2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Unsupervised Depth and Confidence Prediction from Monocular Images using Bayesian Inference

Vishal Bhutani[1], Madhu Vankadari[1], Omprakash Jha[1], Anima Majumder[1], Swagat Kumar[2] and Samrat Dutta[1]

***Abstract*— In this paper, we propose an unsupervised deep learning framework with Bayesian inference for improving the accuracy of per-pixel depth prediction from monocular RGB images. The proposed framework predicts confidence map along with depth and pose information for a given input image. The depth hypotheses from previous frames are propagated forward and fused with the depth hypothesis of the current frame by using Bayesian inference mechanism. The ground truth information required for training the confidence map prediction is constructed using image reconstruction loss thereby obviating the need for explicit ground truth depth information used in supervised methods. The resulting unsupervised framework is shown to outperform the existing state-of-the-art methods for depth prediction on the publicly available KITTI outdoor dataset. The usefulness of the proposed framework is further established by demonstrating a real-world robotic pick-and-place application where the pose of the robot end-effector is computed using the depth predicted from an eye-in-hand monocular camera. The design choices made for the proposed framework is justified through extensive ablation studies.**

I. Introduction

Depth estimation from RGB images is an active field of research finding application in a wide range of fields such as Augmented Reality [1], 3D graphics [2],[3] and robotics [4] . The deep learning based methods have been shown to outperform traditional methods that use hand crafted features and exploit camera geometry and/or motion to estimate depth. These learning based methods could be broadly classified into two categories - supervised or unsupervised depending on whether or not they require explicit ground truth depth information obtained from range sensors such as LiDAR. Since the availability of explicit ground truth poses a constraint which could not be met in many real world situations, there is a growing interest for unsupervised learning methods over the years aiming to overcome this limitation. These methods exploit the temporal and/or spatial consistencies present in the images to extract structural and motion information in the absence of ground truth depth data [5], [6], [7], [8]. The constraints for spatial consistency are derived from stereo or multi-view images while the constraints for temporal consistency is obtained by having a sequence of images. Monocular methods [6], [9], [10] that rely only on temporal consistency (optical flow motion) are shown to be inferior to stereo methods [7], [11] that additionally incorporate spatial consistency into their learning process. While the depth prediction accuracies have been increasing over the years, there is still enough scope for improvement as the current depth predictions are still not close to what is available from range sensors.

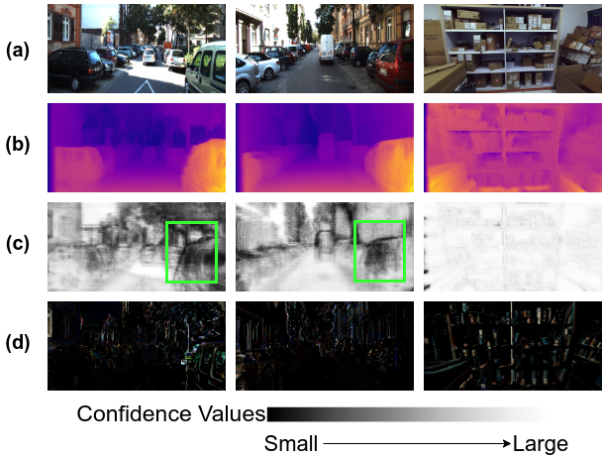

Fig. 1: A visual demonstration of depth estimation results applied to a few RGB monocular images. (a) shows the input image to the network - first two are randomly selected from KITTI outdoor dataset and last one is taken from our own indoor dataset. (b) shows the per-pixel depth predicted using our network (c) shows the confidence maps predicted by the network and (d) shows the image reconstruction error (darker pixels indicate lower error). The lighter regions in the confidence map signifies high confidence of predicted depth and darker regions represent low confidence. One such example can be observed in the highlighted rectangular region that, due to reflective window of the car in the RGB image, the network is not able to correctly predict disparity, thereby resulting into low confidence value.

It has been demonstrated recently that having a measure of model uncertainty (or confidence) can greatly influence the decision making process [12] [13]. The depth hypotheses from previous frames could be propagated and combined with the depth hypothesis for the current frame according to their respective uncertainty maps to smooth out abrupt errors, thereby, improving the accuracy of depth prediction for the current frame [12]. These confidence maps (inverse of model uncertainty) are predicted along with depth and pose information from RGB images using deep networks. The loss function necessary for training the confidence map prediction usually requires ground truth depth information as the supervision signal which might not be available in many real world situations. This constrains the applicability of the current approach and hence, provides motivation for our work. Instead of using ground truth depth data, we use reconstructed images (constructed using predicted depth maps) to compute the loss function required for training the confidence map prediction. The intuition for this comes from the fact that the quality of image reconstruction usually

The authors are associated with TCS Research, [1]TATA Consultancy Services, India and [2]Edge Hill University, UK. Email ID: {vk.bhutani, madhu.vankadari, omprakash.jha, anima.majumder and d.samrat}@tcs.com, swagat.kumar@edgehill.ac.uk

depends on the accuracy of the depth prediction. Higher depth accuracy leads to lower image reconstruction error and vice-versa. Thus, the image reconstruction error could be used to construct the ground truth for confidence map prediction. During the inference phase, the estimated depth and uncertainties (or confidences) of the previous frame are then propagated to the current frame by using a camera pose, proposed by Mur-Artal et al. [14]. The resultant multiple depth and uncertainty hypotheses generated for the current frame are then fused using the Bayesian inference framework for more robust and accurate depth estimation.

Unlike the previous approach, the proposed deep learning framework is trained using stereo pair of images to compensate for the absence of explicit depth information required for learning. Use of image reconstruction error for computing ground truth confidence map data overcomes an important limitation of the previous work while making it completely unsupervised. It is to be noted that the computation of image reconstruction loss does not necessarily require having stereo (or multi-view) images. It could also be computed from a temporal sequence of monocular images and hence, the above approach is applicable to all methods that either use monocular and stereo images during the training phase. Computationally, the proposed framework is more efficient compared to [13] that use a separate deep network to predict confidence map. The output of our network is shown in Figure 1. The first row in this figure shows the input images for which the predicted depth is shown in the second row. The confidence map and the corresponding image reconstruction error is shown in the third and fourth row respectively. The efficacy of the proposed approach is further established by applying it to a real world robotic pick-and-place application where the robot end-effector pose required for picking objects is determined by using the predicted depth information obtained from our network.

In short, the main contributions made in this paper are as follows:

- We provide an unsupervised version of Bayesian Depth Network that creates a new benchmark by outperforming existing state-of-the-art methods in this field.
- Creating ground truth confidence map using image reconstruction error is a novel contribution that has not been reported earlier in the literature.
- We demonstrate the applicability of the proposed framework in a real-world robot pick-and-place task using only a single monocular camera. Such applications usually make use of depth or RGBD cameras for reliable performance.
- Exhaustive performance analysis and ablation studies have been carried out to establish the superiority of our approach and justify the choice of various design parameters.

Rest of this paper is organized as follows. An overview of related works is provided in the next section. A detailed description of proposed deep learning framework is provided in Section III. Experimental results and the analysis are described in IV. Finally, Section V concludes the paper.

II. Related Work

Depth estimation is a long-lasting field of research in the community of computer vision and robotics. Existing literature in this area can be broadly categorized into four different groups; depth estimation using conventional computer vision techniques, supervised stereo approaches, deep network based supervised and unsupervised approaches using monocular images. A brief literature survey on each of these categories is given as follows.

A. Conventional and Supervised Stereo Approaches

One of the earliest conventional approaches for depth estimation without using additional hardware or depth sensors was by taking multiple images of the same scene with slight displacements. This was accomplished by matching key-points [15] that are common with each image and reconstructing a 3D model of the scene. Later in the year 2006, Saxena et al. [16] first introduced the concept of estimating per-pixel depth map of a scene from a single monocular images using supervised learning approach. They used Markov Random Field (MRF) that considers multi-scale local and global-image features as input to the training model. In a successive work [17], the authors incorporate the assumption of smoothness function in order to enforce the neighboring constraint. Few more works in this direction, include [18], [19] and [20]. However, one of the notable limitations of these approaches is that, most of these work rely on strong geometric constraints and hand-crafted features, thereby limiting its possibility to provide a generalized solution. Some other recent approaches are structure from motion (SfM)[21], monocular Simultaneous Localization and Mapping (SLAM)[22], binocular, and multi-view stereo [23]. However, all these approaches need explicit availability of multiple observations of the scene of interest taken at different viewing angles.

B. Supervised and Unsupervised Monocular Approaches

Recent advancements in deep learning techniques have motivated researches to estimate depth from a single image by training the deep learning models with monocular sequences of images. Literature in this direction can be categorized into supervised and unsupervised approaches. Few state-of-the-art works that use supervised training approaches include [24], [12], [25], [26], [27], [28]. All these approaches have shown promising results. However, supervised approaches need absolute depth data as ground-truth which is an important limitation. Most of the recent works therefore emphasize on an unsupervised way of solving this problem that avoids need for expensive ground-truth depth data, which is replaced by multiple views of the scene. Supervision signals are obtained by synthesizing images using predicted depth map and/or camera pose. Garg et al. [5] first introduced concept of unsupervised deep learning based depth estimation using only monocular sequences of images. Training include a pair of images (left-right stereo pair).

Given left images as input to the network, they generated inverse warp of the right image using the predicted depth and known inter-view displacement. Reconstruction of the input image is done by using the photometric error in the reconstruction as loss function. Further improvement is done by Godard et al. [29] by introducing a spatial transform network and then by [30] using a left-right consistency loss. Based on these aforesaid concepts lots of improvements are made for achieving better depth estimation. Some of the those include [6] [10], [7] and [11]. The work of Zhou et al. [6] is the first to introduce monocular depth estimation using images sequences from a single camera. This was further improved in [9], [31], [32], where the authors included additional cues, such as point-cloud alignment [9], differentiable DVO [31] and multi-task learning [32].

III. Method

This section gives a detailed description of the proposed unsupervised deep learning framework to estimate depth, pose, and confidence map for a given sequence of monocular images. The proposed framework is shown in Figure 2 and it comprises of three main modules namely, Disp-Net, Pose-Net [7] and Bayesian Inference module. The Disp-Net takes a monocular image as an input and predicts per-pixel depth-map D and a confidence-map C. The Pose-Net takes a sequence of $n-$monocular images and predicts $n-1$ 6-DOF pose vectors, P, between the consecutive frames. Finally, the Bayesian inference module is designed to fuse the previous $t-n$ temporally aligned depth maps with the depth obtained using the predicted uncertainty maps and pose-vectors at time t. All of these modules are delineated in the following subsections.

A. Monocular Depth Estimation (Disp-Net)

One of our main objectives is to predict a per-pixel depth map, D by learning a mapping function F in an unsupervised manner for a given monocular image I. Instead of predicting depth maps directly, we predict the correspondence-map between left and right image i.e, disparity map d to obtain the depth. For instance, given a rectified stereo image pair $\{I_l, I_r\}$, the function F takes the left image I_l and predicts a dense correspondence maps d_{rl}(right-to-left disparity) and d_{lr}(left-to-right disparity). The predicted disparities are then used to reconstruct the stereo images from one another using the spatial reconstruction module [7]. The left image is reconstructed using the right image I_r and left-to-right disparity d_{lr}. Similarly, right image is also reconstructed using left-image I_l and right-to-left disparity d_{rl}. The depth map D can be calculated as $D = \frac{bf}{d}$ where, b is the baseline distance of the stereo rig, f is the rectified camera focal length and $d = d_{lr}$ is the predicted disparity. These estimated left I'_l and right I'_r are compared against the original left and right images to calculate the photo-metric losses required for training the network. Since the model takes only left-image as input, it can work as a monocular-depth estimation network during the inference time.

B. Pose estimation from monocular image sequence (Pose-Net)

Given a pair of consecutive monocular images, the Pose-Net predicts 6-DOF ego-motion of the camera between them. For instance, given a pair of images (I_{t-1}, I_t) from either left or right camera, the Pose-Net predicts a vector P_t^{t-1} comprising of translation (t_x, t_y, t_z) and rotation (ρ, θ, ψ) of the camera between the frames. The predicted pose information P_t^{t-1} and the depth of the t-th frame D_t is used to reconstruct the I_t using a temporal reconstruction module [7] from I_{t-1}. The reconstructed image I'_t is compared against the original I_t to calculate the losses necessary for the training.

C. Confidence and Uncertainty Maps Estimation

The confidence map C is predicted along with the disparities from the Disp-Net. The confidence map values lie between [0,1] where, higher the confidence value signifies network is much more certain about its predictions and correspondingly for the lower confidence value. Previous approaches such as [12] used ground-truth depth-map to estimate a reference confidence to evaluate the predicted-confidence map. However, our approach is a completely unsupervised learning method and the ground-truth depth is not available. To circumvent this problem, we use a negative exponential of the photo-metric error calculated using the reconstructed left and right images as shown in Equation 1. By doing this, pixel having large photometric error will map to a smaller confidence value and pixel having small photo-metric error will map to a larger confidence value. The motivation behind this comes from the fact that the accuracy of the estimated disparity defines the reconstruction quality of the image. In other words, if the predicted disparity is accurate, the images will get reconstructed accurately and the error between the original and reconstructed image will be small. Mathematically, the ground-truth confidence is calculated as:

$$C_g = e^{-|I_l - I'_l|} + e^{-|I_r - I'_r|} \tag{1}$$

We assume that the depth measurement of a pixel (u, v) belongs to a normal distribution with a mean $D(u, v)$ and variance σ^2. The per-pixel uncertainty-map U is obtained using the predicted confidence C as $\sigma^2 = ln^2(C) = U$.

D. Detailed Network Architectures

The Disp-Net is a fully convolutional encoder-decoder style network with four output layers. Each of the output layer predicts two disparity maps (left-right and right-left disparities) and a confidence map. The predicted disparities are normalized to have maximum value as 30% of input-image width using a sigmoid activation function and the confidence values are normalized to be between 0 and 1. There are multiple skip-connections attached to different stages of decoder from the encoder in-order to facilitate the exchange of low-dimensional information while estimating disparity and confidence maps. The Pose-Net is composed

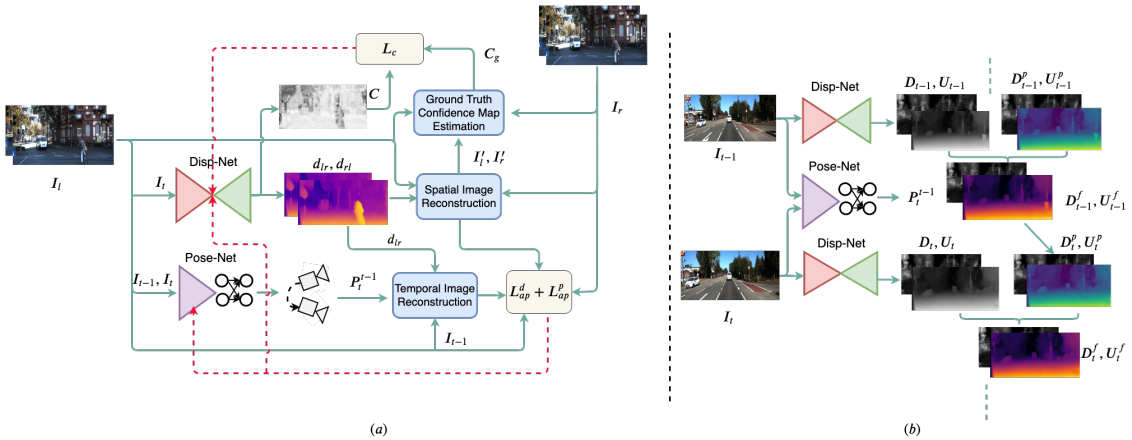

Fig. 2: An architectural overview of the proposed method. (a) Disp-Net takes left images($I_t \epsilon I_l$) as input and predicts left-right and right-left disparities $d = (d_{lr}, d_{rl})$ and obtains confidence map C. Pose-Net takes a pair of images $(I_t, I_{t-1})\epsilon I_l$ and estimates a 6-DOF pose-vector P_t^{t-1} of the camera between the frames. The predicted disparities and pose-information are given to view-reconstruction modules for image reconstruction. The reconstructed images are compared against their original images for appearance losses. The reconstructed images are further used to estimate ground-truth confidence which is compared against the predicted confidence to calculate the loss. (b) The Bayesian inference module takes a sequence of images (I_{t-n}, I_{t-n-1},I_{t-n-2},.....I_{t-1},I_t) and predicts fused depth map at t.

of a convolutional encoder followed by two fully connected layers that output a 6-DOF pose vectors for a given pair of images. Relu is used as an activation function in both of the networks except in the output layers. The outputs of the Pose-Net are not processed through any activation. A detailed architectural overview of the proposed network is shown in Figure 2.

E. Bayesian Inference Module

The Bayesian inference module is introduced to make the predicted depth D_t at t more accurate by using $t-n$ previous depth, pose and uncertainty information. Given a pair of estimated depth-maps D_{t-1}, D_t, uncertainty-maps U_{t-1}, U_t and the transformation between the frames $P_t^{t-1} = [R, T]$, a pixel $(u, v)^T$ in depth-map D_{t-1} can be propagated into D_t using the camera calibration matrix K as follows

$$P = K^{-1}(u, v, 1)^T$$
$$D_t^p(u', v') = K[RD_{t-1}(u, v)P + T] \tag{2}$$

In the same way, the uncertainty map is also propagated from $t-1$ to t-th frame as

$$U_t^p = JU_{t-1}J^T \tag{3}$$

where, $J = \frac{\partial D_t^p}{\partial D_{t-1}}$ and it is mathematically difficult to calculate due to the non-linear relationship between D_t^p and D_{t-1}. For simplicity and efficiency, we assume that the change in rotation of the camera between two consecutive frames is very small which is often the case in real-world and hence, the rotation matrix R can be replaced with an identity matrix. With this assumption, the equations 2, 3 can be written as follows

$$D_t^p = D_{t-1} + T_z \tag{4}$$
$$U_t^p \approx U_t + \sigma_w^2 \tag{5}$$

where T_z is the translation of the camera in z-direction and σ_w^2 is white Gaussian noise with mean zero and variance of σ^2. The noise is added to supplant the inherent noises that are present in the pose-estimation.

F. Depth and Uncertainty Fusion

The predicted depth map D_t and uncertainty map U_t at time t, are fused with the propagated depth map D_t^p and uncertainty map U_t^p to get a final fused depth map D_t^f and uncertainty map U_t^f. The detailed overview of the proposed method is depicted in Figure 2. Using the Bayesian inference, the fused depth is defined as:

$$P\left(D_t^f | D_t\right) = \frac{P\left(D_t | D_t^p\right) P(D_t^p)}{P\left(D_t\right)} \tag{6}$$

Probability of the propagated depth is represented by prior probability $P\left(D_t^p\right)$ and the probability of the predicted depth is represented by the likelihood $P\left(D_t | D_t^p\right)$. Based on the assumption that the depth measurement of a pixel belongs to a normal distribution, $P\left(D_t | D_t^p\right) \sim \mathcal{N}(D_t, U_t)$, $P\left(D_t^p\right) \sim \mathcal{N}(D_t^p, U_t^p)$, $P\left(D_t\right)$ is constant and $P\left(D_t^f | D_t\right)$ $\sim \mathcal{N}(D_t^f, U_t^f)$.

$$P\left(D_t^f | D_t\right) \propto \frac{1}{\sqrt{2\pi U_t}} \exp\left(-\frac{(x - D_t)^2}{2U_t}\right) \times \frac{1}{\sqrt{2\pi U_t^p}} \exp\left(-\frac{(x - D_t^p)^2}{2U_t^p}\right) \tag{7}$$

$$P\left(D_t^f | D_t\right) \propto \frac{1}{\sqrt{2\pi U_t^f}} \exp\left(-\frac{\left(x - D_t^f\right)^2}{2U_t^f}\right) \tag{8}$$

Using equation (7) and (8) the fused depth and uncertainty maps at time t can be written as $D_t^f = \frac{D_t^p U_t + D_t U_t^p}{U_t^p + U_t}$ and $U_t^f = \frac{U_t^p U_t}{U_t^p + U_t}$. But, it is important to note that the assumptions of Jacobian and rotation matrices fail when there is a large camera rotation between the consecutive image frames and hence the Bayesian fusion can not be used.

G. Loss Functions:

Multiple loss functions are used to train the proposed model. Each of these are explained below in this section.

1) Appearance Loss: Appearance loss is calculated between the reconstructed images obtained from the spatial and temporal reconstruction modules and their respective original images. This loss function makes sure the that the reconstructed images and their original images are same in all the aspects such as color and structure. Two metrics namely, L1 and Structural Similarity Index (SSIM)[33] are used to calculate this loss. The convex combination of these two metrics is referred as appearance or photo-metric loss and mathematically, it can be written as follows.

$$L_{ap} = \frac{1}{N}\sum_{ij}\frac{\alpha}{2}\tilde{\rho}(1-SSIM(I_{ij}, I'_{ij}))+(1-\alpha)\tilde{\rho}(I_{ij}-I'_{ij}) \tag{9}$$

where, $I'\epsilon(I'_l, I'_r, I'_t)$ and I is its respective original image. $\alpha = 0.85$ is a weighted factor, and $\tilde{\rho}$ is the charbonnier penalty function [7].

2) Disparity Smoothness Loss: Smoothness loss encourages the predicted disparities to be locally smooth between the strong gradient regions like edges to preserve the spatial layout of the scene. To achieve this, the disparity gradients ∂d are weighted with negative exponential image gradients ∂I and the same can be mathematically written as:

$$L_{ds} = \frac{1}{N}\sum_{ij}\tilde{\rho}(\partial_x d_{ij} e^{-|\partial_x I_{ij}|}) + \tilde{\rho}(\partial_y d_{ij} e^{-|\partial_y I_{ij}|}) \tag{10}$$

where, $d\epsilon(d_{lr}, d_{rl})$ and $I\epsilon(I_l, I_r)$ respectively.

3) Left-Right consistency loss: Proposed Disp-Net predicts two (left-right and right-left) disparity-maps [30]. The left-right consistency loss enforces the cycle-consistency constraint between the predicted disparities by predicting left-right disparity from right-left disparity and vice versa. This loss makes the predicted disparities to be coherent with each other and improves the overall performance. The loss is calculated for both the disparities. The loss with left-right disparity can be mathematically written as follows,

$$L^l_{lr} = \frac{1}{N}\sum_{ij}\left\| d^{ij}_l - d^{ij+d^{ij}_l}_r \right\| \tag{11}$$

4) Confidence Loss: Confidence loss is calculated as L_1 difference between the ground-truth and predicted confidences. As the direct ground-truth confidence is not available, we used the confidence map calculated in section III-C as a reference. The loss can be defined as:

$$L_{cl} = \sum_{ij}|C_{ij} - (C_g)_{ij}| \tag{12}$$

We define a loss function L_{total} combining all the aforesaid loss functions and it can be written as:

$$L = \mu_1 L_{ap} + \mu_2 L_{ds} + \mu_3 L_{lr} + \mu_4 L_{cl} \tag{13}$$

where, $L_{ap} = (L^d_{ap} + L^p_{ap})$ is the appearance loss calculated with the disparity L^d_{ap} and predicted poses L^p_{ap}. $L_{ds} = (L^l_{ds} + L^r_{ds})$ is disparity smoothness loss, $L_{lr} = L^l_{lr} + L^r_{rl}$ is left-right consistency loss and $L_{cl} = L^l_{cl} + L^r_{cl}$ is confidence loss. The μ's are weight co-efficients given to each loss. In the next section, we delineate the experiments and evaluations performed to prove the efficacy of the propose method.

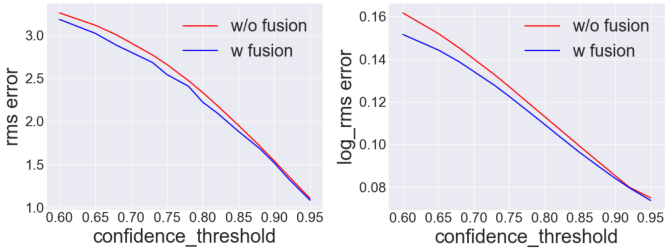

Fig. 3: Average RMS and Log RMS error plots for depth estimation on a continuous sequence of images taken from KITTI dataset. These are obtained by varying the threshold value of the predicted confidence. Comparisons have been made for both using and without using the Bayesian inference.

IV. Experiments and Results

A. Implementation Details

The proposed network architecture is implemented in Tensorflow[39]. There are around 12M trainable parameters and it takes 30 hours to train the network on a GTX 1080 GPU machine. The input image resolution is set to 256×512 and the batch size is set to 4. Adam optimizer[40] is used for minimizing the loss function, with $\beta_1 = 0.9$, $\beta_2 = 0.99$. The initial learning rate is set to $1e^{-04}$. It gets reduced by half after completing $(3/5)^{th}$ of total iteration and further reduced by half after completing $(4/5)^{th}$ of the total number of iterations. The weights μ_1, μ_3 and μ_4 in the loss function are set to 1.0 and μ_2 to 0.1. We have used data augmentation to reduce the possibility of over fitting which include change in brightness, change in saturation and change in gamma in a range of [0.5,2.0], [0.8, 1.2] and [0.8,1.2] respectively. In the following subsections, details about both qualitative and quantitative evaluation on KITTI and Indoor Datasets are given.

B. Depth and Confidence Evaluation using KITTI and Indoor Datasets

KITTI-Dataset: KITTI[41] is a popular outdoor driving dataset to benchmark the efficacy of the depth-estimation methods in this domain. It is comprised of 61 different outdoor driving sequences with 42,382 images of resolution 1242×345. As per the literature, the dataset is divided into two splits, namely *kitti-split* and *eigen-split* which are commonly used for evaluating the depth estimation accuracy. We demonstrate the performance of the proposed method using the *eigen-split* [24] for a fair comparison with the state-of-the-art methods. Further details on Eigen split can be found in [7].

Confidence map is evaluated indirectly using the ROC curve analysis as there is no respective ground-truth information is available. This is carried out by considering the

TABLE I: Performance comparison of the proposed network architecture on eigen split with existing state-of-the-art methods. The depth maps for the eigen split are computed using velodyne laser data and Eigen[24] is used for fair comparison. The table is divided into two sections, first section considers maximum depth of 80m using Garg crop and the second section uses maximum depth of 50m. The column *Supervision* refers D as Depth, M as Monocular, MS as Monocular Stereo. These are the supervisions used while training. The method 'Ours w fusion' is our proposed network with Bayesian inference and 'Ours w/o fusion' is the proposed network without Bayesian inference.

Method	Supervision	Abs Rel	Sq Rel	RMSE	logRMSE	δ <1.25	δ <1.25^2	δ <1.25^3
Train set mean	D	0.361	4.826	8.102	0.377	0.638	0.804	0.894
Eigen et al.'14 [24] Fine	D	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al.'15 [26]	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
SfM Learner'17 [6]	M	0.208	1.768	6.856	0.283	0.678	0.885	0.957
UnDeepVO'18 [34]	MS	0.183	1.73	6.57	0.283	-	-	-
Vid2Depth'18 [9]	M	0.1631	1.240	6.220	0.250	0.762	0.9160	0.968
Monodepth'17 [30]	MS	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Garg et al.'16 [5]	MS	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Chen et al.'20 [13]	MS	0.145	1.267	5.786	0.244	0.811	0.925	0.965
DeepFeat-VO'17 [6]	MS	0.144	1.391	5.869	0.241	0.803	0.928	0.969
UnDEMoN'18 [7]	MS	0.139	1.174	5.59	0.239	0.812	0.930	0.968
GeoNet'18 [10]	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
GAN-VO'18 [35]	M	0.150	1.141	5.448	0.216	0.808	0.939	0.975
Vankadari et al.'19 [36]	MS	0.1269	0.9982	5.309	0.226	0.827	0.93	4 0.971
EPC ++'18 [37]	MS	0.128	0.935	5.011	0.209	0.831	0.945	0.979
SuperDepth'18 [38]	S	0.112	0.875	4.958	0.207	0.852	0.947	0.977
Bayesian DeNet'19 [12]	S	0.112	-	4.867	0.184	-	-	-
Monodepth2'18 [29]	MS	0.106	0.818	4.750	0.196	0.874	0.957	0.979
Ours w/o fusion	MS	**0.091**	**0.742**	4.389	0.188	0.901	0.958	0.978
Ours w fusion	MS	0.095	0.788	**4.386**	**0.182**	**0.902**	**0.960**	**0.980**
Garg et al.'16 [5]	MS	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Monodepth'17 [30]	MS	0.140	0.976	4.471	0.232	0.818	0.931	0.969
SfM Learner'17 [6]	M	0.201	1.391	5.181	0.264	0.696	0.900	0.966
Vid2Depth'18 [9]	M	0.155	0.927	4.549	0.231	0.781	0.931	0.975
Zhan et.al'18 [8]	MS	0.135	0.905	4.366	0.225	0.818	0.937	0.973
Chen et al.'20 [13]	MS	0.138	0.937	4.399	0.231	0.825	0.933	0.969
UnDEMoN'18 [7]	MS	0.132	0.884	4.290	0.226	0.827	0.937	0.972
Vankadari et al.'19 [36]	MS	0.1207	0.7490	4.051	0.214	0.840	0.941	0.975
Ours w/o fusion	MS	**0.086**	**0.542**	3.331	0.178	**0.910**	0.961	0.980
Ours w fusion	MS	0.090	0.574	**3.326**	**0.172**	**0.910**	**0.963**	**0.982**

TABLE II: An ablation study on performance comparison of the proposed Bayesian inference module. As the number of past n frames increases for the fusion, RMSE and logRMSE error increases and accuracy decreases. Our proposed network using Bayesian inference is given as 'Ours w fusion'.

Method	Abs Rel	Sq Rel	RMSE	logRMSE	δ <1.25	δ <1.25^2	δ <1.25^3	n
Ours w fusion	0.0830	0.4635	3.379	0.160	0.912	0.965	0.986	1
Ours w fusion	0.1035	0.5582	3.527	0.168	0.899	0.963	0.986	2
Ours w fusion	0.1311	0.6992	3.751	0.185	0.872	0.958	0.984	3
Ours w fusion	0.1608	0.8736	4.009	0.207	0.833	0.950	0.981	4
Ours w fusion	0.1915	1.0756	4.283	0.230	0.781	0.940	0.977	5

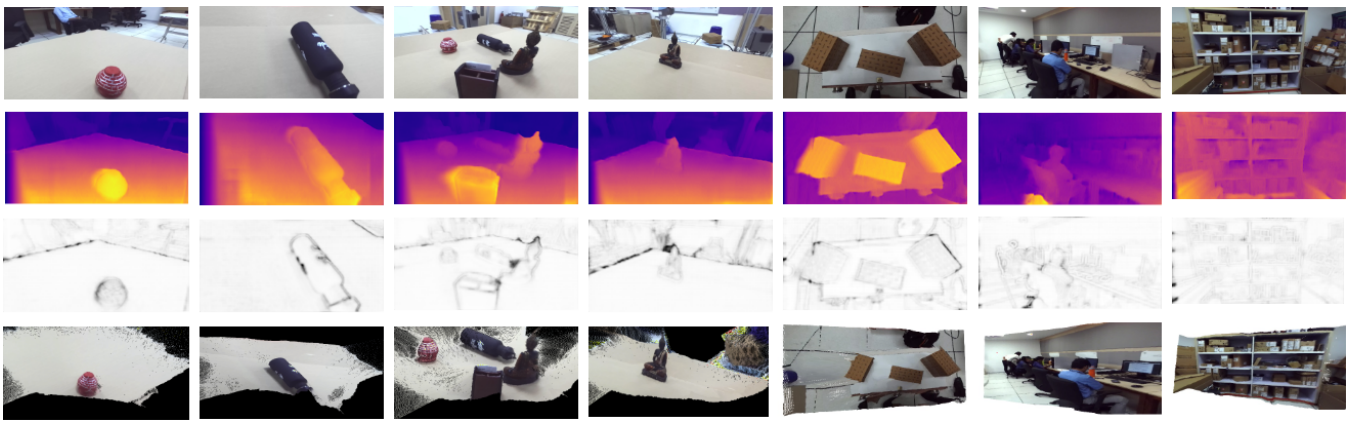

Fig. 4: An illustration of a few images taken from the Indoor database and the corresponding prediction results. (a) Few randomly selected images from the Indoor dataset, (b) corresponding disparity maps, (c) corresponding confidence maps and (d) generated point-clouds for each image.

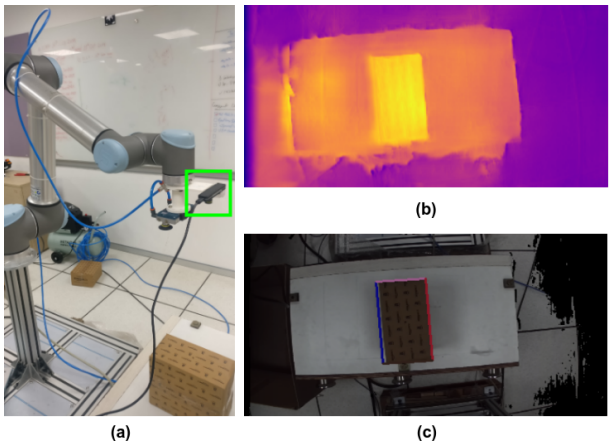

Fig. 5: Camera and manipulator setup for pick-and-place task (a) ZED-M camera mounted at the end effector of the manipulator (b) Disparity map predicted from our trained model (c) Detected bounding box using region growing algorithm.

pixels which have confidence values grater than a threshold τ. Two different error metrics: RMSE and Log RMS error are used to demonstrate the consistency of the estimated confidence with the predicted depth. The same is depicted in Figure 3. One can observe, that with the increase of confidence threshold value the error decreases. This is also reflected in the depth prediction results, where the proposed net with fusion outperforms the network without fusion. This is also illustrated in Figure 1 that for high re-projection errors the predicted confidences become low. The snippet of sequence of images taken from KITTI dataset for evaluation of confidence is mentioned in section IV-C.

Pick and Place Operation: To further demonstrate the usability of the estimated depth and confidence maps in the real-world applications, a pick-and-place task is performed using an industrial manipulator UR10. We collected a dataset of 22K images with a resolution of 720×1280 having 5 different object configurations using a ZED-Mini stereo camera. The estimated depth-maps, confidence maps and 3D-reconstructed point clouds are depicted in Figure 4. One can observe that the proposed method is able to generate plausible depth maps and retain most of the structural attributes of the objects. The reconstructed point clouds are further processed using the region-growing algorithm [42] to estimate a possible position for picking the object. An illustration for this is shown in Figure 5. Once the picking position is known, the manipulator plans the trajectory to carry out the picking operation to a pre-defined placing point. The entire demonstration can be found at `https://youtu.be/oHjdYephI7k`

C. Ablation Study

Along with the above mentioned experiments, we have also performed rigorous ablation studies. First ablation study is carried out by changing the number of frames n that are used for fusion. A random snippet of images is taken from KITTI dataset (with sequence number $2011\_09\_26\_drive\_0093\_sync$). The depth evaluation metrics are calculated on that snippet by varying the number of previous n frames used for fusion. Experimental results are shown in Table II. It can be observed that, on increasing the number of past n frames, the corresponding error increases, thereby reduces the depth prediction accuracy. It is due to the fact that, with increasing number of frames, overlapping areas/correspondences between the past and current frame decreases and hence the Bayesian inference cease to work. Second ablation study is carried out for indoor environment depth prediction. We collected a dataset of 18K images with a resolution of 720×1280 of an indoor environment of our workplace. The model is trained to predict depth maps, then 3D point cloud is reconstructed using the camera intrinsic parameter K. Qualitative results for the indoor scene is shown in 6^{th} and 7^{th} column of Figure 4.

V. Conclusions

This work presents an unsupervised approach for depth and confidence-map estimation from a monocular image using Bayesian inference. A loss function is proposed by combining four different losses; appearance loss, disparity smoothness loss, left-right consistency loss and confidence loss. We validate and demonstrate the efficacy of our approach by performing several experiments on both KITTI and our own collected indoor dataset. We compare the performance of our network with several existing state-of-the-art techniques. The presented results show that our depth estimation accuracy outperforms the current state-of-the-art techniques, [29], [12] and [13], thereby creating a new benchmark in the field. The proposed framework is validated through several ablation studies. It is experimentally observed that incorporation of predicted depths and uncertainties (inverse of confidence) from immediate previous frame provides improved depth results at the time of Bayesian inference. Whereas, on considering more previous frames, the performance degrades as the correspondence/overlapping regions between current frame and previous frames decreases. Currently, we have used image sequences of either day-time or taken in an indoor environment. However, we did not address environmental conditions, such as night-time images with sudden changes in lighting conditions, or sequences of images with a very-low illumination. An extension of our work can be done in this direction by taking advantages of GANs based domain adaptation technique.

References

[1] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2016.

[2] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium*, Baden-Baden, Germany, June 2011.

[3] Minwoo Park, Jiebo Luo, Andrew C Gallagher, and Majid Rabbani. Learning to produce 3d media from a captured 2d video. *IEEE Transactions on Multimedia*, 15(7):1569–1578, 2013.

[4] Carlos Sampedro, Alejandro Rodriguez-Ramos, Ignacio Gil, Luis Mejias, and Pascual Campoy. Image-based visual servoing controller for multirotor aerial robots using deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 979–986. IEEE, 2018.

[5] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.

[6] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

[7] V Madhu Babu, Kaushik Das, Anima Majumdar, and Swagat Kumar. Undemon: Unsupervised deep network for depth and ego-motion estimation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1082–1088. IEEE, 2018.

[8] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian 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*, pages 340–349, 2018.

[9] Reza Mahjourian, Martin Wicke, and Anelia 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*, pages 5667–5675, 2018.

[10] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018.

[11] Anima Majumder Madhu Vankadari, Swagat Kumar and Kaushik Das. Unsupervised learning of monocular depth and ego-motion using conditional patchgans. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, 2019.

[12] Xin Yang, Yang Gao, Hongcheng Luo, Chunyuan Liao, and Kwang-Ting Cheng. Bayesian denet: Monocular depth prediction and frame-wise fusion with synchronized uncertainty. *IEEE Transactions on Multimedia*, 2019.

[13] Long Chen, Wen Tang, Tao Ruan Wan, and Nigel W John. Self-supervised monocular image depth learning and confidence estimation. *Neurocomputing*, 381:272–281, 2020.

[14] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[15] Martin Bujnak, Zuzana Kukelova, and Tomas Pajdla. 3d reconstruction from image collections with a single known focal length. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1803–1810. IEEE, 2009.

[16] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

[17] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Learning 3-d scene structure from a single still image. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[18] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. 3-d depth reconstruction from a single still image. *International journal of computer vision*, 76(1):53–69, 2008.

[19] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1253–1260. IEEE, 2010.

[20] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–96, 2014.

[21] Sid Yingze Bao and Silvio Savarese. Semantic structure from motion. In *CVPR 2011*, pages 2025–2032. IEEE, 2011.

[22] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016.

[23] Adarsh Prakash Murthy Kowdle and Richard S Szeliski. Depth estimation using multi-view stereo and a calibrated projector, December 31 2015. US Patent App. 14/319,641.

[24] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[25] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[26] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015.

[27] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3917–3925, 2018.

[28] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[29] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *arXiv preprint arXiv:1806.01260*, 2018.

[30] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[31] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.

[32] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–53, 2018.

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[34] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7286–7291. IEEE, 2018.

[35] Yasin Almalioglu, Muhamad Risqi U Saputra, Pedro PB de Gusmao, Andrew Markham, and Niki Trigoni. Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. *arXiv preprint arXiv:1809.05786*, 2018.

[36] Madhu Vankadari, Swagat Kumar, Anima Majumder, and Kaushik Das. Unsupervised learning of monocular depth and ego-motion using conditional patchgans. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5677–5684. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[37] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *arXiv preprint arXiv:1810.06125*, 2018.

[38] Sudeep Pillai, Rares Ambrus, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. *arXiv preprint arXiv:1810.01849*, 2018.

[39] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, and Matthieu Devin. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[41] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[42] Shilpa Kamdi and RK Krishna. Image segmentation and region growing algorithm. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, 2(1), 2012.