Depth Estimation of Monocular Images using Transfer Learning based Unet Model

Suchitra A. Patil
Electronics and Telecommunication Engineering
Pillai College of Engineering, Ramrao Adik Institute of Technology
D. Y. Patil Deemed to be University
New Panvel, India
spatil@mes.ac.in

Chandrakant Gaikwad
Electronics and Telecommunication Engineering
Ramrao Adik Institute of Technology,
D. Y. Patil Deemed to be University
Nerul, India
cjgaikwad@gmail.com

***Abstract*—The problem of depth estimation using monocular images is very challenging contrasted to methods for estimating depth that involve several pictures, like stereo depth perception. Previous studies in this field have usually focused on utilizing geometrical priors or relied on other data collection techniques. Various machine learning methods, notably deep convolutional neural networks (CNN) integrated with artificial intelligence (AI) approaches, have recently produced new marks for a variety of visual applications.**

In this paper, a convolution neural network is used for estimating high-resolution depth image. We have used a pre-trained model DenseNet-169 which is trained on ImageNet. The encoder-decoder model used is a simple Unet model, used in biomedical image analysis. The proposed model is more accurate and efficient with reduced complexity in terms of the fewer parameters used for training. This model is also noteworthy when comparing a state of art and qualitatively it performs well and captures better edges and corners of the depth map, which is the most important factor in the depth estimation.

***Index Terms*—Convolutional neural network, Deep learning, Unet, Pretrained model, Monocular Depth estimation**

I. Introduction

Measurement of 3D depth information from monocular 2D images is a fundamental and important task in machine learning that may be put to a variety of applications, including simultaneous localization and mapping (SLAM), navigation, object recognition, autonomous driving, 3D reconstruction, classification approaches and Augmented reality [1]- [6]. Lately deep learning neural network based approaches have sparked a new wave in the field of image processing. A wide number of deep learning approaches have recently been presented in literature, and they have shown considerable promise in dealing with the conventional ill-posed issue [7]. The majority of currently available methods for monocular depth measurement frequently result in low-resolution, hazy approximations. Although the performance of recent algorithms has been gradually improving, there are still significant issues with estimation of depth map's quality and resolution. These significant issues with monocular depth estimation are addressed in this paper.

Certain well-known sensors are available for depth estimation purpose like stereo cameras, LIDAR and time of flight based depth cameras. These depth sensors have several limitation like, light sensitivity, high costs for maximum depth precision, and extremely low resolution (with just a low lines of resolution in the upright direction) as compared to time-of-flight dependent depth cameras (such as the Kinect v2). For stereo cameras, accurate estimation requires rigorous calibration and a significant amount of calculation, which frequently fails to estimate in specific situations [8]. Due to such drawbacks, we outline a strategy depending on convolutional neural network based Unet model with pretrained DenseNet-169 to create a high-resolution depth map reconstruction and low computational complexity.

In this work, we offer two contributions:

1. CNN based Unet that predicts depth from monocular RGB images. We provide a straightforward network design based on transfer learning that generates depth estimation with greater precision and quality. With fewer parameters and fewer training rounds, compared to depth maps created using earlier methods, the ones that result reflect object boundaries more correctly.

2. To obtain a high resolution depth map, instead of simply using upsampling by transposed convolution we use bilinear interpolation which employs linear interpolations to get each pixel's value from all neighbouring pixels.

Following is how the rest of the article is organised. Section II describes about a related work based on depth estimation using CNN supervised learning. Section III outlines the suggested model's fundamental ideas. In Section IV, it is discussed how the proposed model performed both subjectively and objectively. In section V conclusion is included.

II. Related work

Even though depth estimate from several views has a long history in computer vision, depth estimation from a monocular image was only thought viable with the introduction of deep learning algorithms [9]. Indeed, extracting depth values, it is difficult to draw conclusions from a single image, as the input image might have been formed by an endless number of real world layouts [10]. Learning-based approaches, particularly deep learning, have been acclaimed to be competent in dealing with the challenge [11]. Here a deep learning method using CNN are only considered in the following literature review.

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

A. Monocular Depth Estimation Using Convolution Neural Network

Convolutional Neural Networks (CNNs) were motivated by the structure of the visual cortex, specifically the theories described in [12]. Convolutional layer, pooling layer, fully connected layer, and activation function are the four key layers that make up CNN. The two-dimensional spatial characteristics of the input image can be learned by CNN using these layers. The input is transformed into depth features by the convolutional layer. The pooling layer uses maximum or average pooling to reduce the size of the input feature map. Typically, the fully connected layer is discovered at the very end of the CNN to output the results. To avoid utilising only pure linear combinations, the activation function is often a continuously differentiable nonlinear function [13] [14].

CNNS are used in supervised approaches [15]- [21], unsupervised approaches [22]- [27], and semi-supervised techniques [28] for monocular depth calculation using contextual information. During evaluation, depth networks forecast depth maps from individual images, despite the fact that the unsupervised and semi-supervised approaches depend on monocular videos frames or stereo image pairings for learning. Supervised approaches have been described in detail in following subsection.

B. Monocular Depth Estimation Using Supervised Model

Monocular depth computation might be considered a regressive challenge since the supervisory signaling of supervised learning algorithms is predicated on the underlying data of depth maps. Deep neural networks are intended to predict depth information from single pictures. First, Eigen et al. [15] use CNNs to tackle depth estimate through monocular vision process. They suggested structure, which is made up of two constituent stacks (the worldwide rough networking and the localized fine-scale networking), is meant to forecast the depth image from such a single image. Dijk et al. [16] research demonstrates how neural networks detect depth. The lack of diversity in the training set, which might be remedied by adding more data, or characteristics of convolutional neural networks (such as their invariance to translation but not to scaling), are likely explanations in this paper. Laina et al. [17] suggest a fully convolutional architecture that includes residual learning to represent the complex relationship between depth maps and monocular images. A method of adversarial patch attack for monocular depth was put by Yamanaka et al [18] by employing deep learning techniques for estimating depth. In this study, artificial patterns (adversarial patches) were created with the intention of deceiving the approaches that are intended to estimate an incorrect depth for the areas where the patterns are placed.

It is necessary to use the ground truth (GT) depth maps as a source of data for the supervised depth estimation model in order for it to learn 3D mapping and scale information. However, since real-world scenarios are challenging to acquire GT depth maps in, researchers developed GAN [29] to provide depth maps that are clearer and more accurate than those produced by other methods. A deep adversarial network with two streams was proposed by praful et al. [30] for single picture depth estimation in RGB images.

C. Datasets used in literature review

1) Cityscapes: The Cityscapes dataset [31] is primarily concerned with text categorization. This dataset consists 5,000 photos having fine descriptions and 20,000 images with rough annotation [25] . Furthermore, this data comprises of a collection of stereoscopic surveillance videos collected over so many months in 50 locations.

2) Make3D: In comparison to the preceding statistics, the Make3D dataset [9] only contains monocular RGB and depth pictures which does not include stereo images. Because this dataset contains no monocular sequencing or stereoscopic picture pairings, semi-supervised and unsupervised learning techniques do not see that as a training data set, whereas supervised methods can. Instead, it's extensively utilised as a collection of unsupervised techniques for evaluating the capacity of network to generalise across diverse dataset [22].

3) NYU-Depth V2 dataset: Consists of a dataset that offers 640 x 480-pixel resolution pictures and depth maps for various indoor situations [32]. There are 654 testing samples and 120K training samples in the dataset.

4) KITTI: This dataset consists of stereo photos and related outdoors using 3D laser scanning settings that were photographed while a moving vehicle was mounted with the necessary equipment. The RGB images have a resolution of about 1241X376 [33].

III. Proposed method

Here, we outline our procedure for generating a one RGB image to produce a depth map. In this study, through supervised learning we discovered how to directly regress depth from monocular information in 2D images by minimising a regression loss. A basic block diagram is explained in Figure 1. Some of the generic concepts used in the architecture are described in following subsections. We start by outlining the encoder-decoder architecture that is being used is shown in Figure 3. Then, we go over our findings about the network behaviour of the encoder and decoder and how it relates to performance followed by an algorithm. Next, we explain a loss function that is suitable for the task. The training process is considerably aided by effective augmentation strategies, which we detail in the following sections.

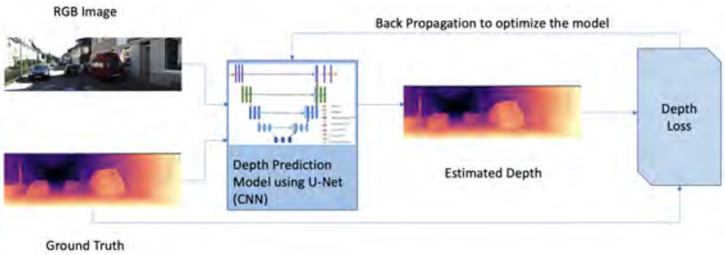

RGB Image
Back Propagation to optimize the model
Depth Prediction Model using U-Net (CNN)
Estimated Depth
Depth Loss
Ground Truth

Fig. 1. Illustration of the proposed Depth estimation by Unet.

A. *Transfer learning*

In this work, monocular depth estimation used the transfer learning and CNN Unet model, showcasing the remarkable capability of transfer learning, which is reusing a previously trained model on a different problem [34]. In transfer learning, the early and intermediate layers are utilised, and the later layers are only retrained. It helps in making use of the labelled data from the first task it was trained on. A deep learning model that has been pre-trained for a difficult image classification job like the ImageNet [36], 1000-class image classification competition is frequently used for transfer learning. Deep neural networks are employed to complete tasks involving images since they are effective at recognising intricate characteristics of the image. Our approach is greatly influenced by the concept of transfer learning and makes use of image encoders that were initially developed to solve the issue of image classification.

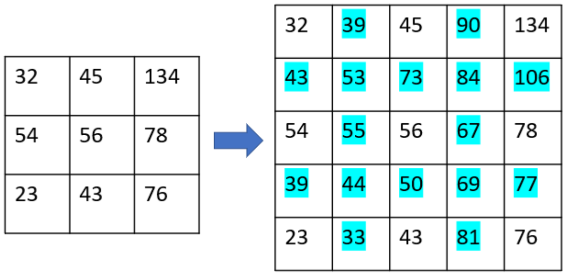
B. *Unet model*

Another crucial element is a local details in a image, which focuses on object boundaries and edges with fine details and tries to produce sharp depth maps. The decoder only makes use of concatenation and a straightforward convolution to combine high-level data coming from early stage of Unet and low level coming from later stage of Unet data. Unet framework makes advantage of its architecture providing the same for the depth estimation problem [35].

The layers in the encoder component are skip connected to the layers in the decoder part and concatenated. In order to create a image in the decoder section, the Unet must utilise the precise information they learnt in the encoder part.

A proposed network absorbs the advantages of CNN, Unet and Transfer learning.

C. *Bilinear upsampling*



32	45	134
54	56	78
23	43	76

32	39	45	90	134
43	53	73	84	106
54	55	56	67	78
39	44	50	69	77
23	33	43	81	76

Fig. 2. Illustration of Bilinear interpolation used in Decoder.

In order to increase the receptive field, CNNs frequently use pooling, which has the benefit of having a low cost of computation. Pooling, however, can result in information loss and is therefore damaging to subsequent operations like feature extraction and analysis. Instead of simply employing upsampling by transposed convolution or pooling here, we apply bilinear interpolation, which uses all neighbouring pixels to calculate the value of the pixel using linear interpolations, in order to generate a high resolution depth map. A technique of bilinear interpolation is explained in detail for understanding in Figure 2. To ensure that the characteristics of each scale, the appropriate layers of the decoder and encoder are concatenated with skip-connections.

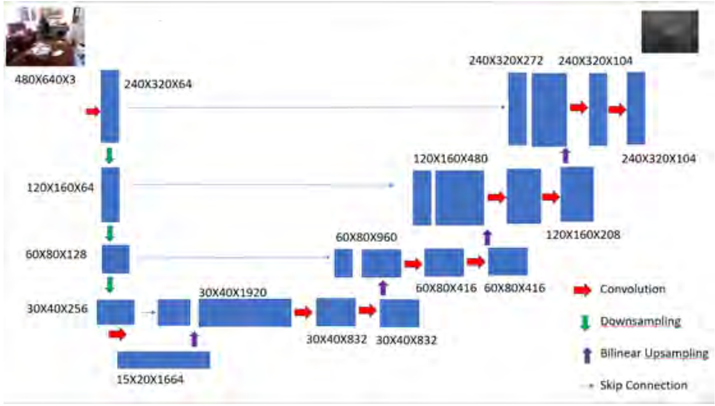


Fig. 3. A proposed pretrained Unet model for depth estimation

D. *Proposed Network*

The algorithm for detecting the image is contained in the dense layers, therefore altering the higher layers won't change the underlying logic. DenseNet-169 [19], that infers sharp geometric layout of the indoor or a outdoor scene is used as a network at encoder pretrained on ImageNet. This vector is subsequently sent through a number of upsampling layers in succession [25] to create the final depth map at a resolution that is half that of the original map. Our decoder is made up of these upsampling layers and the corresponding skip connections. By upsampling the lower-dimensional representations, this feature allowed the decoder to view feature maps at various resolutions and made it simpler to reconstruct the finer elements of the image.

The model learns to create a geometric map that more accurately depicts fine features and item boundaries from a single colour image.

An encoder-decoder configuration particular to a U-shaped design is as follows: Every layer's spatial dimensions are decreased by the encoder, but the number of channels is increased. In contrast, the decoder lowers the channels while raising the spatial dimensions. Bottleneck refers to the tensor that is fed into the decoder. In order to produce a prediction for each pixel in the input image, the spatial dimensions are ultimately recovered. These models have a significant role in practical applications. CNN based Unet that predicts depth from monocular RGB images. A novel Transfer learning with the pretrained DenseNet-169 model is implemented preserving details and reducing the number of trainable parameters.

E. *Algorithm*

- **Input**: Input x, Loss function L, Ground truth y, Epochs K, Weights W.

```
• **Output**: Network prediction using Unet with pretrained
  model DenseNet-169 trained on ImageNet.
• **Dataset** : Training set and Validation set.
• **Initialization**: Random weights, epochs count $i = 0$
• **While** $i <= K$ **do**
  Forward propagate on training data through network
  Generate prediction $y$ at output
  Evaluate prediction $L(y, \hat{y})$
  Backpropagation of loss gradient $\frac{\delta L(y,\hat{y})}{\delta W}$
  Update W and b with gradient based optimization method
  Evaluate network.
• **End while**
• Generate predictions of test set with the network with
  best performance on test set.
• Evaluate performance of network predictions.
```

F. Error function

A common error function for depth regression issues takes into consideration the discrepancy seen between ground truth depth map y as well as the estimate of the depth regression network $\hat{y}$ [15].

A loss when training a network between y and $\hat{y}$, L is described as the weighted combination of three loss functions:

$$L(y, \hat{y}) = \lambda_1 L_{depth}(y, \hat{y}) + \lambda_2 L_{grad}(y, \hat{y}) + \lambda_3 L_{SSIM}(y, \hat{y}) \tag{1}$$

The point-wise L1 loss defined on the depth values is the first loss term L_{depth}:

$$L_{depth}(y, \hat{y}) = \frac{1}{n} \sum_{p}^{n} |y_p - \hat{y_p}| \tag{2}$$

The $L1$ loss, which is defined over the depth picture's image gradient g, is the second loss term L_{grad}:

$$L_{grad}(y, \hat{y}) = \frac{1}{n} \sum_{p}^{n} |g_x(y_p - \hat{y_p}| + |g_y(y_p - \hat{y_p}| \tag{3}$$

where g_x and g_y correspondingly, determine the disparities in the x and y components for the depth image gradients of y and $\hat{y}$.

L_{SSIM} also uses the widely used measure for image reconstruction tasks known as the Structural Similarity Index Measure (SSIM) [38]. It has recently been demonstrated that this is an appropriate loss term for estimating the depth of CNNs [22]. In contrast to image-domain structures, depth-domain structures display geometric structures while excluding low-level cues that are irrelevant to depth, such as painted designs and textures on objects. By using SSIM with regard to statistical restrictions, image structure similarity may be determined. The depth distribution modelled by the mean, variance, and covariance for similarity measures also correlates with depth-domain structures. So, using the negative metric as the loss term, we compute the SSIM using the depth map inputs y and $\hat{y}$.

Since SSIM has an upper bound of one, the following is how we define it as a loss L_{SSIM}.

$$L_{SSIM}(y, \hat{y}) = \frac{1 - SSIM(y, \hat{y})}{2} \tag{4}$$

Weight parameters for the loss term L_{depth}, L_{SSIM} and L_{grad} empirically found as 0.1, 0.85, and 0.9 respectively.

IV. Experiments and Results

In this section, we present experimental results for monocular depth estimation and contrast the proposed Transfer learning Unet model with state-of-the-art CNN methods. On the benchmark NYU-Depth V2 dataset, we used various approaches to examine the resilience of our model.

A. Implementation Details

RGBD frames recorded with a Kinect sensor make up NYU-Depth V2 120K raw frames were scanned from different indoor situations. We employ a 50K sample subset of the complete dataset where depth is inpainted using the inpainting technique developed by Levin et al [37]. The 654 test frames are used to execute the NYU-Depth V2 evaluation.

Proposed depth estimation network is implemented on NVIDIA 3060 GPU with 12GB memory. The model is trained using Tensorflow, with a learning rate of 0.0001 and parameter values of $\beta 1$ = 0.9, $\beta 2$ = 0.999, the ADAM optimizer is utilised.

Model is trained for 20 epochs. They were trained with a batch size of 4 and an input resolution of 640X480 and output resolution of 320X240. The model estimates at full resolution after bilinearly upsampling but predicts outcomes with half the input resolution.

Layers up to resolution 15X20X1664 are exact replicas of DenseNet-169 layers. Every second convolutional layer in the decoder that we follow is preceded by a leaky ReLU activation function with $alpha$=0.2 as a parameter.

There are roughly 23 M parameters altogether that can be trained across the entire network. Training requires 24 hours to finish. Evaluation metric is a loss function.

B. Data Augmentation

We use a 50K subset of NYU-Depth V2 to train our algorithm. The inpainting method [37] of filling in missing depth values is used. The upper limit of the depth maps is 10 metres. Our network generates predictions with a resolution of 320X240, which is half that of the input. We downsample the ground truth depths to 320X240 for training while keeping the input images at their original resolution.

A collection of methods called data augmentation are used to fictitiously expand the volume of data. This includes applying deep learning models by making minor adjustments to the data. It also leads to increase the volume of data. Through the geometric and photometric alteration of training images, this lowers overfitting and improves generalisation performance.

Horizontal flipping, or the mirroring of images, was used as a data augmentation strategy in this work with a probability of 0.5. Another method, which is particularly effective,

TABLE 1

Method	Parameters(M)	δ_1	δ_2	δ_3	Rel	RMS	Log(10)
Eigen [15]	141.1	0.611	0.887	0.971	0.215	0.907	
Laina [17]	63.4	0.811	0.953	0.988	0.127	0.573	
Kumar [20]		0.809	0.928	0.971	0.137		0.218
Godard et al [22]		0.836	0.943	0.974	0.127	-	-
Proposed method	23	0.8505	0.9732	0.9932	0.1223	0.4614	0.053

Results obtained for proposed approach on the NYU-Depth V2 dataset

involves applying several colour channel permutations, such as switching the input's red and green channels. This technique boosts performance. The likelihood for this colour channel augmentation is now set to 0.25.

C. Evaluation metrics

We calculate depth estimation scores for the NYU-Depth V2 dataset using the accepted metrics proposed by Eigen et al [15], Abs Rel, RMSE, Average log_{10} error, Threshold accuracy (δ_i) $\delta_1 = \delta < 1.25$, $\delta_2 = \delta < 1.252$, and $\delta_3 = \delta < 1.253$.

A frequently recognised assessment approach is proposed in using 6 review identifiers:to assess and examine the effectiveness of several depth estimation methods. The error metrices are defined as:

Average relative error(rel):

$$\frac{1}{n}\sum_{p}^{n}\frac{|y_p - \hat{y}_p|}{y} \tag{5}$$

Root mean squared error (RMSE):

$$\sqrt{\frac{1}{n}\sum_{p}^{n}(y_p - \hat{y}_p)^2} \tag{6}$$

Average log_{10} error:

$$\frac{1}{n}\sum_{p}^{n}|log_{10}(y_p) - log_{10}(\hat{y_p})| \tag{7}$$

Threshold accuracy (δ_i) : % of y_p s.t.

$$max(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}) = \delta < thr \tag{8}$$

for $thr = 1.25, 1.25^2, 1.25^3$, where y_p is the pixel in depth image y, $\hat{y_p}$ is pixel of predicted depth image $\hat{y}$,and n denotes the aggregate amount of pixels with real-depth values.

D. Evaluation and Comparative Analysis

We assess our model's performance on the NYU-Depth V2 dataset using the metrics specified in the evaluation metrics section. The quantitative results are summarized in Table 1. All of the evaluation metrics show that our strategy performs better than cutting-edge approaches. RMSE, REL, log10-lower is better and for δ_1, δ_2, δ_3, higher is better. Qualitative results are shown in Figure 4 and 5 with different set of images. Our method generates depth estimations at a higher quality level with much fewer artefacts and depth edges that more closely resemble those of the ground truth.

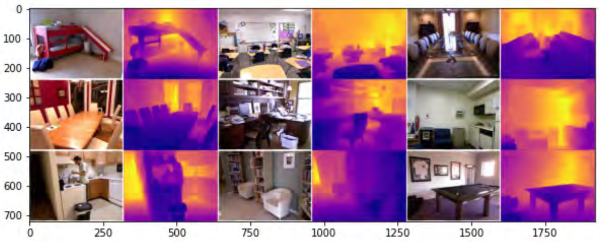


Fig. 4. Qualitative NYU-Depth V2 results. All methods were trained on indoor NYU-Depth V2 using monocular supervision.

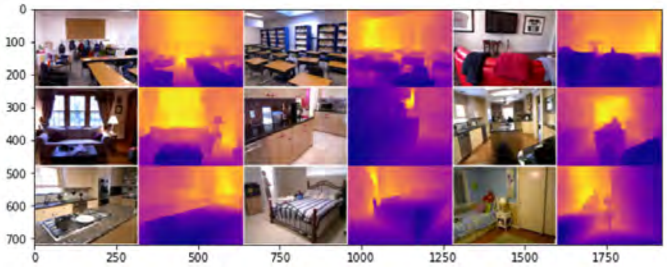


Fig. 5. Qualitative NYU-Depth V2 results. All methods were trained on indoor NYU-Depth V2 using monocular supervision.

V. Conclusion

For depth estimation, the proposed method using convolutional neural network Unet architecture pretrained on DenseNet-169 is analyzed. This method indicates good resolution with less computational complexity as compared to the state of art models. Future research will explore further experimentation with the proposed method and its applications.

References

[1] G. Hu, S. Huang, L. Zhao, A. Alempijevic, and G. Dissanayake, "A robust rgb-d slam algorithm," in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012, pp. 1714– 1719.

[2] Z. Zhu, A. Su, H. Liu, Y. Shang, and Q. Yu, "Vision navigation for aircrafts based on 3d reconstruction from real-time image sequences," Science China Technological Sciences, vol. 58, no. 7, pp. 1196–1208, 2015.

[3] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3061– 3070, 2015.

[4] Laidlow, T.; Czarnowski, J.; Leutenegger, S. DeepFusion: Real-Time Dense 3D Reconstruction for Monocular SLAM using Single-View Depth and Gradient Predictions. In Proceedings of the 2019 Sensors

2020, 20, 2272 14 of 16 International Conference on Robotics and Automation (ICRA), Montreal, Canada, 20–24 May 2019; pp. 4068–4074.

[5] Yu, F.; Gallup, D. 3D Reconstruction from Accidental Motion. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; Columbus, OH, USA, 23–28 June 2014; pp. 3986–3993.

[6] Chen, L.; Tang, W.; John, N.W.; Wan, T.R.; Zhang, J.J. Augmented Reality for Depth Cues in Monocular Minimally Invasive Surgery. arXiv Prepr. 2017, arXiv:1703.01243.

[7] Yue Ming, Xuyang Meng, Chunxiao Fan, Hui Yu, Deep learning for monocular depth estimation: A review,Neurocomputing,Volume 438, 2021, Pages 14-33, ISSN 0925-2312

[8] C. Godard, O. Mac Aodha, G.J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 270–279.

[9] Saxena, A.; Sun, M.; Ng, A.Y. Make3d: Learning 3d scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. 2009, 31, 824–840.

[10] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, Monocular Depth Estimation Based On Deep Learning: An Overview, Feng Qian East China University of Science and Technology, Shanghai, China, 200237, 3 Jul 2020

[11] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 10, pp. 2024–2039, 2015.

[12] Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, Mohammed Bennamoun, Gerard Medioni, Sven Dickinson, "A Guide to Convolutional Neural Networks for Computer Vision" , Morgan and Claypool, 2018.

[13] M. Tygert, J. Bruna, S. Chintala, Y. LeCun, S. Piantino, and A. Szlam, "A mathematical motivation for complex-valued convolutional networks," Neural Computation, vol. 28, no. 5, pp. 815–825, 2016.

[14] Y. L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in Proceedings of the ICML, 2010.View at: Google Scholar

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In Advances in Neural Information Processing Systems 27, pages 2366–2374. Curran Associates, Inc., 2014.

[16] T. v. Dijk and G. d. Croon, "How do neural networks see depth in single images?" in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2183–2191.

[17] 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). IEEE, 2016, pp. 239–248.

[18] K. Yamanaka, R. Matsumoto, K. Takahashi and T. Fujii, "Adversarial Patch Attacks on Monocular Depth Estimation Networks," in IEEE Access, vol. 8, pp. 179094-179104, 2020

[19] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. 2017

[20] A.C. Kumar, S.M. Bhandarkar, M. Prasad, Depthnet: a recurrent neural network architecture for monocular depth prediction, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 283–291 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269, 2017.

[21] Ali Zia, Jun Zhou, Yongsheng Gao, "Exploring Chromatic Aberration and Defocus Blur for Relative Depth Estimation From Monocular Hyperspectral Image", IEEE Transaction of Image Processing, VOL. 30, 2021

[22] 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, 2017, pp. 270–279

[23] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in Proceedings of the IEEE Conference

[24] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5667–5675.

[25] Gao H, Liu X, Qu M, Huang S. PDANet: Self-Supervised Monocular Depth Estimation Using Perceptual and Data Augmentation Consistency. Applied Sciences. 2021; 11(12):5383. https://doi.org/10.3390/app11125383

[26] C. Godard, O. Mac Aodha, M. Firman, G.J. Brostow, Digging into selfsupervised monocular depth estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3828–3838.

[27] Ravi Garg, Vijay B.G. Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, European Conference on Computer Vision, pages 740–756, Cham, 2016. Springer International Publishing.

[28] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6647–6655.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. (2014) 2672–2680.

[30] Praful Hambarde, Akshay Dudhane, and Subrahmanyam Murala, "Single image depth estimation using deep adversarial training," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 989–993.

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.

[32] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In Computer Vision – ECCV 2012, pages 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg

[33] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3354–3361

[34] A. R. Zamir, A. Sax, W. B. Shen, L. J. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling task transfer learning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3712–3722, 2018.

[35] Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015.

[36] Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In NIPS, pp. 1106–1114, 2012.

[37] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. ACM Trans. Graph., 23:689–694, 2004

[38] J. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13:600– 612, 2004