DepthNet: A Monocular Depth Estimation Framework

Anunay
Department of Mechanical Engineering
Delhi Technological University (DTU)
Delhi, India
anunay2608@gmail.com

Pankaj
Department of Computer Science and Engineering
Delhi Technological University (DTU)
Delhi, India
Pankajrajput020010@gmail.com

Chhavi Dhiman
Department of Electronics and communication Engineering
Delhi Technological University (DTU)
Delhi, India
chhavi.dhiman@dtu.ac.in

***Abstract*—Depth estimation is an important technique in computer vision for applications such as pose modelling, activity recognition, etc. Common methods of depth estimation include stereo vision which aims to trace the corresponding feature displacement to reconstruct a geometry as a depth map. To reduce the reliance on stereo vision systems, training-based deep learning methods have been utilized to generate depth maps using RGB images from a single camera. In this paper we propose a DepthNet framework to predict the relative depth of objects placed in the scene of an RGB image with respect to each other, with the help of multiple resizing upskip connections and up-convolutional layers, which further enhance its depth estimation abilities. The results specialize on the edges and the gradients of the objects existing within the scene by introduced as an Edge Loss. The single layered model with limited number of parameters which enables it to be implemented on any platform which has certain limiting factors of space occupancy and computational power. To validate the performance of the proposed architecture, the experiments are conducted on three publically available datasets: NYU depth dataset V2 [1], KITTI depth dataset [2], SUN-RGBD dataset. The projected work exhibits superior depth estimation results victimizing single RGB images, from the opposite state-of-the-arts. The performance of proposed work is evaluated on the following victimizing parameters: REL, RMS, Squared REL, and RMS_log10.**

Keywords- monocular cameras, RGB single image, depth estimation, edge loss, upskip connections

I. INTRODUCTION

Monocular Depth Estimation, known to be read for a long time period with its vast applications both in field of electronics and Computer Vision, concerning Robot Vision [3], Simultaneous Localization and Mapping(SLAM) [4], 3D modeling [5], human computer interaction [6], in Self-Driving Cars [7], etc. However to fulfill these widespread applications, depth estimation techniques are restricted to utilization of Depth Predicting Sensors such as Activity Vision or LIDAR [8] which measures the target distance by illuminating it with a laser light or by using multi-view observations by stereo matching. Application of these methods [9] [10] [11] being extremely costly and space occupying tackles the real life problem, which gave rise to solutions to overcome the problem, using a single sensor (camera), as well as to the present, a haul of obtaining a high-quality dense depth map from a single RGB image has entered the scene. So, during this work we have a tendency to propose a replacement network architecture named DepthNet and a custom Edge Loss to beat this unwell display downside, by rising this progressive paper for Depth Estimation on two publicly obtainable datasets for indoor (NYU depth dataset V2 [1]) and outdoor (KITTI depth dataset [2]) environments. As of now we have perceived that the depth values of indoor environments are distinctly different from that of outdoor ones. This is an intrinsic property of design that is mostly induced by the perspective-effect during the depth acquisition process.

A commonly observed problem during the time of depth estimation is of predicting the objects and differentiating them separately so that the depth-map of objects kept at the same distance, did not get merged with each other. For example, knowing the shape of an object together with its edge detection will be mutually beneficial for predicting the relative depth map of the RGB image. To address the above challenges, our main contributions are as, the proposed a deep neural network, called DepthNet predicts the relative depth of objects placed in the scene of an RGB image with respect to each other. The flow of the architecture includes the resizing upskip connections [12]and up-convolutional layers has further enhanced its depth estimation ability for monocular cameras. The custom Edge Loss [13] while estimating depth maps, helps to discriminate the edges of different objects so that they do not get merged when placed next to or at the same depth with respect to another object.

The rest of paper is organized as follows: The related work of the proposed model is present in Section II. Followed by Section III which describes the detailed proposed work of Depth Estimation Model, Depth Aware and Edge Loss concepts together with the depth map prediction from single RGB images employing a Monocular camera. Section IV represents the Experimental Results obtained by evaluating the performance of proposed work and then comparing them with the current state-of-the-art papers. Finally, conclusion and future work area are stated in Section V.

II. RELATED WORKS

Monocular Depth Estimation is a task of estimating the relative depth of objects placed in a scene with respect to other objects or surfaces present in the same RGB images obtained from a monocular camera. It can be performed by many methods, most common of them are discussed as: CNN based methods [14] that contemplate task of depth estimation a regression model training which outputs a depth map corresponding to a one RGB image. Another work [15] used Encoder-Decoder primarily based structures that have created vital contributions in several vision related problems like image segmentation [16], optical flow estimation [17], and image reconstruction [18]. The encoder downsamples the image into bottleneck and also the decoder upsample bottleneck to the final output size of depth-map, most generally, one or more encoder-decoder networks are used as a subset of a wider network., followed by residual

2021 International Conference on Engineering and Emerging Technologies (ICEET) | 978-1-6654-2714-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICEET53442.2021.9659770

978-1-6654-3897-1/21/$31.00 ©2021 IEEE

connections between I_{th} layer and $(N\text{-}I)_{th}$ layer, where N is total number of layers.

The prototypic architecture for dense depth map estimation is square measured by fully-convolutional networks Fully-convolutional networks [19, 20]. Over the years, some variations on this particular pattern have been proposed. However, in order to learn multi-scale function representations, all current architectures use convolution and subsampling as fundamental components. The works proposed progressively upsample representations which are pooled at different stages, whereas others expanded convolutions recovers fine-grained predictions making certain sufficiently large context. Newer architectures [21] [22] maintain a high-resolution illustration beside multiple lower-resolution representations throughout the network.

It can also be solved by using Multi Task Learning [23], includes a knowledge transfer mechanism which improves the generalization of tasks by sharing the particular dominion of information between complementary tasks. During this sort of technique [13] that shares the corresponding knowledge of depth prediction together with to the linguistics segmentation that helps in classification of objects in several categories and shares the corresponding information in between training processes by using Information Sharing Units. Whereas, alternative strategies search the best descent direction of the gradient by adaptively choosing the coefficient factors throughout the training process. L.L. et al. [24] proposed a Multi-Layered Depth model to perform Depth estimation and semantic segmentation all together efficiently.

Attention-based models [25] and in particular Transformers [26] are set-to-set models depending on Multi-Head Attention mechanism which empowers the models to encode multiple relationships among the features corresponding to each and every layers and have been successful when are trained on very large datasets. These network architectures are vastly used for learning and training of Natural Language Processing [27]problems. It has been invented that the transformer architecture which have been extremely successful in NLP, are functional when applied to images in tokes format, i.e. an image is further divided into certain number of patches and they are numbered accordingly and are passed through the transformer network hence resulting in the feature training of model from created tokens and patches, and can yield competitive performance in image classification. For example, DPT Hybrid [26] is one of the recently proposed works which uses Vision Transformers [28] for the task of Monocular Depth Estimation.

III. PROPOSED WORK

Network architecture: The proposed Depth Estimation Model, DepthNet, as shown in Fig. 1, is capable of predicting the depth-map of any indoor (NYU V2 [1]) and outdoor (KITTI [2]) scene using single RGB image. Depth Estimation Model consists of a backbone made of ResNet50 Convolutional Neural Network which downsamples the feature map of the input image of size (240×320) to final size of (30×40) but as observed in the native ResNet50 [29] model the final feature map size is reduced to (8×10), but according to the requirement feature map's size is not reduced to that extent which makes it difficult for our final model to learn the corresponding features while resizing them back to our output shape of dimensions$(240 \times 320 \times 1)$. The benefit of using the ResNet50 model is the upskip connections which helps the model to remember the information shared in the complete long tailed network. Four up-convolutional layers were introduced to increase the size of the feature maps received form the output of ResNet50 by the ratio of (2: 2) so that each up-convolutional layer outputs doubles the size of the input given to it, i.e. $(30 \times 40) \rightarrow (60 \times 80) \rightarrow (120 \times 160) \rightarrow (240 \times 320) \rightarrow (480 \times 640)$.

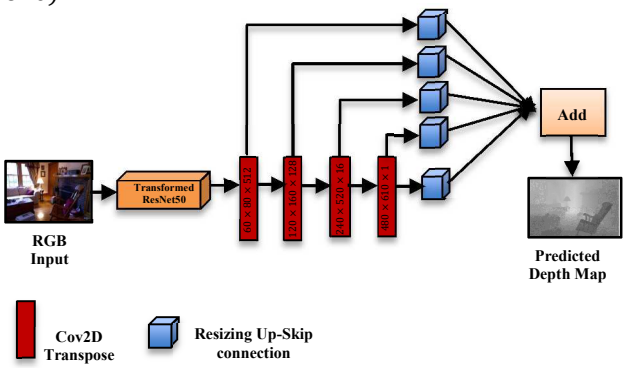

Fig. 1: A brief outline of the proposed specification, a RGB image is given as an input to the ResNet-50 architecture (Orange), deconvolutioned ResNet-50 output to the final size of depth map to be created (Red), besides resizing up-skip connections (Blue) to recollect all the antecedently shared information of the network.

On paying attention towards the output channels obtained from the customized ResNet50 those are 2048, and according to the required depth map to be of a grayscale image, channels of ResNet50 were reduced to 1. To do so, some of the optimum ratios were researched so that output scene gets converged without losing any information while reducing the channel count. Hence, there is a high possibility of the important feature information being lost, to prevent that five resizing up-skip connections are being used to reduce the output feature map size to$(240 \times 320 \times 1)$.

Table I: Layered structure of the proposed DepthNet architecture. The terms 'conv', 'trans-conv', 'relu' stands for Convolution Layer (Conv2D), Convolutional Transpose Layer (Conv2DTranspose) and 'Rectified Linear Unit' activation function respectively.

Serial Number	Layer	kernel	Strides	Input Layer	Output
I.	Input	-	-	-	240x320x3
II.	Transformed ResNet-50	-	-	I.	30x40x2048
III.	trans-conv + relu	(2, 2)	2	II.	60x80x512
IV.	trans-conv + relu	(2, 2)	2	III.	120x160x128
V.	trans-conv + relu	(2, 2)	2	IV.	240x320x64
VI.	trans-conv + relu	(2, 2)	2	V.	480x640x1
VII.	Resize Layer	(240, 320)	-	II.	240x320x2048
VIII.	conv + relu	(1, 1)	1	VII.	240x320x1
IX.	Resize Layer	(240, 320)	-	III.	240x320x512
X.	conv + relu	(1, 1)	1	IX.	240x320x1
XI.	Resize Layer	(240, 320)	-	IV.	240x320x128
XII.	conv + relu	(1, 1)	1	XI.	240x320x1
XIII.	Resize Layer	(240, 320)	-	V.	240x320x64
XIV.	conv + relu	(1, 1)	1	XIII.	240x320x1
XV.	Resize Layer	(240, 320)	-	VI.	240x320x1
XVI.	conv + relu	(1, 1)	1	XV.	240x320x1
XVII.	Add Layer	-	-	VIII, X, XII, XIV, XVI.	240x320x1

The Resizing Up-Skip Connections gets there inputs from ResNet50 and all four up-convolutional layers, resulting in

five outputs having the same size of feature map (240×320) and equal numbers of channels($equal\ to\ 1$), which is done by a resize layer which resizes the input feature map obtained from ResNEt-50 and up-convolutional layers to the output size of (240×320) followed by a convolutional layer which converts the channels of resizing layer output to 1, as seen in Fig. 2. These resizing up-skip connections are in general used for remembering the previously shared information that has the probability to be lost. At the end, these five resized outputs are added and as a result our model

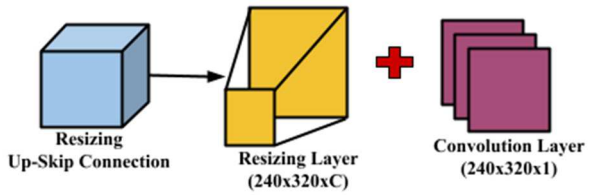

Fig. 2: The Resizing Up-Skip connections includes a resizing layer which resizes the output feature maps obtained from the inputs to (240×320) and then convolutional layer reduces the number of output channels present in the outputs procured from the resizing layers to 1.

will output this added information, and predict the relative depth between the two objects present in RGB images.

Losses: In order to better illustrate the custom losses, we tend to train the architecture for preserving the useful information obtained as features and these losses will help the network to train where the network should focus lots on. Here we use the heat map characteristics to show the final resulting feature maps of the network on predicted monocular depth, where the darker portion specify lower values (closer to zero) or closer objects and the brighter portions specify the higher values (closer to one) or the far away objects.

i) Depth aware loss: Depth Aware Loss is used to minimize the pixel-wise difference between the feature focusing terms (attention term) and regularization terms calculated by the predicted depth values and ground truth scene corresponding to the RGB images on the training datasets, which helps to supervise the monocular depth estimation task, defined in Eq. (1).

$$L_{DA} = 1/N\ (\sum(\phi_{Di}^{GT} - \psi_{Di})\ |\phi_{Di}^{GT} - \psi_{Di}|) \qquad (1)$$

where ϕ_{Di}^{GT} is an attention term, equal to normalized ground truth value, ψ_{Di} is a regularization term which is employed for learning of the features of the nearby objects and also tackles the gradient vanishing problems which leads to termination of edges of the distant objects, for example: an object rectangular in shape seems to be elliptical or oval in shape, throughout the complete duration of the training, as defined in Eq. (2):

$$\psi_{Di} = 1 - \frac{(min(log(d_i), log(d_i^{GT})))}{(max(log(d_i), log(d_i^{GT})))} \qquad (2)$$

Where d_i and d_i^{GT} are predicted depth values and ground truth values respectively (unnormalized) and N is total numbers of pixel values in predicted depth map.

ii) Edge Loss: For architecture to focus expeditiously on the native gradients and edges of structures and on the surfaces present in the scene and preserve information regarding it, we tend to propose the edges in the output scenes obtained from the projected architecture and introduce the Edge Loss as shown in Eq. (3).

$$L_E = \frac{1}{N}(\sum(|d_{ix} - d_{ix}^{GT}| + |d_{iy} - d_{iy}^{GT}|)) \qquad (3)$$

Where N is total numbers of pixel values in predicted depth map, d_{ix} and are d_{iy} sobel gradients of the predicted depth scene d_i in x and y direction, defined in Eq. (4) and (5) below.

$$d_{ix} = d_i \times \begin{matrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{matrix} \qquad (4)$$

$$d_{iy} = d_i \times \begin{matrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{matrix} \qquad (5)$$

The main reason to use this Edge Loss is to differentiate and maintain a high inter-class separation among the different objects. Hence, it resolves the problem faced during the prediction of grayscale depth map when the objects placed next to each other and at same distance from the monocular camera, their depth map do not get merged with each other.

The final attention loss applied to the proposed Depth Estimation Model can be defined as the summation of the outputs of Depth-Aware Loss and Edge Loss, as given below, Eq. (6) i.e.

$$L_{Depth} = L_{DA} + L_E \qquad (6)$$

IV. Experimental Results

Experimental Details: The proposed DepthNet architecture is implemented in tensorflow. In the training process, an adaptive momentum optimizer (Adam) is used which can handle sparse gradients on noisy problems. The model is trained for a total of 70 epochs. For initial 50 epochs Combined Depth Loss is used and for next 20 epochs depth edge loss is used while training, to further specify the gradients in the predicted depth maps. The learning rate is set to e^{-4} and batch size taken 4 images for each step per epoch. One epoch took approximately 39 minutes with Tesla T4. The entire proposed model consists of 28M parameters, 23.5M parameters in ResNet50 backbone and remaining in CNN Layers. The training loss curves for both NYU Depth Datasets V2 and KITTI Depth Dataset are shown in Fig. 5. *Datasets*: To evaluate the performance of the proposed model, experiments are conducted on three publicly available datasets: *NYU Depth Dataset V-2* [1], *KITTI Depth Dataset* [2], *SUN RGB-D Dataset* [30]. All datasets consisting of RGB images captured from a monocular and stereo camera. *NYU Depth Dataset V-2* [1] provides images and depth charts for various indoor scenes shot at pixel resolution of(640×480), which includes images of Bedrooms, Bathrooms, Work Places, Kitchen, etc. The dataset contains fifty thousand and six hundred and fifty-four training and testing samples respectively. We trained our network on thirty-five thousand subset images. The sample images are shown in Fig. 3. *KITTI Dataset* [2] includes depth maps from projected LiDAR point clouds that were matched against the depth estimation from the stereo cameras. The depth maps have an upper bound of 80 meters. The network was trained on the twelve thousand subset images to generalize the results considering indoor and outdoor environments both. The sample frames are shown in Fig. 4. SUN RGB-D Dataset

[30], is a dataset of 10335 real RGB-D images of room scenes. Each RGB image has a corresponding depth and segmentation map. The training and testing sets contain 5285 and 5050 images, respectively. Subset of this dataset consists of NYU depth dataset, which were already used for training the model on indoor dataset, hence it was further used for testing.

Fig. 3: Some examples showing the monocular depth estimation performed real time on NYU Depth Dataset-V2 [1].

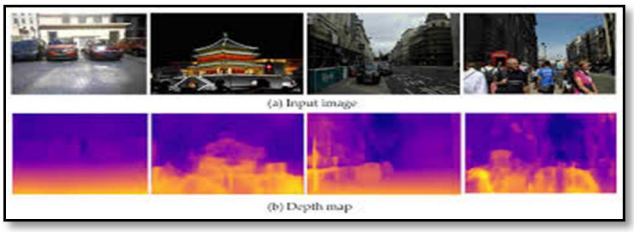

Fig. 4: Some examples showing the monocular depth estimation performed on KITTI Depth Dataset [2].

Evaluation Metrics : To evaluate the prediction quality of the proposed Depth Estimation methods, we used four evaluation metrics: Root Mean Square (RMS) [31] Error, Average Relative Error(REL) [32], Logarithmic Root Mean Square Error(RMSE_log) [31] and Squared Relative Error(Squared REL) [31] to check the similarity measure between predicted depth value and ground truth as defined in Eq. (7)-(10).

$$RMSE = \sqrt{(1/n(\sum(y_p - \hat{y}_p)))} \quad (7)$$

$$REL = 1/n\ (\sum(|y_p - \hat{y}_p|/y) \quad (8)$$

$$\text{Squared } REL = 1/n\ (\sum(|y_p - \hat{y}_p|\text{^}2\ /y) \quad (9)$$

$$RMSE\_\log = \sqrt{(1/n\ (\sum(||log\ (y_p) - log\ (\hat{y}_p)||\text{^}2)))} \quad (10)$$

Where y_p, $\hat{y}_p$ and y are normalized ground truth value, predicted depth map pixel values and testing image's pixel value respectively, n is total numbers of pixel values in predicted depth map, $p \in (0,1)$.

The overall performance of the proposed Results work for KITTI depth dataset [2], NYU depth dataset V-2 [1] is given in Table II. The obtained results are also compared with the recent state-of-the-arts in Table III and IV for NYU depth dataset V2 [1] and KITTI depth dataset [2] respectively.

From Table III, it is clearly noticeable that the proposed work is outperforming the present state-of-the-art papers by a large margin. However, DPT-Hybrid [26] achieved 0.045 log10 value. Whereas, the proposed work is able to achieve 88.59 % better performance for RMS value and similarly 31.52 %for average relative error values. These results, as shown in Fig. 6, assure the depth estimation preciseness and accuracy of the proposed work on NYU depth dataset V2 [1].

From Table IV, it is concluded that the proposed work is outperforming the current state-of-the-arts by a large margin. However, AdaBins [33] achieved 0.058 average relative error value. Whereas, the proposed work is able to achieve 31.52 % better performance for Squared Relative Error value. These results, as shown in Fig. 4, assure the depth estimation preciseness and accuracy of the proposed work on KITTI depth dataset [2]. The qualitative results on the proposed work on testing datasets of NYU V2 and KITTI depth dataset and hence conclude that the final results are the standard and comparable to the ground truth images where the darker pixels represent the closer whereas lighter pixels represent the farther objects and boundaries respectively, as shown in Fig.3.

Table II: Performance of the proposed work for KITTI depth dataset [11] and NYU depth dataset V2 [10] datasets

Dataset	**RMSE** [31]	**REL** [32]	**Squared REL** [31]	**RMSE_log** [33]
KITTI [2]	0.102	0.079	0.063	1.218
NYU Depth V2 [1]	0.0407	0.0376	-	-

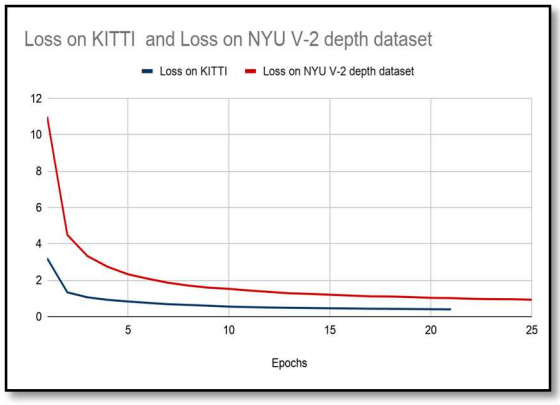

Fig. 5: Representation of continuously degrading loss curves on the KITTI depth dataset [2] and NYU Depth Dataset V2 [1].

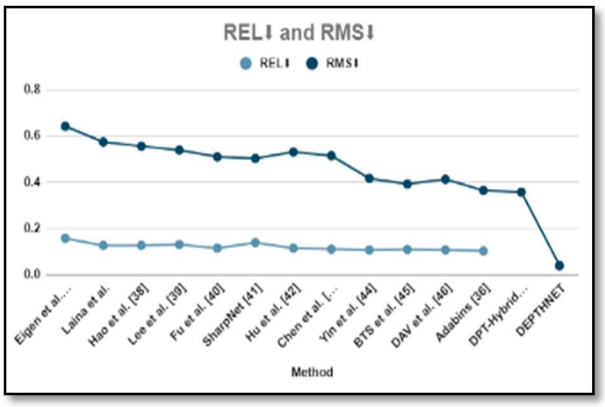

Fig. 6: Average Relative Error(REL) [32] and Root Mean Square(RMS) Error [31], evaluation metrics graph for NYU Depth Dataset V2 [1], showing Ours Model indicating less error values on testing data as compared to corresponding state-of-the-arts.

Table III: Quantitative results of proposed work on the NYU Depth Dataset V2 [1] in comparison with previously stated state-of-the-art papers.

Method	**REL↓**	**RMS↓**	**log10↓**
Eigen et al. [34]	0.158	0.641	-
Hao et al. [35]	0.127	0.555	0.053
Fu et al. [36]	0.115	0.509	0.051
SharpNet [37]	0.139	0.502	0.047
Hu et al. [38]	0.115	0.530	0.050
Chen et al. [39]	0.111	0.514	0.048

Yin et al. [40]	0.108	0.416	0.048
BTS et al. [41]	0.110	0.392	0.047
DAV et al. [42]	0.108	0.412	-
Adabins [33]	0.103	0.364	**0.044**
DPT-Hybrid [26]	-	0.357	0.045
DepthNet (our)	**0.0376**	**0.0407**	1.165

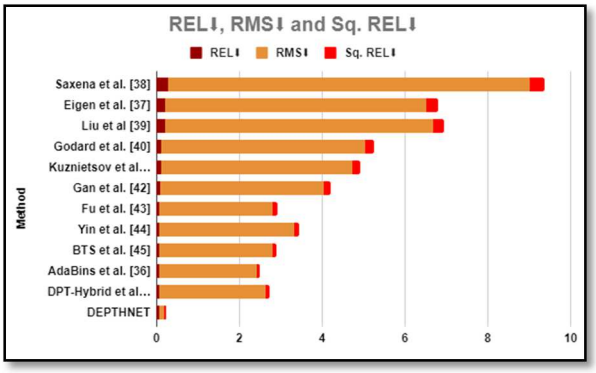

Fig. 7: Graphical comparison of the previously stated and the proposed network architecture on basis of Root Mean Square Error [31], Average Relative Error [32] and Squared Relative Error [31] on KITTI depth dataset [2], indicating that our model outperformed by giving less values for thee error than the corresponding state-of-the-art papers.

Table IV: Quantitative results of the projected work on the KITTI depth dataset [2] and comparing it with previously stated state-of-the-arts.

Method	REL⬇	RMS⬇	Sq. REL⬇	RMS_log⬇
Eigen *et al.* [34]	0.203	6.307	0.282	0.282
Liu *et al* [43]	0.201	6.471	0.273	0.273
Godard *et al.* [44]	0.114	4.935	0.206	0.206
Kuznietsov *et al.* [45]	0.113	4.621	0.189	0.189
Gan *et al.* [46]	0.098	3.933	0.173	0.173
Fu *et al.* [36]	0.072	2.727	0.120	0.120
Yin *et al.* [40]	0.072	3.258	0.117	0.117
BTS *et al.* [41]	0.059	2.756	0.096	0.096
AdaBins *et al.* [33]	**0.058**	2.360	0.088	**0.088**
DPT-Hybrid *et al.* [26]	0.062	2.573	0.092	0.092
DepthNet (our)	0.079	**0.102**	**0.063**	1.218

Fig. 8: Qualitative results of applied network architecture and Edge Loss while testing on NYU Depth Dataset V2 [1], the output depth map is at resolution$(240 \times 320 \times 1)$.

In case of testing of KITTI dataset, we represented the results in the plasma color map where, dark bluish color shows closer and as the pixels become reddish and yellowish it shows far away objects respectively, as expressed in Fig. 8.

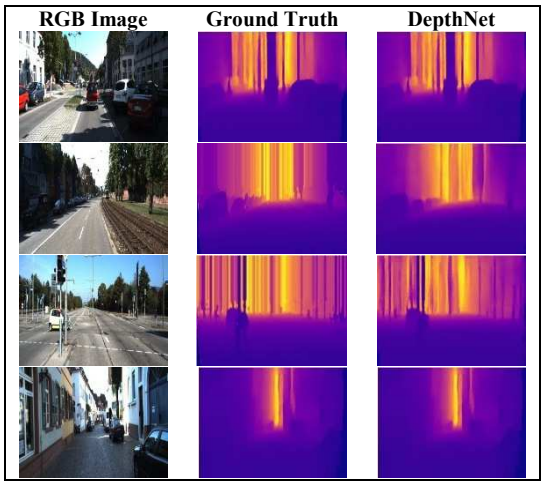

Fig. 9: Qualitative results of applied network architecture and Edge Loss while testing on KITTI Depth Dataset [2], the output depth map is at resolution$(240 \times 320 \times 1)$.

Considering the qualitative and quantitative results of the proposed architecture and loss algorithm, it can be said that the aim, to trace the corresponding features of a RGB frame to reconstruct a geometry as a depth map is fulfilled efficiently. As shown in the Table IV and Table V, the comparison results on the commonly used matrices outperforms that of AdaBins [33], the current state-of-the-art architecture.

V. CONCLUSION

On comparison with the current state-of-the-art papers, our proposed DepthNet architecture predicts the relative depth of an object with respect to another object or surface and the custom Edge Loss thereby differentiates them from their overlapping surfaces, and hence preserves the better dense depth maps for predicting on the two publicly available datasets, for indoor: NYU depth dataset V2 [1], as well as for outdoor: KITTI depth dataset [2], environments. The projected methodology is computationally economical and is more inclined towards the implementation in form of any mobile application or places wherever the weight constraints are applied, as an example, within the field of Autonomous Aerial or Ground Vehicles.

REFERENCES

[1] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, "Indoor segmentation and support inference from rgbd images.," in *European Conference on Computer Vision*, 2012.

[2] A. Geige, P. Lenz and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite.," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[3] L. Nalpantidis and A. Gasteratos , "Stereo Vision Depth Estimation Methods for Robotic Applications.," in *Depth Map and 3D Imaging Applications: Algorithms and Technologies.*, 2012.

[4] K. Tateno, . F. Tombari, . I. Laina and N. , "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction.," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] X. Wang, W. Yin, T. Kong, Y. Jiang, L. Li and C. Shen, "Task-Aware Monocular Depth Estimation for 3D Object Detection.," in *AAAI Conference on Artificial Intelligence*, 2020.

[6] Y. Duan, H. Deng and F. Wang, "Depth Camera in Human-Computer Interaction. Intelligent Networks and Intelligent

System," in *Fifth International Conference on Intelligent Networks and Intelligent Systems*, 2012.

[7] N. Appiah and N. Bandaru, "Obstacle detection using stereo vision for self-driving cars".

[8] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell and K. Q. Weinberger, "Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] V. Harisankar, V. V.V Sajith and K. P. Soman, "Unsupervised Depth Estimation From Monocular Images For Autonomous Vehicles," in *Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020.

[10] A. Sabnis and L. Vachhani, "Single image based depth estimation for robotic applications," in *IEEE Recent Advances in Intelligent Computational Systems*, 2011.

[11] Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. van der Maaten, M. Campbell and K. Q. Weinberger, "Anytime Stereo Image Depth Estimation on Mobile Devices," in *International Conference on Robotics and Automation (ICRA)*, 2019.

[12] N. Zheng, Y. Shi, W. Rong and Y. Kang , "Effects of Skip Connections in CNN-Based Architectures for Speech Enhancement," in *Journal of Signal Processing Systems*, 2020.

[13] J. Jiao, Y. Cao, Y. Song and R. Lau, "Look Deeper into Depth: Monocular Depth Estimation with Semantic Booster and Attention-Driven Loss," in *European Conference on Computer Vision (ECCV)*, 2018.

[14] R. Garg, V. Kumar B.G, G. Carneiro and I. Reid, "Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue," in *European Conference on Computer Vision(ECCV)*, 2016.

[15] C. Godard, O. M. Aodha, M. Firman and . G. J. Brostow, "Digging Into Self-Supervised Monocular Depth Estimation," in *International Conference on Computer Vision (ICCV)*, 2019.

[16] F. Sultana, A. Sufian and P. Dutta, "Evolution of Image Segmentation using Deep Convolutional Neural Network.," in *arXiv:1907.11111v1*, 2019.

[17] J. Hur and S. Rot, " Optical Flow Estimation in the Deep Learning Age," in *Modelling Human Motion*, 2020.

[18] P.-Y. Liu and E. Y. Lam, "Image Reconstruction Using Deep Learning.," in *arXiv:1809.10410*, 2018.

[19] J. Long, E. Shelhamer and T. Darrel, "Fully Convolutional Networks for Semantic Segmentation," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] J. Dai, Y. Li, K. He and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," in *arXiv:1605.06409* , 2016.

[21] Y. Arora, I. Patil and T. Nguyen , "Fully Convolutional Network for Depth Estimation and Semantic Segmentation.," in *stanford.edu*.

[22] B. Kang, Y. Lee and T. Q. Nguyen, "Depth-Adaptive Deep Neural Network for Semantic Segmentation," in *IEEE Transactions on Multimedia*, 2018.

[23] Y. Zhang and Q. Yang, "A Survey on Multi-Task Learning," in *arXiv:1707.08114v3*, 2021.

[24] L. Liebel and M. Körner, "MultiDepth: Single-Image Depth Estimation via Multi-Task Regression and Classification," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019.

[25] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *arXiv:1409.0473*, 2016.

[26] R. Ranftl, A. Bochkovskiy and A. Vladlen Koltun, "Vision Transformers for Dense Prediction.," in *arXiv:2103.13413v1*, 2021.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need.," in *arXiv:1706.03762v5* , 2017.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. ,. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," in *arXiv:2010.11929*, 2020.

[29] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] S. Song, S. P. Lichtenberg and J. Xiao, "SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite.," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] A. Botchkarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology," in *arXiv:1809.03006* , 2018.

[32] C. Chen, J. Twycross and J. M. Garibaldi, "A new accuracy measure based on bounded relative error for time series forecasting," in *PloS one.*, 2017.

[33] S. Farooq Bhat,, . P. Wonka and I. Alhashim, "AdaBins: Depth Estimation using Adaptive Bins," in *arXiv:2011.14141.*, 2020.

[34] D. Eigen, C. Puhrsch and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *ICLR*, 2014.

[35] Z. Hao, y. li, S. You and F. Lu, "Detail preserving depth estimation from a single image using attention guided networks.," in *International Conference on 3D Vision (3DV)*, 2018.

[36] H. Fu, . M. Gong, C. Wang, K. Batmanghelich and D. Tao, "Deep Ordinal Regression Network for Monocular Depth Estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] M. Ramamonjisoa and V. Lepetit, "Fast and accurate recovery of occluding contours in monocular depth estimation.," in *International Conference on Computer Vision*, 2019.

[38] J. Hu, M. Ozay, Y. Zhang and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *IEEE Winter Conference on Applications of Computer Vision*, 2019.

[39] X. Chen, Z. -Jun Zha and . X. Chen, "Structure Aware residual pyramid network for monocular depth estimation," *arXiv:1907.06023.,* 2019.

[40] W. Yin, Y. Liu, C. Shen and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[41] J. H. Lee, M.-K. Han, I. H. Suh and D. Wook Ko, "From big to small: Multi-scale local planar guidance for monocular depth estimation," in *arXiv:1907.10326*, 2019.

[42] L. Huynh, P. Nguyen-Ha, J. Matas, E. Rahtu and J. Heikkilä, "Guiding monocular depth estimation using depth-attention volume," in *European Conference on Computer Vision*, 2020.

[43] F. Liu, C. Shen, G. Lin and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[44] C. Godard, O. M. Aodha and G. J. Brostow, "Unsupervised Monocular Depth Estimation With Left-Right Consistency," in *IEEE Conference on Computer Vision and Pattern Recognition* , 2017.

[45] Y. Kuznietsov, J. Stuckler and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] Y. Gan, X. Xu, W. Sun and L. Lin, "Monocular depth estimation with affinity, vertical pooling, and label enhancement," in *European Conference on Computer Vision (ECCV)*, 2018.