth of each one.We observe a key premise in this work that facilitates the simplification of the SIDE problem. Increasing the input image's pixel count and assuming coplanarity assumption. Many scientists superpixelate the input image and estimate the depth of the superpixels based on the presumption that a 3D scene may be described with small planar surfaces.This assumption has the advantage of minimising the number of estimation points, which lowers the computational cost. Superpixels with similar RGB values are more likely to be found together. Usually, this presumption is used to smooth the model's estimation.

III. PROPOSED METHOD

We look at the efforts that are primarily concerned with improving metric results. Naturally, some works simply apply new expansions and developments to the existing approaches while closely imitating earlier works. However, some works utilise human expertise to improve performance. We also look at methods that combine the learning of many tasks in order to achieve better results.

DEEP LEARNING APPROACHES:-

A. **Convolutional Neural Network:-**

A Convolutional Neural Network [12] is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the ability to learn these filters/characteristics. When compared to earlier approaches, Eigen et al. [13] introduce CNNs to the SIDE issue in 2014 and get reasonably high performance. At the time, CNN-based solutions were already producing outcomes that were largely suitable for a variety of vision issues. In order to solve the SIDE problem, Eigen et al. draw on their prior expertise with CNNs and combine it with the understanding of the particular challenge at hand. They formulate the issue as a supervised regression issue, which they then resolve using their framework, which consists of two networks—a coarse network and a fine network. Convolutional layers and completely connected layers are found near the conclusion of the coarse network. The Fine network only takes into account the local regions of the image and is a fully convolutional network.

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

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

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

Convolutional neural networks are primarily used on images with the major components

convolutional layers, pooling layers (max pooling and average pooling), and activation functions which together allow these networks in learning 2D spatial features of input images.CNN's are used for extracting depth features from images, reducing the size of these extractions using the pooling layers, and reconstructing the depth maps using their activation functions and FC layers.

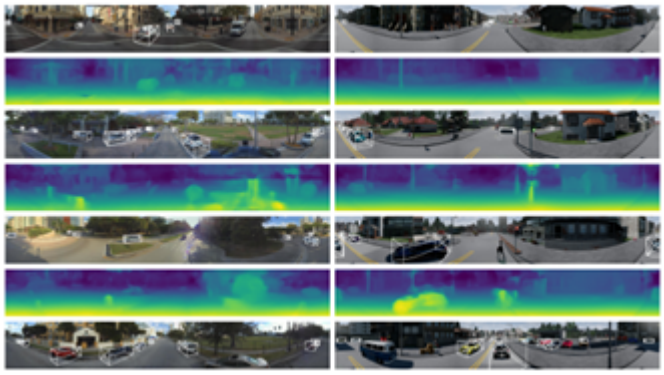

Fig. 3. Depth recovery from panoramic imagery using the approach described in [14]

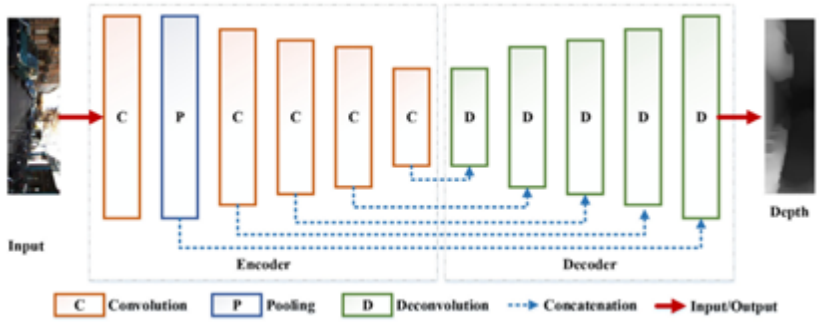

Fig. 4. The general pipeline of deep learning for monocular depth estimation using CNNs. Source: [15]

B. Recurrent Neural Network:-

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

The RNN is composed of three parts: an input unit, an output unit, and a hidden unit. The outputs of both the present input unit and the previously concealed unit make up the input of a hidden unit. Feedback connections enable LSTMs, a particular kind of RNNs, to learn the temporal dependencies between data points. This can be taken advantage of by leveraging video-based datasets, where the LSTM networks can be fed advancing frames, and depth maps can then be derived.

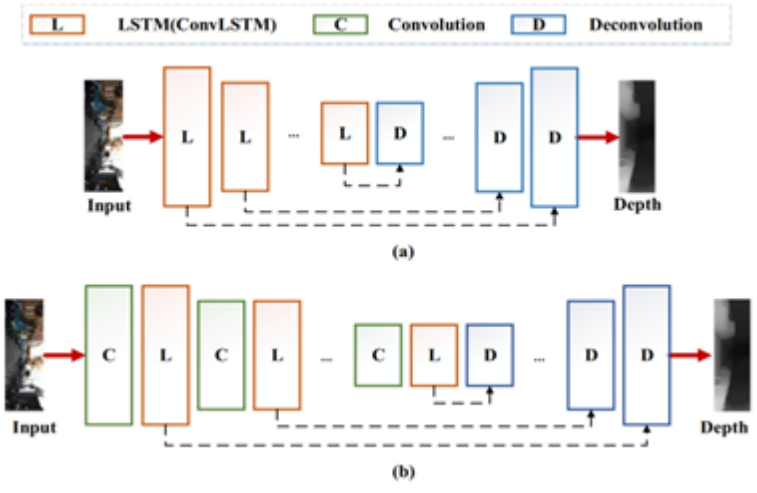

Fig.The two major architectures of RNN based networks for predicting monocular depth maps. (a) shows the architecture containing only LSTMs in the encoder while (b) uses convolutions in addition to the LSTMs.

DEEP LEARNING TRAINING:-

A. Supervised Learning:-

Using the Ground Truth depth maps, supervised learning networks for monocular depth estimation are developed. The goal of learning is to penalise the discrepancy between the ground truth depth map and the prediction using an inverse Huber function and the log depth-based loss function (Berhu)[16]. The anticipated depth value should be as close to GT as feasible for the depth model to converge. Supervised learning networks for monocular depth estimation are created using the Ground Truth depth maps. An inverse Huber function and a loss function based on log depth are used to penalise the difference between the prediction and the ground truth depth map (Berhu). For the depth model to converge, the projected depth value should be as close to GT as is practical.

B. Unsupervised Learning:-

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

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 the conventional belief propagation method. [18] suggest a CNN-based design in which the model picks up depth information from the right and left viewpoints.Use of monocular sequences is a different method for training unsupervised learning-based networks. Because monocular depth estimate datasets are more readily available and easier to obtain, this is an extremely appealing study issue. It also stays clear of the projection and left-right source mapping problems that stereo matching raises.

C. Self-Supervised Learning:-

Since the depth value in real-world applications is far higher than the value that these neural networks can reliably produce, a good depth representation will significantly boost performance. Therefore, In depth learning and self-supervised monocular motion, it is crucial to select the right depth representation to assist feature representation learning. A machine learning technique is called SSL (self-supervised learning). It obtains its knowledge from sample data that are not labelled. This type of learning falls somewhere in the middle of supervised and unsupervised learning.

The learning process has two phases. Pseudo-labels, which help with the initialization of network weights, are first used to tackle the problem. Second, the assignment is finished using either supervised or unsupervised learning. Self-supervised learning has produced impressive results recently. The fundamental advantage of SSL is that it enables training with lower-quality data rather than emphasising increasing results. The availability of several long-range datasets for outdoor depth measurement makes it challenging to build generalizable deep learning models. Here, self-supervised learning enters the picture by creating fresh data to train on using the sparsely annotated portion of the dataset.

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] 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.