S2DNet: Depth Estimation from Single Image and Sparse Samples

Praful Hambarde and Subrahmanyam Murala, *Member, IEEE*

***Abstract*—Depth prediction from single image is a challenging task due to the intra scale ambiguity and unavailability of prior information. The prediction of an unambiguous depth from single RGB image is very important aspect for computer vision applications. In this paper, an end-to-end sparse-to-dense network (S2DNet) is proposed for single image depth estimation (SIDE). The proposed network processes single image along with the additional sparse depth samples for depth estimation. The additional sparse depth sample are acquired either with a low-resolution depth sensor or calculated by visual simultaneous localization and mapping (SLAM) algorithms. In first stage, the proposed S2DNet estimates coarse-level depth map using sparse-to-dense coarse network (S2DCNet). In second stage, the estimated coarse-level depth map is concatenated with the input image and used as an input to the sparse-to-dense fine network (S2DFNet) for fine-level depth map estimation. The proposed S2DFNet comprises of attention map architecture which helps to estimate the prominent depth information. The quantitative and qualitative performance evaluation of the proposed network has been carried out using the error metrics. We perform complete evaluation of S2DNet on four publically available benchmark data sets *i.e.* NYU Depth-V2 indoor dataset [1], KITTI odometry outdoor dataset [2], KITTI depth completion test database [12] and SUN-RGB database [13]. Further, we have extended the proposed S2DNet for image de-hazing. The experimental analysis shows that the proposed S2DNet outperforms the existing state-of-the-art methods for both single image depth estimation and image de-hazing.**

***Index Terms*—Depth estimation, S2DNet, Coarse-level depth, Fine-level depth.**

I. Introduction

DEPTH prediction plays an important role in the technology field including, autonomous car driving, advanced driver-assistance systems, robotics, simultaneous localization and mapping (SLAM), augmented reality, virtual reality (VR)/mixed reality (MR), semantic segmentation, 3D object recognition, and vision in low visibility [3], [4], [5]. Also, it is useful for image de-hazing, human action recognition, scene recognition, object detection, endoscopic medical image depth analysis, etc. Mostly, depth information is captured by Light Detection and Ranging (LiDAR) and Kinect camera sensor for outdoor and indoor scene respectively. The LiDAR sensor has its regulations which include mild sensitiveness, strength consumption. [6], [7]. In case of stereo camera, depth is estimated using a disparity map. The disparity map is referred as horizontal lateral pixel-level difference between left and right image of stereo camera *i.e.* $D(x,y) = x_r - x_l$. Afterwards, depth map $d(x,y)$ is determined by $\frac{f*b}{D(x,y)}$ where, f is the focal length and b is distance between cameras center. The base path of stereo sensors needs a careful calibration for accurate camera triangulation and large computing power. To solve this problem, nowadays researchers are been focusing on a single image depth estimation. Different techniques have been created in the literature to accomplish this assignment based on the methods of mirror and concentrate camera triangulation [8]. But, triangulation technique is very difficult to deploy in consumer electronic products such as smart phone camera, wearable device and medical imaging instrumentation because of compact and low cost demand of products. Single image depth estimation is effective solution to fulfil the above mentioned requirements.

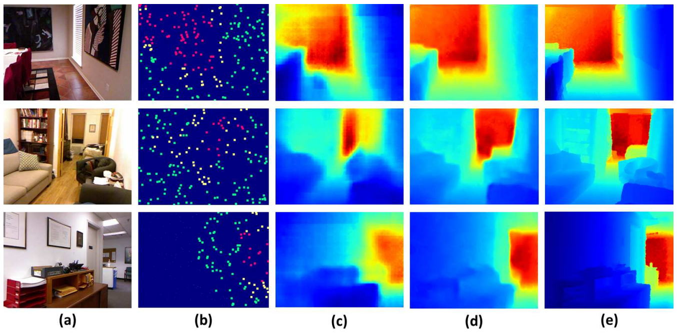

Fig. 1. Result of the proposed S2DNet. (a) Input RGB image, (b) Sparse depth samples, (c) Estimated depth map from sample RGB image, (d) Estimated depth map from sample RGB image with sparse depth samples, and (e) Ground truth depth map.

As we already discussed in previous paragraph, an energy efficient, precise and real time depth prediction is very essential task for broad area of scene analysis applications. Recently, Kinect v2 depth sensor is being used for depth sensing purpose but it absorbs ~ $15W$ of power and work for indoor scene with restricted range of ~ $0.5m$ near and ~ $4.5m$ far. The response of Kinect v2 sensor decreases as ambient light increases [9]. Also, the VR/MR mounted depth sensor absorbs the power and have 1-80m range for indoor and outdoor scene. This type of depth sensor also have contingency to conjointly develop energy-efficient depth prediction models and depth sensing product. Single image depth estimation (SIDE) also suffers with intra scale ambiguity problems [10], [11]. To overcome this limitation, we have provided additional sparse depth samples along with single image for SIDE. In proposed work, due to an unavailability of low resolution depth map, we have generated sparse depth samples from target depth image

Praful Hambarde and Subrahmanyam Murala are with Computer Vision and Pattern Recognition Laboratory, Department of Electrical Engineering, IIT Ropar, Punjab, India. Phone No: +91-9646455334; fax: +91-1881-242267; (e-mail: 2018eez0001@iitrpr.ac.in; subbumurala@iitrpr.ac.in).

by using uniform sampling strategy. A low resolution sparse depth map is obtained from Time-of-Flight (ToF) sensor, Low cost LiDAR sensor, confident stereo match.

In this paper, an end-to-end sparse to dense network (S2DNet) is proposed to estimate the scene depth map from a single image. The proposed network processes single image and additional sparse depth samples to estimate the scene depth map. The low resolution depth sensor or SLAM algorithms can be used to acquire the sparse depth samples. Fig. 1 shows the result of the proposed network for depth estimation. Fig. 1 witnessed the close appearance of the results of the proposed approach with the actual ground truth depth map. The major contributions of the proposed approach are given below:

1) A sparse depth samples based two-stream cascaded pyramid network is proposed for SIDE. In which, a sparse-to-dense coarse network (S2DCNet) is proposed for coarse-level depth estimation. Also, a sparse-to-dense fine network (S2DFNet) is proposed for fine-level depth estimation. Both S2DCNet and S2DFNet are cascaded and the overall network is named as S2DNet.
2) In the proposed S2DNet, encoder features are combined with the decoder features through the attention block. This process refines the encoder features and gives an appropriate weightage to the extracted features.
3) The proposed S2DNet is extended for the single image de-hazing.

The qualitative and quantitative comparison of the proposed S2DNet with existing state-of-the-art methods has been carried out with the help of three benchmark datasets *viz.* NYU-Depth [1], KITTI odometry [2], KITTI depth completion test database [12]. Also, SUN-RGB database [13] is used for cross-dataset evaluation.

Rest of the manuscript is organized as, Section I and II illustrate the introduction and related work on SIDE, respectively. Section III depicts the proposed approach for SIDE. Further, experimental analysis is carried out in Section IV. Section V concludes the proposed approach for single image depth estimation.

II. Related Work

In this Section, we have discussed various depth estimation approaches which are based on single image, single image with sparse depth samples and sensor modality fusion methods.

A. Single Image Modality

In the existing literature, researchers make use of probabilistic models [14], hand-crafted features, surface orientation [15] etc. for depth estimation. Saxena *et al.* [16] proposed Markov random field (MRF) model for depth estimation. In [17], [18], authors proposed Non-parametric data-driven approaches to anticipate the depth of a probe image by using photometric information given in dataset.

It is well known fact that hand-crafted features or techniques fails in complex cases. Since last decade, deep learning is state-of-the-art approach due to its learning capability. In literature [19], two stream deep network is proposed for depth estimation. Further, Eigen and Fergus [10] have proposed multi stream convolution neural network (CNN) for surface normal estimation, scene depth estimation, and extended it for semantic labelling. To achieve the refined edge details, Liu *et al.* [20] proposed a deep network and combined it with the continuous-conditional random field (C-CRF). Laina *et al.* [11] proposed a usage of residual learning for depth estimation. To overcome the unavailability of large-scale training datasets, Roy *et al.* [21] combined the random forest and CNNs for depth estimation. Further, Chakrabarti *et al.* [22] proposed a deep network to differentiate between the scene geometry of each image pixel location, scales and orientations. It is obsereved that the depth sensor fails to estimate the scene depth when captured scene contains smooth, sunny, semi-transparent surfaces [23]. Thus, Zhang *et al.* [23] proposed a deep neural network for depth channel filling. Zhang *et al.* [24] proposed progressive hard-mining network (PHN) for single image depth prediction task. The PHN continuously enhances estimated depth map representation by incorporating intra and inter scale refining operation and a refinement loss function. Haim *et al.* [25] proposed the SIDE approach, to predict the depth-map by utilizing the phase coded mask. Currently, unsupervised [26], [27], [28] and semi-supervised [28] techniques are put forward for depth estimation from stereo images. Godard *et al.* [28] proposed unsupervised approach for disparity map estimation from left and right image of stereo sensor. Ummenhofer *et al.* [29] proposed CNN based manifold bank of encoder-decoder framework for depth estimation and camera motion calculation from two consecutive images. In order to operate the autonomous vehicle safely, Mancini *et al.* [30] proposed deep CNN based method for depth estimation. Zhan *et al.* [31] proposed unsupervised technique for SIDE and visual odometry using stereo images. Moreover, Zhan *et al.* also proposed feature reconstruction warping loss for scale ambiguity free SIDE and visual odometry task. Chen *et al.* [32] presented self-supervised learning based GLNet to jointly predict the depth map, optical flow, camera pose from input video frames. To train the GLNet effectively for four different tasks, Chen *et al.* proposed adaptive photometric consistency and geometric constraints loss. Mancini *et al.* [33] proposed a deep CNN based technique to collectively learn the depth estimation and obstacle detection from vehicle navigation application. Abarghouei *et al.* [34] proposed an unsupervised learning-based depth prediction approach using a style transfer technique to minimize the difficulty in disparity map prediction. Mahjourian *et al.* [35] proposed an unsupervised learning-based technique for depth and ego-motion prediction from video. Mancini *et al.* [36] proposed encoder-decoder architecture for depth prediction and obstacle detection.

B. Sparse Samples Modality

Recently, CNN achieves a significant improvement in the SIDE. But, deep CNN based methods also suffer from spatial domain loss of predicted depth map and scene ambiguity. To address these issues, researchers introduced sparse depth map which is captured by a low resolution sensor or calculated using the SLAM algorithms. Hawe *et al.* [37] proposed wavelet based approach for dense disparity maps estimation from sparse disparity map. Liu *et al.* [38] proposed an efficient approach for depth estimation from sparse measurements. Afterwards, they integrated the wavelet and contourlet wavelet dictionary for getting more fine information from depth prediction. Secondly, Liu *et al.* [38] also proposed an alternative direction method of multipliers to estimate the depth map in an efficient and faster manner. Ma *et al.* [39] proposed a sparse-to-dense depth map estimation approach by using second order derivative of scene depth map. Also, they proposed deep residual network for dense depth map prediction from single RGB image and sparse depth samples [4]. Cheng *et al* [40] proposed a convolutional spatial propagation network to predict the depth from input image and sparse depth samples. In order to minimize the noise in sparse depth samples, Uhrig *et al.* [12] proposed sparse convolution network. Ma *et al* [41] proposed a self-supervised learning-based depth completion approach for input spatial LiDAR patterns. Qiu *et al* [42] proposed dense depth-map prediction pipeline from an input image and sparse LiDAR depth-map.

C. Sensor Modality Amalgamation

In literature, several techniques have been proposed to enhance the depth prediction response by integrating the different image modalities. For instance, Mincini *et al.* [36] proposed a deep CNN to estimate the scene depth map. They integrated input RGB and optical modality images SIDE. Liao *et al.* [43] proposed residual deep network by introducing single image and sparse planar 2D laser camera images for depth estimation. Because of the sparse laser modality fused with single image modality, residual deep network attained high accuracy as compared to single image alone. Cadena *et al.* [44] proposed multi-modal auto-encoder model to reconstruct the depth and scene segmentation map from RGB image, semantic label map and sparse depth samples from output of structure from motion (SfM) algorithm. Zhang *et al.* [45] proposed a raw depth-map completion pipeline using an input image, predicted surface normal-map and occlusion boundaries-map.

III. Proposed Approach

In this Section, we have described the proposed approach for SIDE in detail. The sparse-to-dense network (S2DNet) is discussed in the first part followed by the depth sampling technique in the second part. Inspired from [46], [47], we have designed a S2DCNet and S2DFNet. The proposed S2DNet consists of sparse-to-dense coarse network (S2DCNet) and sparse-to-dense fine network (S2DFNet) to learn the coarse and fine level depth information respectively.

A. S2DCNet Architecture

Fig. 2 shows the proposed S2DCNet architecture. From left to right, first two pyramids illustrates the proposed S2DCNet. The first encoder pyramid maps the given input image $x(i,j) \in X \subset \mathbb{R}^{q_0}$ to a feature space $z \in Z \subset \mathbb{R}^{q_c}$, then decoder takes this feature map and generate the output image *i.e.* $y(i,j) \in Y \subset \mathbb{R}^{q_0}$ (q_0 represents the number of channels of input and output images). The S2DCNet have the same number of levels for encoder and decoder pyramid *i.e.* $k = 5$. Let, ε represents the pyramid encoder/decoder layers. Features of a current pyramid encoder/decoder are given as an input to the next pyramid encoder/decoder as given in Eq. 1.

$$\varepsilon^c : \mathbb{R}^{q_{c-1}} \mapsto \mathbb{R}^{q_c} \tag{1}$$

where, $c \in [k]$ *i.e.* $k = \{1, 2, .., n\}$, n represents number of pyramid encoder/decoder levels, q_c represents number of output channels of c^{th} layer. The input for the c^{th} coarse level pyramid encoder layer comes from $(c-1)^{th}$ layer output as given in Eq. 2.

$$\xi^{c-1} = \left[\left(\xi_1^{c-1}\right) \; \dots \; \left(\xi_{q_{c-1}}^{c-1}\right)\right] \in \mathbb{R}^{q_{c-1}} \tag{2}$$

where, $\xi_j^{c-1} \in \mathbb{R}^{q_{c-1}}$ represents the j^{th} channel output of $(c-1)^{th}$ (j^{th} channel input of c^{th}) layer. The c^{th} layer of coarse-level pyramid encoder generates q_c output channels pyramid encoder feature map ξ_j^c by utilizing convolution operation.

In case of pyramid decoder, the input for c^{th} coarse-level pyramid decoder layer comes from $(c-1)^{th}$ output layer from pyramid encoder. Afterwards, the c^{th} layer of pyramid decoder layer generates q_c output channel pyramid decoder map $\tilde{\xi}_j^c$ by utilizing convolution operation.

Further, in proposed S2DCNet, we are using a skip connections from coarse-level pyramid encoder levels to coarse-level pyramid decoder levels. In a coarse-level pyramid encoder, we are performing pooling *i.e.* down-sampling operation, their will be loss of information. To recover this information, skip connections are used in the proposed S2DCNet. The generated output of $(N-c)^{th}$ coarse-level pyramid encoder levels is concatenated with the c^{th} pyramid decoder levels as given in Eq. (3).

$$\tilde{\xi}^c = \left[\tilde{\xi}^c \otimes \xi^{N-c}\right] \; ; \; c \in \left[\frac{N}{2}, N\right] \tag{3}$$

where, N represents the total number of encoder-decoder levels in the S2DCNet, $\otimes$ represent the concatenation operation.

B. S2DFNet Architecture

Further, the estimated coarse-level depth map is processed through the proposed sparse-to-dense fine network (S2DFNet) to predict the fine level depth map as shown in Fig 2. From left-to-right, the second two pyramids of Fig 2 shows the S2DFNet architecture. Decoder of the S2DFNet is designed using the principles of the attention map [48]–[50]. The pyramid encoder architecture of S2DFNet is similar as pyramid encoder architecture of S2DCNet. But, S2DFNet pyramid decoder architecture is different from S2DCNet pyramid decoder architecture. Along with skip connections, attention map is proposed in S2DFNet pyramid decoder architecture. The proposed S2DFNet estimates the fine-level depth map and suppress the irrelevant background information. The output of S2DCNet is concatenated with the input image and given

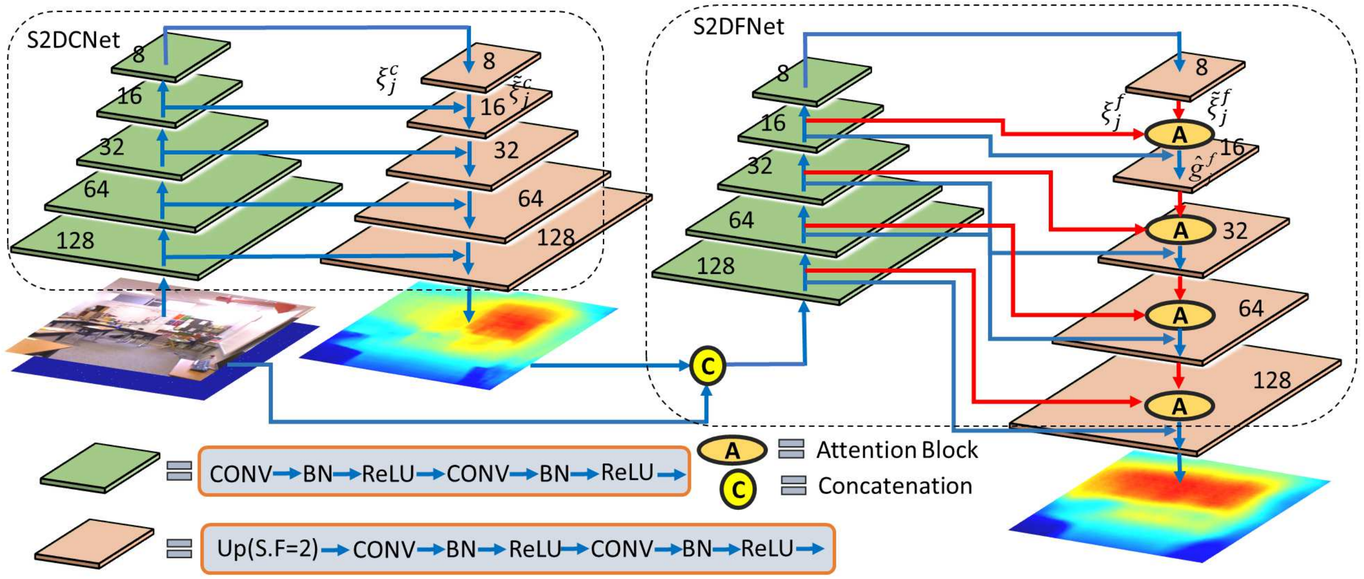

Fig. 2. The proposed S2DNet architecture with S2DCNet and S2DFNet architectures for SIDE.

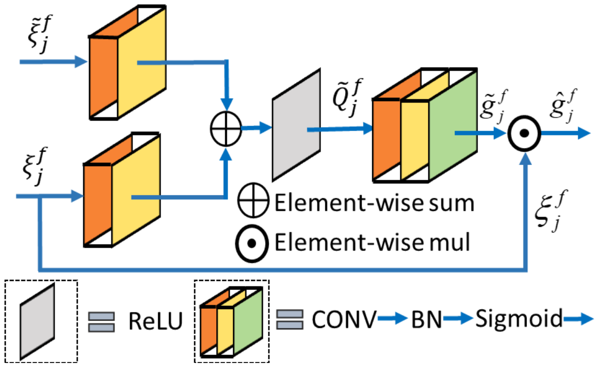

Fig. 3. The proposed attention block architecture for S2DFNet.

as input to the S2DFNet. In S2DFNet, pyramid encoder and decoder generate the fine-level pyramid encoder map ξ^f_j and pyramid decoder feature map $\tilde{\xi}^f_j$ using convolution operation.

The attention block shown in Fig. 3 takes the fine-level pyramid encoder (ξ^f_j) and pyramid decoder ($\tilde{\xi}^f_j$) layers features as an input. The estimated output from attention map architecture is given in Eq. (4).

$$\tilde{Q}^f_j = \sigma\left(\tilde{B}^f \sum_{k=1}^{q_f}\left(\xi^{N-f}_j * \varphi^f_{j,k}\right) \oplus \tilde{B}^f \sum_{k=1}^{q_f}\left(\tilde{\xi}^f_j * \tilde{\varphi}^f_{j,k}\right) + b_g\right) \tag{4}$$

where, $j \in \left[1, q_f\right]$, N represents the total number of encoder-decoder levels in the S2DFNet, $\varphi^f_{j,k}$ represents the j^{th} channel output of f^{th} layer filter, $*$ represents the convolution operation, $\oplus$ represents the element-wise sum and b_g represent the bias term. After getting the dense level information from generated output $\tilde{Q}^f_j$, we have used the sigmoid activation function to get the attention coefficients as given in Eq. (5).

$$\tilde{g}^f_j = \sigma_1\left(\left(\tilde{B}^f\right)\sum_{k=1}^{q_f}\left(\tilde{Q}^f_j * \tilde{\varphi}^f_{j,k}\right)\right), j \in \left[q_f\right] \tag{5}$$

where, $\sigma_1(.)$ denote the sigmoid activation function. The output of attention block is the element wise multiplication of pyramid encoder layer generated output and the output of sigmoid activation function as given in Eq. (6).

$$\hat{g}^f_j = \xi^f_j \odot \tilde{g}^f_j \tag{6}$$

where, $\odot$ represents the element-wise multiplication. Further, the $(N-f)^{th}$ encoder features are concatenated as skip connections with output generated from attention block as given in Eq. (7). The resultant response of Eq. (7) is given as input to the f^{th} pyramid decoder layer.

$$\hat{\xi}^{f-1}_j = \sigma\left(\sum_{k=1}^{q_f}\left(\left(\tilde{\Gamma}^f\left(\tilde{B}^f\left(\hat{g}^f_j \otimes \xi^{N-f}_j\right)\right)\right) * \tilde{\varphi}^f_{j,k}\right)\right) \tag{7}$$

At the end, we have got fine level depth information from S2DFNet as shown in Fig 2.

C. Depth Sampling Technique

In the proposed S2DNet, we have used depth sampling technique proposed by [4]. In this technique, sparse depth samples are generated from ground truth depth images. During the training of S2DNet, sparse depth image $S(i,j)$ is sampled inconstantly from ground truth depth image $S^*(i,j)$.

$$S(i,j) = \begin{cases} S^*(i,j) & With\ p \\ 0 & Otherwise \end{cases} \tag{8}$$

where, x represents the target number of depth samples (user defined parameter and should be fixed at the time of training), p is a Bernoulli probability *i.e.* $p = \frac{x}{y}$ where, y is the total number of legal pixel in $S^*(i,j)$. The sparse depth image with valid depth pixels (x) is selected randomly using the p is given in Eq. (8).

D. Loss Function

In the literature, the mean absolute error (MAE) and mean squared error (MSE) are commonly used as loss functions for regression problems. The MAE loss (ℓ_1 *loss*) is absolute difference between the target image $S^*(i,j)$ and predicted image $P^*(i,j)$ and it is more robust to the outlier. ℓ_1 loss measures an average magnitude error as given in Eq. (9).

$$\ell_1 = \frac{1}{N}\sum_{i,j=1}^{N} \|S^*(i,j) - P^*(i,j)\|_1 \tag{9}$$

The MSE loss (ℓ_2 *loss*) is the sum of squared distance between the target image $S^*(i,j)$ and predicted image $P^*(i,j)$ and it is more sensitive to outlier. ℓ_2 loss measures the average sum of squared error as given in Eq. (10).

$$\ell_2 = \frac{1}{N}\sum_{i,j=1}^{N} (S^*(i,j) - P^*(i,j))^2 \tag{10}$$

The total loss function for the S2DNet as given in Eq. (11).

$$\ell_{total} = \ell_1 + \beta\ell_2 \tag{11}$$

where, β is a weight factor. Empirically, we set $\beta = 0.1$.

IV. Experimental Analysis

The detailed experimental evaluation of the proposed S2DNet is conducted in this Section. Initially, experimental details are discussed then, quantitative and qualitative analysis has been carried out. Further, robustness of the proposed S2DNet is tested for image de-hazing.

A. Experimental Details

1) Dataset: The performance of the proposed S2DNet has been validated on NYU-Depth-V2 [1], KITTI odometry [2] KITTI depth completion [12] and SUN-RGB [13] datasets for depth estimation.

The NYU Depth-D v2 dataset consists of RGB-D images taken from 464 distinct indoor scenes using the Microsoft Kinect Sensor. We have used standard training and testing spilt for training and testing data generation *i.e.* 249 indoor scene are utilized for training and remaining 215 indoor scene for testing. Especially, for the sake of simplicity we are using small labelled 654 image dataset for testing, which has also used in earlier work [11], [19]–[22]. In training case, we have used training data generation strategy used by [4]. They randomly sampled each raw video sequence from training set and introduced 48K RGB-D image pairs. The depth image generated by projecting raw depth pixel values on RGB image and in-painted with a cross-bilateral filters by utilizing official toolbox [1]. The original images are of the size 640 × 480, initially down-sampled by the down-sampling factor two followed by the center crop to remove the extra boundaries where, depth information is not present. Finally, we reach to a size of 304 × 228 for a sample RGB-D image.

The KITTI odometry dataset consists of different urban scenes obtained by car mounted sensor and LiDAR sensor. All scenes are captured in outdoor environment. Hence, RGB image in KITTI odometry dataset is significantly different from RGB image of NYU-Depth-V2 dataset. Thus, we train the proposed S2DNet on KITTI odometry dataset to show the capability of the proposed S2DNet to predict the depth map for an outdoor scene. This dataset consist of 22 video sequences. Training split consists 11 video sequences (46k images) while remaining videos belongs to testing split [4]. In this work, we have used all 11 training video sequences (46K images) for training. For testing, 3200 images were used which are selected randomly from the testset (remaining 11 video sequences) for performance evaluation of the proposed S2DNet. The original RGB image is of size 1224×368. Then, we cut off upper boundary of each image because of the sky region *i.e.* sky region can not be scanned by LiDAR sensor. Hence, we have taken bottom crop of each image of size 912 × 228 for training and testing task.

We have also used KITTI depth completion dataset [12]. The dataset consists of 85,898 training data, 1,000 selected validation data, and 1,000 test data with RGB image and sparse LiDAR depth-map.

2) Training Details: We have implemented the proposed S2DNet using PyTorch[1] deep learning framework. The proposed S2DNet is trained on the NYU-Depth-V2, KITTI odometry and KITTI depth completion datasets utilizing stochastic gradient descent (SGD) back propagation algorithm. Training hyper-parameters are updated on NVIDIA DGX station with processor 2.2 GHz, Intel Xeon E5-2698, NVIDIA Tesla V100 1×16 GB GPU. We have used small batch size of 16 and epoch 20 for NYU-Depth-V2 and batch size of 8 and epoch 20 for KITTI odometry and KITTI depth completion datasets. During S2DNet training, learning rate start at 0.01 and decreased to 20% for each 4 epochs. For regularization purpose we use 10^{-4} as weight decay.

3) Evaluation Metrics: We have used the following evaluation metrics for quantitative analysis of predicted depth results.

- RMSE: Root mean squared error
$$RMSE = \sqrt{\frac{1}{N}\sum_{i\in N} \|S_i{}^* - P_i{}^*\|^2} \tag{12}$$
- $\log_{10}$: Average $\log_{10}$ error
$$\log_{10} = \frac{1}{N}\sum_{i\in N} \left|\log_{10}(S_i{}^*) - \log_{10}(P_i{}^*)\right| \tag{13}$$
- REL : Mean absolute relative error
$$REL = \frac{1}{N}\sum_{i\in N} \left(\frac{|S_i{}^* - P_i{}^*|}{S_i{}^*}\right) \tag{14}$$
- δ_i : Predicted pixel accuracy with relative error in threshold.
$$\delta_i = \frac{\left|\left\{P_i{}^* \in \{1,..,N\} : \max\left(\frac{P_i{}^*}{S_i{}^*}, \frac{S_i{}^*}{P_i{}^*}\right) = \delta_i < Thr\right\}\right|}{N} \tag{15}$$

$Thr = 1.25, 1.25^2, 1.25^3$

where, S^* and P^* are ground truth and predicted image. N is

[1] https://pytorch.org

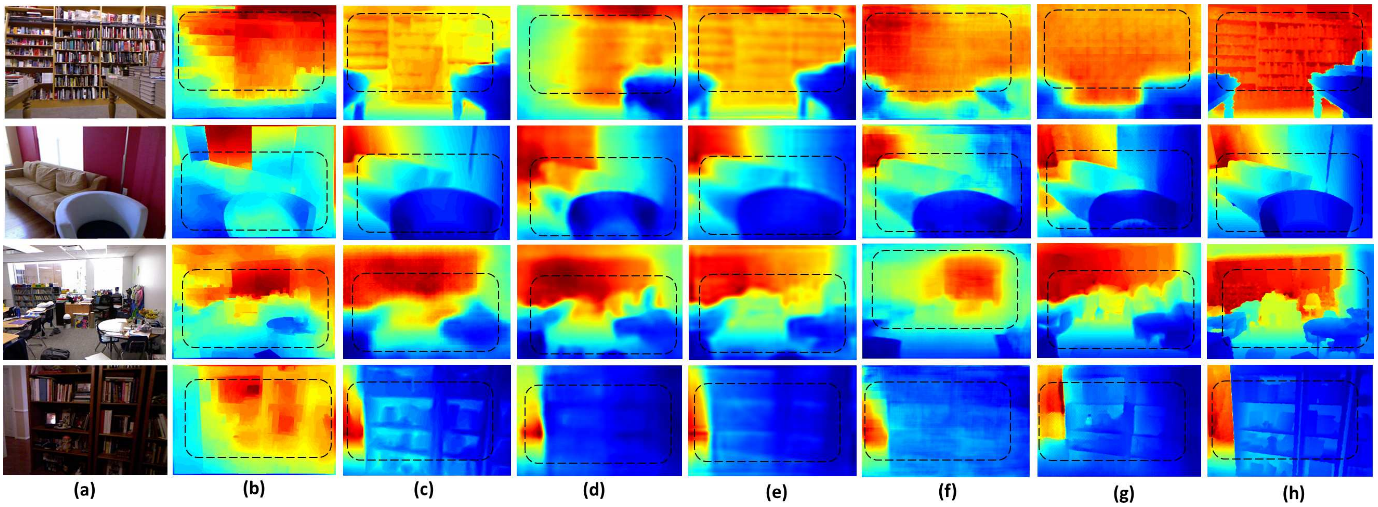

Fig. 4. Comparison between the proposed (S2DCNet, S2DNet) and existing methods on NYU-Depth-V2 dataset for single image depth estimation (when considering only RGB image). (a) Input RGB image, (b) Results of Liu *et al.* [20], (c) Results of chakrabarati *et al.* [22], (d) Results of Xu *et al.* [51], (e) Results of Lee *et al.* [52], (f) Results of the proposed S2DCNet, (g) Results of the proposed S2DNet, and (h) Ground truth depth map. *(Best viewed in bounding colors.)*

total number of valid pixels. The better depth prediction shows the higher value δ_i and lower value of errors.

TABLE I
QUANTITATIVE DEPTH ESTIMATION RESULTS EVALUATION OF THE PROPOSED S2DNET AND STATE-OF-THE-ART METHODS ON NYU-DEPTH-V2 DATASET FOR RGB IMAGE. NOTE: *RMSE, REL and* $\log_{10}$ *- Lower is better and* δ_1*,* δ_2 *and* δ_3 *- Higher is better.*

Method	RMSE	REL	$\log_{10}$	δ_1	δ_2	δ_3
Karsch [14]	1.12	0.374	0.134	-	-	-
Ladicky [56]	-	-	-	54.2	82.9	94.1
Liu [57]	1.06	0.335	0.127	-	-	-
Li [58]	0.821	0.232	0.094	62.1	88.6	96.8
Liu [59]	0.824	0.23	0.095	61.4	88.3	97.1
Wang [53]	0.745	0.22	0.094	60.5	89	97
Eigen [19]	0.907	0.215	-	61.1	88.7	97.1
Roy [21]	0.744	0.187	0.078	-	-	-
Chakrabarti [22]	0.62	0.149	-	80.6	95.8	98.7
Eigen [10]	0.641	0.158	-	76.9	95	98.8
Cao [60]	0.615	0.148	-	80.0	95.6	98.8
Xu [51]	0.586	**0.121**	0.052	**81.1**	95.4	98.7
Laina [11]	0.573	0.127	0.055	81.1	95.3	98.8
Xu [54]	0.593	0.125	0.057	80.6	95.2	98.6
S2DCNet	0.646	0.198	0.080	71.1	94.5	96.1
S2DNet (W/o AB)	0.583	0.192	0.057	74.5	95.1	97.9
S2DNet	**0.543**	0.160	**0.052**	77.3	**95.9**	**98.9**

B. Comparison with State-of-The-Art Methods

We have compared the proposed S2DNet both qualitatively and quantitatively with the state-of-the-art methods on NYU-Depth-V2 [1], KITTI odometry [2] KITTI depth completion [12], and SUN-RGB [13] datasets. The proposed S2DNet is evaluated quantitatively using RMSE, $\log_{10}$, REL and δ_i evaluation metrics. We consider the following baseline setting for fair comparative result analysis with the proposed S2DNet.
1] Comparative result analysis and evaluation of proposed S2DNet on test set of NYU-Depth dataset for RGB image we consider the [53], [22], [11], [51], [54], and [52] as baseline methods. In the case of RGB image with sparse depth samples, we consider the [4], [40], and [55] as baseline methods and randomly choose (200) sparse depth samples.
2] In the KITTI odometry dataset for RGB image we consider the [4], and [55] as baseline methods. In the case of RGB image with sparse depth samples, we consider the [4], [40], and [55] as baseline methods and randomly choose (500) sparse samples.

1) NYU-Depth-V2 dataset: The quantitative results of the proposed S2DNet and other existing methods on NYU-Depth-V2 dataset for depth estimation are given in Table I. The values of the error metrics are taken from the respective research article. Compared with the method [11], the proposed S2DNet absolute REL and δ_1 metrics results are slightly lesser but, RMSE, $\log_{10}$, δ_2 and δ_3 metrics gives better results. The values of the best performing indexes in Table are bolded. The visual results of the proposed S2DNet without using sparse depth samples on sample images from NYU-Depth-V2 dataset are illustrated in Fig 4. Here, we have compared results of the proposed method with the existing state-of-the-arts methods. The existing state-of-the-arts methods suffer from textural information loss, intra-scale ambiguity noise, etc. In contrast, the proposed S2DNet estimates the robust depth map and approaches towards the ground truth depth map. Fig. 4 and Table I shows that the proposed S2DNet outperforms the other existing approaches for depth estimation on NYU-Depth-V2 database.

Estimated depth maps for a input RGB image with sparse depth samples using the proposed S2DNet and existing methods are shown in Fig 5. It is observed from Fig 5 that the proposed S2DNet have enhanced the edge quality, texture information details and local level information in very efficient manner as compared to [4], [40], and [55] method. The quantitative results of the proposed S2DNet on RGB image with sparse depth samples of NYU-Depth-V2 dataset are given

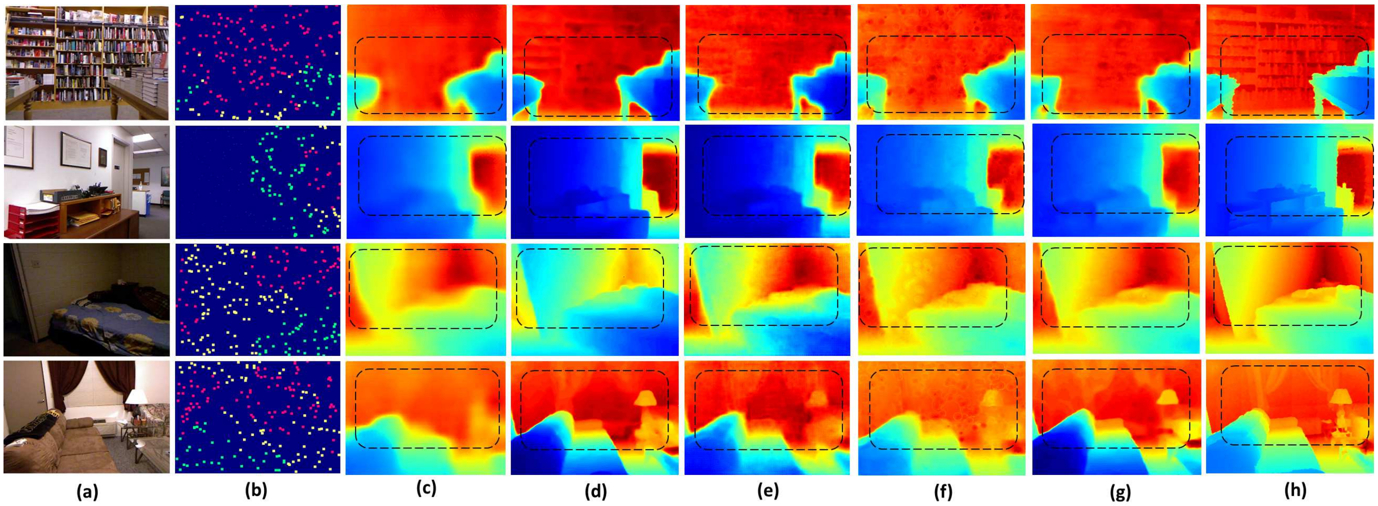

Fig. 5. Comparison between the proposed (S2DCNet, S2DNet) and existing methods on NYU-Depth-V2 dataset for single image depth estimation (when considering only RGB image with sparse depth samples). (a) Input RGB image, (b) Sparse depth samples, (c) Results of Ma [4], (d) Results of Cheng [40], (e) Result of Fu [55]), (f) Results of the proposed S2DCNet, (g) Results of the proposed S2DNet, and (h) Ground truth depth map. *(Best viewed in colors.)*

TABLE II
QUANTITATIVE DEPTH ESTIMATION RESULTS EVALUATION OF THE PROPOSED S2DNET AND STATE-OF-THE-ART METHODS ON NYU-DEPTH-V2 DATASET FOR RGB IMAGE WITH SPARSE DEPTH SAMPLES. NOTE: *RMSE, REL and* $\log_{10}$ *- Lower is better and* δ_1*,* δ_2 *and* δ_3 *- Higher is better.*

Method	RMSE	REL	$\log_{10}$	δ_1	δ_2	δ_3
Liao (225) [43]	0.442	0.104	-	87.8	96.4	98.9
Ma (50) [4]	0.281	0.059	-	95.5	99.0	99.7
Ma (200) [4]	0.230	0.044	-	97.1	99.4	99.8
Cheng (200) [40]	0.221	0.040	-	97.0	99.1	99.3
Fu (200) [55]	0.203	0.040	-	97.6	99.5	99.9
S2DCNet(200)	0.226	0.053	0.021	96.1	99.3	99.7
S2DNet(200)(W/o AB)	0.215	0.052	0.023	96.7	98.2	99.0
S2DNet(200)	**0.193**	**0.036**	**0.015**	**97.8**	**99.5**	**99.9**

in Table II. The Table II witnessed the improvement in the evaluation parameters of the S2DNet over the other existing methods [4], [40], and [55] which have utilized same number of spatially uniform sparse depth samples *i.e.* (200). The reason behind this is the proposed combination of coarse and fine-level S2DNet architecture with attention mechanism. In Table I, we have also given results of the proposed S2DNet by removing the attention block (AB). From Table I, II, Fig 4, 5 we can conclude that the proposed S2DNet with only RGB image and RGB with sparse depth samples outperforms the other existing methods for depth estimation.

2) *KITTI odometry dataset:* The KITTI odometry dataset is more challenging dataset as compared to NYU-Depth-V2 dataset for depth estimation. Since, the maximum distance is 100 meters as opposed to 10 meters in NYU-Depth-V2 dataset. The value of error metrics have taken from the respective research articles. In the proposed work, we have used sparse labelled depth map projected from LiDAR scanner instead of using disparity map determined from stereo depth as in [19], [38], [44], [28], [61]. The quantitative results of the proposed S2DNet (without sparse depth samples) and other existing methods on KITTI odometry dataset are given in Table III. From Table III, we can observe that the proposed S2DNet outperforms other existing methods. The error metric results of the proposed S2DNet are much better as compared to method [4] and [55] in all aspects. The visual results of the proposed S2DNet (without sparse depth samples) are shown in Fig 6. Fig 6 witnessed that the propsoed S2DNet outperforms the most related work [4] and [55]. As illustrated in Fig 6, the estimated depth map using the proposed S2DNet are very smooth as compared to the results of method [4] and close to ground truth LiDAR image.

The proposed S2DNet also predict the depth map from RGB image with sparse depth samples as shown in Fig 7. We have compared result of the proposed method with most related existing method [4], [40] and [55]. Visual results of proposed S2DNet outperforms the existing method [4] (*i.e.* large object and boundaries) and closer towards the ground truth. The error metric results on KITTI odometry dataset for RGB image with sparse depth samples are given in Table IV. Even though we have used (500) sparse samples in the proposed S2DNet, it outperforms the existing [44] which have used (~ 650) sparse depth samples by structure from motion (SfM) algorithm. We give this credit to the proposed coarse and fine level S2DNet architecture. An additional (500) sparse depth samples bring REL and RMSE values approx half of the RGB image approach and also enhance the predicted pixels accuracy values. We have also given the results of the proposed S2DNet by removing the AB presents in S2DFNet. Due to the presence of AB, performance evaluation of the proposed S2DNet has significantly improved. In Table III and IV, we have given a comparison of existing methods and proposed S2DNet results with and without AB. From Table III, IV, Fig 6, 7, we can conclude that estimated depth map using proposed S2DNet from RGB image with sparse depth samples outperforms the other existing methods.

TABLE III
QUANTITATIVE DEPTH ESTIMATION RESULTS EVALUATION OF THE PROPOSED S2DNET AND STATE-OF-THE-ART METHODS ON KITTI ODOMETRY DATASET FOR RGB IMAGE. NOTE: *RMSE, REL and* $\log_{10}$ *- Lower is better and* δ_1*,* δ_2 *and* δ_3 *- Higher is better.*

Method	RMSE	REL	$\log_{10}$	δ_1	δ_2	δ_3
Saxena [62]	8.734	0.28	-	60.1	82.0	92.6
Mancini [36]	7.508	-	-	31.8	61.7	81.3
Eigen [19]	7.156	0.19	-	69.2	89.9	96.7
Ma [4]	6.266	0.208	-	59.1	90.0	96.2
Fu [55]	5.920	0.193	-	67.3	91.6	97.6
S2DNet (W/o AB)	5.592	0.163	0.067	77.7	92.2	96.5
S2DNet	**5.285**	**0.142**	**0.063**	**79.7**	**93.2**	**97.5**

TABLE IV
QUANTITATIVE DEPTH ESTIMATION RESULTS EVALUATION OF THE PROPOSED S2DNET AND STATE-OF-THE-ART METHODS ON KITTI ODOMETRY DATASET FOR RGB IMAGE WITH SPARSE DEPTH SAMPLES. NOTE: *RMSE, REL and* $\log_{10}$ *- Lower is better and* δ_1*,* δ_2 *and* δ_3 *- Higher is better.*

Method	RMSE	REL	$\log_{10}$	δ_1	δ_2	δ_3
Cadena (~ 650) [44]	7.14	0.179	-	70.9	88.8	95.6
Ma (50) [4]	4.884	0.109	-	87.1	95.2	97.9
Ma (500) [4]	3.378	0.073	-	93.5	97.6	98.9
Cheng (500) [40]	3.248	**0.059**	-	94.4	97.7	98.9
Fu (500) [55]	3.670	0.072	-	92.3	97.3	98.9
S2DNet(500)(W/o AB)	3.213	0.089	0.050	92.6	96.3	98.1
S2DNet(500)	**3.112**	0.069	**0.038**	**94.5**	**97.8**	**99.1**

3) KITTI depth completion: Here, we have discussed about the performance of the proposed S2DNet on KITTI-depth-completion database (test set) [12]. We utilize an official released error metrics for the evaluation, such as Root-Mean-Square Error (RMSE[mm]), Mean-Absolute Error (MAE[mm]), Inverse-Root-Mean-Square Error (iRMSE[1/km]), Inverse Mean-Absolute Error (iMAE[1/km]). Table V gives the performance of the proposed S2DNet and existing methods on KITTI-depth-completion database (test set). Table V witnessed the superiority of the proposed S2DNet over other state-of-the-art methods in depth estimation.

C. Cross-Dataset Evaluation

In this Section, we have shown the performance of the proposed S2DNet (with sparse depth samples) on SUN RGB-D database [13]. Here, interesting fact is, we have not fine-tuned the proposed S2DNet on SUN RGB-D database. Thus, we call this as a cross-dataset evaluation. Testing split of the SUN RGB-D indoor scene dataset consists of 5,050 images. As per cross dataset evaluation methods in literature [64], [11], [60] we select only 500 images from testing split. We compared the proposed S2DNet (with sparse depth samples) both qualitatively as well as quantitatively with the existing methods [64], [11], [60], [4] as given in Fig 8 and Table VI respectively. The qualitative results of the proposed network are resemble closely to the ground truth depth maps as shown in Fig 8. From Table VI and Fig 8, it is clearly observed that the proposed S2DNet outperforms the other existing methods for depth estimation.

TABLE V
QUANTITATIVE DEPTH COMPLETION RESULTS EVALUATION OF THE PROPOSED S2DNET AND STATE-OF-THE-ART METHODS ON KITTI DEPTH COMPLETION TEST DATASET.

Method	RMSE	MAE	iRMSE	iMAE
Cheng [40]	1019.64	279.46	2.93	1.15
Huang [63]	841.78	253.47	2.73	1.13
Ma [41]	**814.73**	249.95	2.80	1.21
S2DNet	830.57	**247.85**	**2.70**	**1.20**

TABLE VI
CROSS DATASET DEPTH ESTIMATION RESULTS EVALUATION OF THE PROPOSED S2DNET ON SUN RGB-D DATASET FOR RGB IMAGE WITH SPARSE DEPTH SAMPLES. NOTE: *RMSE, REL and* $\log_{10}$ *- Lower is better and* δ_1*,* δ_2 *and* δ_3 *- Higher is better.*

Method	RMSE	REL	$\log_{10}$	δ_1	δ_2	δ_3
Liu [64]	0.931	0.316	0.161	35.6	57.6	83.1
Liana [11]	0.851	0.279	0.138	53.9	70.3	89.0
Cao [60]	0.839	0.256	0.127	56.3	72.7	88.2
Ma(200) [4]	0.847	0.284	0.181	54.2	71.9	80.7
S2DNet(200)	**0.683**	**0.122**	**0.055**	**88.1**	**95.1**	**97.2**

D. Ablation Study

Here, an ablation study has been carried out to analyse the importance of attention block in the proposed S2DNet. We examine the performance of the proposed S2DNet with and without attention block for depth prediction from RGB image with sparse depth samples. The quantitative and qualitative analysis of the proposed S2DNet with and without attention block are shown in Table VII and Fig 9 respectively. It is observed from the Table VII and Fig 9 that the proposed S2DNet with attention block estimates more accurate depth map as compared with the S2DNet without attention block. Observations about the attention block are (1) The Sobel gradient along y-direction of the depth map (which is estimated using the proposed S2DNet with attention block) provides a sharp edge information when compared with that of S2DNet (w/o attention block) as shown in Fig 9 (d), (f), and (h) respectively. (2) From Fig 9 (c), (e) and (g), it is clear that the proposed S2DNet minimizes the noisy information more precisely as compared to the S2DNet (w/o attention block) and gives a fine-level depth information which is closer to the ground truth depth map.

In Table VII, we have given a quantitative evaluation of the proposed S2DNet with and without attention block. It is clearly observed from Table VII that the proposed S2DNet with attention block outperforms the S2DNet without attention block in terms of the evaluation parameters. From Table VII and Fig 9, we can conclude that the presence of AB block in S2DFNet improves the performance of the proposed S2DNet for single image depth estimation.

E. Application: Image De-hazing

We have verified the proposed S2DNet for depth estimation which has shown significant improvement in the accuracy as compared to existing methods. In this section, we have analysed the usefulness of the proposed S2DNet for single

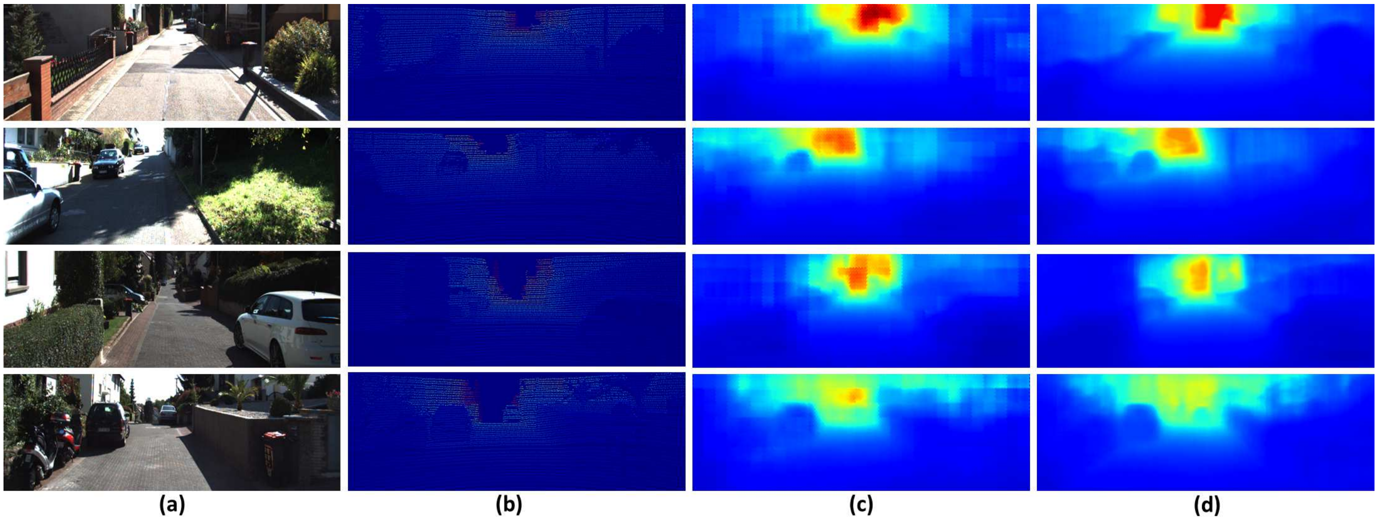

Fig. 6. Qualitative depth estimation results analysis of the proposed S2DNet and state-of-the-art methods on KITTI odometry dataset for RGB image. (a) Input RGB image, (b) Ground truth, (c) Results of Ma [4], and (e) Results of the proposed S2DNet.

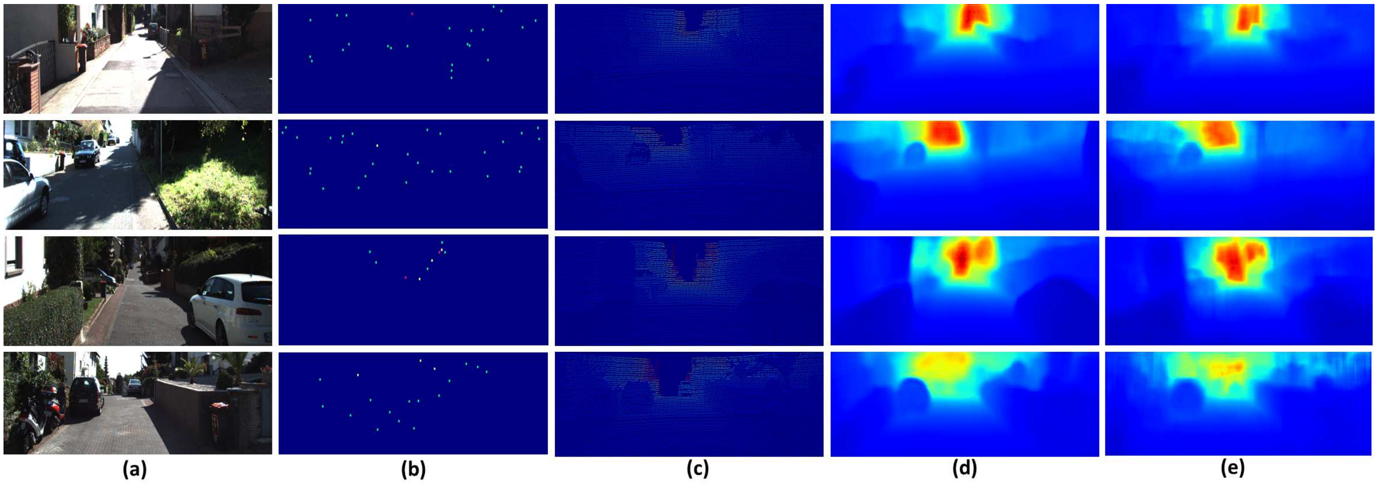

Fig. 7. Qualitative depth estimation results analysis of the proposed S2DNet and state-of-the-art methods on KITTI odometry dataset for RGB image with sparse depth samples. (a) Input RGB image, (b) Sparse depth samples, (c) Ground truth depth map, (d) Results of Ma [4], and (e) Results of the proposed S2DNet.

TABLE VII
QUANTITATIVE ANALYSIS OF THE PROPOSED S2DNET WITH AND WITHOUT AB BLOCK ON NYU-DEPTH-V2 DATASET FOR RGB IMAGE WITH SPARSE DEPTH SAMPLES. NOTE: *REL*, $\log_{10}$-*Lower is better and* δ_2, δ_3-*Higher is better.*

Methods	REL	$\log_{10}$	δ_2	δ_3
S2DNet w/o AB block	0.052	0.023	98.2	99.0
S2DNet with AB block	**0.036**	**0.015**	**98.2**	**99.5**

TABLE VIII
QUANTITATIVE IMAGE DE-HAZING RESULTS EVALUATION OF THE PROPOSED S2DNET ON D-HAZY DATASET FOR HAZY IMAGE WITH SPARSE DEPTH SAMPLES. NOTE: *SSIM and PSNR - higher is better and CIEDE2000 - lower is better.*

Method	SSIM	PSNR	CIEDE2000
DehazeNet [65]	0.7270	13.4005	13.9048
MSCNN [66]	0.7231	12.8203	15.8048
DChP [67]	0.7060	12.5876	15.2499
CAP [68]	0.7231	13.1945	16.6783
AODNet [69]	0.7177	12.4120	16.6565
CycleDehaze [70]	0.6490	**15.4130**	15.0263
S2DNet(200)	**0.7288**	14.0871	**13.6027**

image haze removal. Existing approaches [67], [71], [65] make use of atmospheric scattering model [67] to recover the haze-free scene from hazy scene. Accurate estimation of scene transmission map is a prominent step in image de-hazing. According to [67], scene transmission map is inversely proportional to the scene depth map. Thus, here we have utilized this observation and used proposed S2DNet to estimate the depth map from a given hazy scene. Hazy scene along

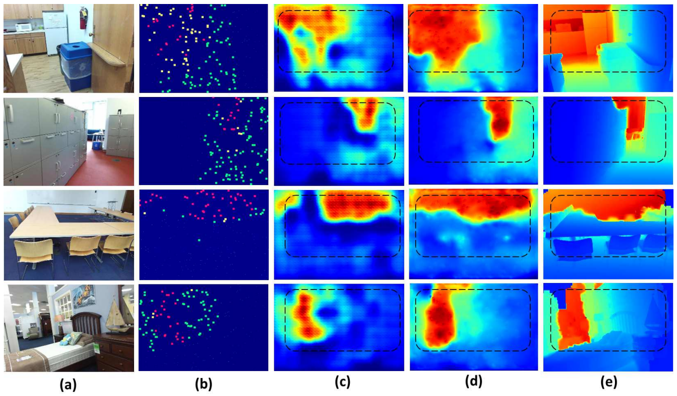

Fig. 8. Cross dataset visual depth estimation results of the proposed S2DNet on SUN RGB-D dataset for RGB image with sparse depth samples. (a) Input RGB image, (b) Sparse depth samples, (c) Results of Ma [4], (d) Results of the proposed S2DNet, and (e) Ground truth depth map. (*Best viewed in colors.)*

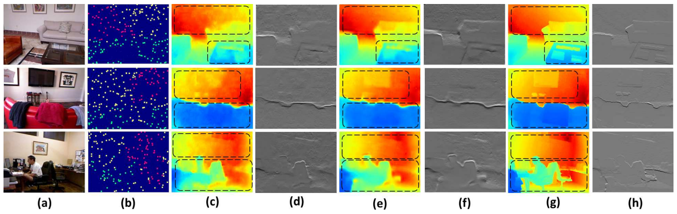

Fig. 9. Visual depth estimation results of the proposed S2DNet on NYU-Depth-V2 dataset for RGB image with sparse depth samples. (a) Input RGB image, (b) Sparse depth samples, (c) Results of the proposed S2DNet (w/o attention block), (d) Sobel gradient of depth map (estimated using the proposed S2DNet w/o attention block), (e) Results of the proposed S2DNet (w/i attention block), (f) Sobel gradient of depth map (estimated using the proposed S2DNet w/i attention block), (g) Ground truth image, and (h) Sobel gradient depth map of ground truth image. (*Best viewed in colors.)*

with the sparse depth samples are given as an input to the proposed S2DNet to estimate the scene depth map. Further, scene transmission map is estimated using relation given in [67]. We followed the similar approach as given in [65], [67], [71] to estimate the scene atmospheric light from input hazy scene and estimated scene transmission map. Fig. 10 shows the input hazy scene, sparse depth samples, recovered haze-free scene using proposed approach and ground truth haze-free scene.

Without any further training, we have utilized the trained S2DNet to estimate depth map from the given hazy scene and (200) sparse depth samples. We utilized entire D-Hazy dataset [72] which consists of 1449 hazy and respective haze-free scenes. We have carried both quantitative and qualitative analysis for image de-hazing task and given in Table VIII and Fig. 10 respectively. Table VIII witnessed the superiority of the proposed approach over the other existing methods for image de-hazing. Also, it can be observed from Fig. 10 that the recovered haze-free scene using proposed approach resembles closely to the ground truth haze-free scenes. From this discussion and results, it is clear that the proposed approach outperforms the existing methods for image de-hazing.

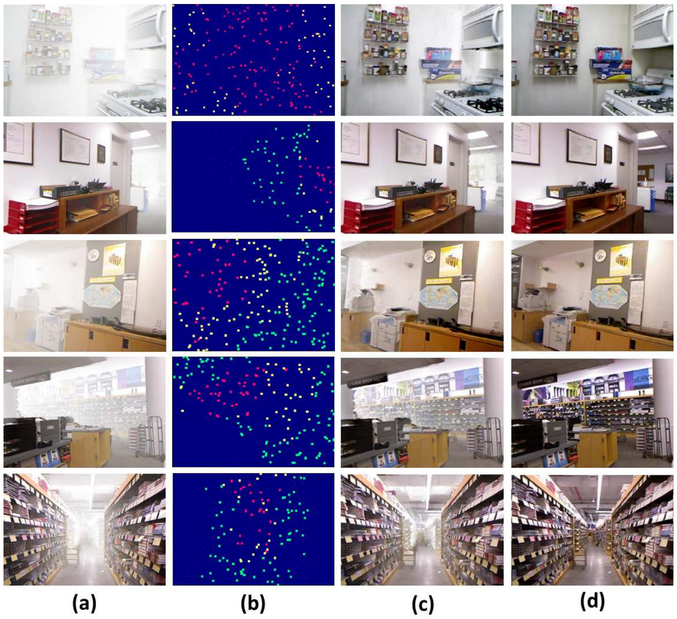

Fig. 10. Visual image de-hazing results of the proposed S2DNet on D-Hazy dataset for RGB image with sparse depth samples. (a) Input hazy image, (b) Sparse depth samples, (c) Results of the proposed S2DNet, and (d) Ground truth haze free image.

V. Conclusion

In this work, we addressed the intra scale ambiguity problem associated with the single image depth estimation (SIDE) by using single image with sparse depth samples. The proposed approach is useful for sensor fusion, self driving cars, SLAM and many other computer vision tasks. The proposed S2DNet overcomes the disadvantage of the existing learning approaches with the help of S2DCNet and S2DFNet architecture and uniform sparse samples strategy. Also, we analysed the effect of proposed S2DCNet and S2DFNet for coarse-level and fine-level depth estimation respectively. Correspondingly, the proposed S2DCNet predicts a coarse-level depth map from input images. Further, the estimated coarse-level depth map is concatenated with the input image and processed with the proposed S2DFNet for fine-level depth map estimation. The proposed S2DFNet comprises of the attention model which extracts prominent features and helps to estimate the precise (fine-level) depth map. Rigorous experiments have been carried out to prove the robustness of the proposed S2DNet to estimate the scene depth map. To analyse the performance of the proposed S2DNet, four benchmark datasets namely: NYU-Depth-V2 (indoor), KITTI odometry (outdoor), KITTI depth completion, and SUN RGB-D database have been utilized. The evaluation parameters used for the quantitative analysis are RMSE, $\log_{10}$, REL, and δ_i. The qualitative analysis has been carried out on benchmark datasets by comparing results of proposed with existing methods for depth estimation. Further, proposed approach of depth estimation is extended for single image de-hazing. Experimental analysis shows that proposed approach is superior than the existing state-of-the-art methods for both single image depth estimation and image de-hazing.

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*. Springer, 2012, pp. 746–760.

[2] A. Geiger, P. Lenz, and 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.

[3] Y. Huang, Y. Quan, Y. Xu, R. Xu, and H. Ji, "Removing reflection from a single image with ghosting effect," *IEEE Transactions on Computational Imaging*, 2019.

[4] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[5] M. Lawson, M. Brookes, and P. L. Dragotti, "Scene estimation from a swiped image," *IEEE Transactions on Computational Imaging*, 2019.

[6] X. Song, Y. Dai, and X. Qin, "Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 360–376.

[7] A. Halimi, R. Tobin, A. McCarthy, J. M. Bioucas-Dias, S. McLaughlin, and G. S. Buller, "Robust restoration of sparse multidimensional single-photon lidar images," *IEEE Transactions on Computational Imaging*, 2019.

[8] A. Lamża, Z. Wróbel, and A. Dziech, "Depth estimation in image sequences in single-camera video surveillance systems," in *International Conference on Multimedia Communications, Services and Security*. Springer, 2013, pp. 121–129.

[9] R. Horaud, M. Hansard, G. Evangelidis, and C. Ménier, "An overview of depth cameras and range scanners based on time-of-flight technologies," *Machine vision and applications*, vol. 27, no. 7, pp. 1005–1020, 2016.

[10] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE ICCV*, 2015, pp. 2650–2658.

[11] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 239–248.

[12] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.

[13] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[14] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2144–2158, 2014.

[15] J. Konrad, M. Wang, and P. Ishwar, "2d-to-3d image conversion by learning depth from examples," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 16–22.

[16] 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.

[17] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," in *ACM transactions on graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 577–584.

[18] S. Suwajanakorn, C. Hernandez, and S. M. Seitz, "Depth from focus with your mobile phone," in *Proceedings of the IEEE Conference on CVPR*, 2015, pp. 3497–3506.

[19] 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.

[20] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE Conference on CVPR*, 2015, pp. 5162–5170.

[21] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proceedings of the IEEE Conference on CVPR*, 2016, pp. 5506–5514.

[22] A. Chakrabarti, J. Shao, and G. Shakhnarovich, "Depth from a single image by harmonizing overcomplete local network predictions," in *Advances in Neural Information Processing Systems*, 2016, pp. 2658–2666.

[23] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," in *Proceedings of the IEEE Conference on CVPR*, 2018, pp. 175–185.

[24] Z. Zhang, C. Xu, J. Yang, J. Gao, and Z. Cui, "Progressive hard-mining network for monocular depth estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3691–3702, 2018.

[25] H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom, "Depth estimation from a single image using deep learned phase coded mask," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 298–310, 2018.

[26] 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 on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[27] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[28] 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.

[29] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5038–5047.

[30] M. Mancini, G. Costante, P. Valigi, T. A. Ciarfuglia, J. Delmerico, and D. Scaramuzza, "Toward domain independence for learning-based monocular depth estimation," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1778–1785, 2017.

[31] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 340–349.

[32] Y. Chen, C. Schmid, and C. Sminchisescu, "Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7063–7072.

[33] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia, "J-mod 2: joint monocular obstacle detection and depth estimation," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1490–1497, 2018.

[34] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2800–2810.

[35] 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.

[36] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia, "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4296–4303.

[37] S. Hawe, M. Kleinsteuber, and K. Diepold, "Dense disparity maps from sparse disparity measurements," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2126–2133.

[38] L.-K. Liu, S. H. Chan, and T. Q. Nguyen, "Depth reconstruction from sparse samples: Representation, algorithm, and sampling," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1983–1996, 2015.

[39] F. Ma, L. Carlone, U. Ayaz, and S. Karaman, "Sparse sensing for resource-constrained depth reconstruction," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 96–103.

[40] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 103–119.

[41] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3288–3295.

[42] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3313–3322.

[43] Y. Liao, L. Huang, Y. Wang, S. Kodagoda, Y. Yu, and Y. Liu, "Parse geometry from a line: Monocular depth estimation with partial laser observation," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5059–5066.

[44] C. Cadena, A. R. Dick, and I. D. Reid, "Multi-modal auto-encoders as joint estimators for robotics scene understanding." in *Robotics: Science and Systems*, vol. 5, 2016, pp. 1–9.

[45] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 175–185.

[46] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1511–1520.

[47] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1078–1085.

[48] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[49] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[51] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5354–5362.

[52] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on fourier domain analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 330–339.

[53] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proceedings of the IEEE Conference on CVPR*, 2015, pp. 2800–2809.

[54] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3917–3925.

[55] C. Fu, C. Mertz, and J. M. Dolan, "Lidar and monocular camera fusion: On-road depth completion for autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 273–278.

[56] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 89–96.

[57] M. Liu, M. Salzmann, and X. He, "Discrete-continuous depth estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 716–723.

[58] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *Proceedings of the IEEE Conference on CVPR*, 2015, pp. 1119–1127.

[59] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE Conference on CVPR*, 2015, pp. 5162–5170.

[60] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2018.

[61] F. Liu, G. Lin, and C. Shen, "Discriminative training of deep fully connected continuous crfs with task-specific loss," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2127–2136, 2017.

[62] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2008.

[63] Z. Huang, J. Fan, S. Yi, X. Wang, and H. Li, "Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion," *arXiv preprint arXiv:1808.08685*, 2018.

[64] 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.

[65] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[66] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 154–169.

[67] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[68] Q. Zhu, J. Mai, and L. Shao, "Single image dehazing using color attenuation prior." in *BMVC*. Citeseer, 2014.

[69] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4770–4778.

[70] D. Engin, A. Genç, and H. Kemal Ekenel, "Cycle-dehaze: Enhanced cyclegan for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 825–833.

[71] A. Dudhane and S. Murala, "Cˆ 2msnet: A novel approach for single image haze removal," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1397–1404.

[72] C. Ancuti, C. O. Ancuti, and C. De Vleeschouwer, "D-hazy: A dataset to evaluate quantitatively dehazing algorithms," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2226–2230.