2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Probabilistic TSDF Fusion Using Bayesian Deep Learning for Dense 3D Reconstruction with a Single RGB Camera

Hanjun Kim and Beomhee Lee

***Abstract*— In this paper, we address a 3D reconstruction problem using depth prediction from a single RGB image. With the recent advances in deep learning, depth prediction shows high performance. However, due to the discrepancy between training environment and test environment, 3D reconstruction can be vulnerable to the uncertainty of depth prediction. To consider the uncertainty of depth prediction for robust 3D reconstruction, we adopt Bayesian deep learning framework. Conventional Bayesian deep learning requires a large amount of time and GPU memory to perform Monte Carlo sampling. To address this problem, we propose a lightweight Bayesian neural network consisting of U-net structure and summation-based skip connections, which is performed in real-time. Estimated uncertainty is utilized in probabilistic TSDF fusion for dense 3D reconstruction by maximizing the posterior of TSDF value per voxel. As a result, global TSDF robust to erroneous depth values can be obtained and then dense 3D reconstruction from the global TSDF is achievable more accurately. To evaluate the performance of depth prediction and 3D reconstruction using our method, we utilized two official datasets and demonstrated the outperformance of the proposed method over other conventional methods.**

I. INTRODUCTION

In autonomous systems, mobile agents require 3D reconstruction to recognize surrounding environments. For instance, to realistically interact with real and virtual worlds, such as AR and MR, dense 3D reconstruction is desired on the environments that the agents can perceive. In recent years, the applications of a single camera to autonomous systems have been actively studied [1]–[3] due to the following advantages: economical, lightweight, and intuitive like a human vision system. However, 3D reconstruction with a single camera has intrinsic problems; dense correspondences between two frames are hardly found in a featureless environment and the absolute scale cannot be measured without a prior knowledge. To overcome these disadvantages, depth estimation from a single image using a deep learning is introduced [4]–[7]. With this method, dense depth image can be estimated even in the untextured environments, and the scale-ambiguity problem can also be resolved.

There are three main ways of using depth prediction for dense 3D reconstruction and tracking. First, dense depth is predicted for the keyframes only and then the depth is refined by finding geometric correspondences between the keyframe and the current frame [8], [9]. Second, sparse depth is obtained by a general visual odometry algorithm and used to improve the performance of depth prediction [10]. Third, RGB-D SLAM is performed using an RGB image and an estimated depth image pair as if using an RGB-D camera [7]. We follow the last stream to perform dense 3D reconstruction using depth prediction. However, unlike the existing method of 3D reconstruction based on point-based fusion [7], we perform a dense 3D reconstruction based on truncated signed distance function (TSDF) fusion. By using TSDF fusion, noise of the data can be relaxed through a weighted moving average, and the 3D surfaces can be readily extractable from the global TSDF map using raycasting or marching cube algorithm. In addition, probabilistic TSDF fusion can be performed considering the sensor noise model for a more robust 3D reconstruction to the noise [11], [12].

However, traditional depth prediction cannot obtain the sensor noise model because deterministic networks with a single output are insufficient to capture the noise, i.e. uncertainty. In this paper, we propose a probabilistic TSDF fusion by capturing the uncertainty of depth prediction using Bayesian deep learning. Bayesian neural network (BNN)

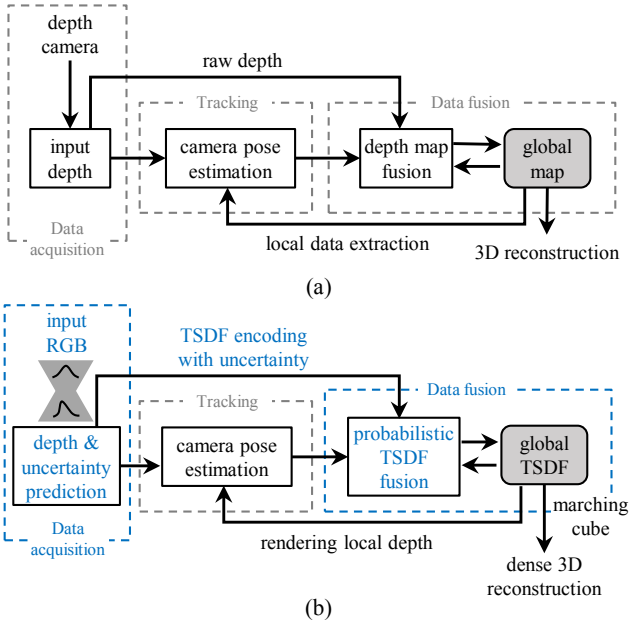

Fig. 1. **Comparison of pipeline between typical 3D reconstruction and the proposed method.** (a) Pipeline of typical 3D reconstruction with a depth camera. (b) Pipeline of 3D reconstruction using probabilistic TSDF fusion with depth and uncertainty prediction.

*This work was supported in part by the Brain Korea 21 Plus Project and in part by a Bio–Mimetic Robot Research Center funded by Defense Acquisition Program Administration, and by Agency for Defense Development (UD190018ID)

Hanjun Kim and Beomhee Lee are with the Automation and Systems Research Institute (ASRI), Department of Electrical and Computer Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea {k3k5good, bhlee}@snu.ac.kr

estimates two main types of uncertainties: aleatoric uncertainty and epistemic uncertainty [13]. Aleatoric uncertainty is implicit noise in the output of the model and is estimated by variance of the output using mixture density network [14]. Epistemic uncertainty is caused by the lack of knowledge (or data), and is modelled by giving a prior distribution to the parameters of the model. BNN is able to capture epistemic uncertainty using dropout variational inference [15], [16].

However, real-time application is difficult in the case of capturing epistemic uncertainty, because Monte Carlo (MC) dropout sampling is hardly implemented in parallel, which consumes a large amount of GPU memory [17]. To estimate epistemic uncertainty and aleatoric uncertainty with a single phase, we propose a simple BNN with U-net structure using summation-based skip connections. Estimated uncertainty in each output is propagated using an incremental Bayesian learning in probabilistic TSDF fusion, and the value in global TSDF is averaged using the uncertainty as weight. As a result, a global TSDF robust to the uncertainty of depth prediction is obtained, and the 3D surfaces composed of meshes can be extracted from the global TSDF. The pipelines of typical 3D reconstruction method and the proposed method are summarized in Fig. 1. In Fig. 1 (b), the depth and the uncertainty are predicted via BNN in the data acquisition step, and the predicted depth is converted to TSDF value for TSDF fusion. In the data fusion step, the depth and the uncertainty are used for probabilistic TSDF fusion, and the dense 3D surfaces are reconstructed accordingly.

In order to verify the proposed method and compare it with other methods, experiments were conducted in two views: depth prediction task of the proposed network and 3D reconstruction task of the proposed weight option for TSDF fusion. To train and validate the depth prediction network, the NYU-Depth-V2 dataset [18] was used; in addition, to verify the accuracy of 3D reconstruction, 3D surface model of ICL-NUIM [19] dataset was used.

The main contributions of this paper are as follows:

- We predict a depth image from a single RGB image using our simple and lightweight BNN which estimates the uncertainty of depth prediction with a single phase through GPU parallelization.
- TSDF values of the global TSDF are updated probabilistically using estimated depth and uncertainty. The uncertainty is propagated by incremental Bayesian learning rule.
- Dense 3D reconstruction, which is robust to the uncertainty of depth prediction, can be obtained by applying marching cube algorithm to the global TSDF.

The remainder of this paper is summarized as follows. Depth and uncertainty prediction tasks and the proposed lightweight BNN are demonstrated in Sec. II. Probabilistic TSDF fusion is performed using Bayesian learning rule in Sec. III. Implementation details for the experiments are explained in Sec. IV. The proposed method is verified qualitatively and quantitatively in Sec. V. Finally, we conclude the paper in Sec. VI.

II. Probabilistic Depth Prediction Task

A. Depth Regression with Uncertainty Using Bayesian Neural Network

There are various streams to predict a depth image from a single RGB image [4]–[7], [17], [20]–[24]. In this work, we use an end-to-end model that directly predicts a depth image from a single RGB image using deep neural network. Also, we use a BNN to capture uncertainty of depth prediction similar to [17]. The uncertainty captured by the model is utilized in the next section for probabilistic TSDF fusion.

Assuming RGB images $I = \{I(\mathbf{x}_1), ..., I(\mathbf{x}_N)\}$ and paired depth images $D = \{D(\mathbf{x}_1), ..., D(\mathbf{x}_N)\}$ with pixel $\mathbf{x}$ are given, the likelihood of depth prediction model can be set to follow Gaussian distribution: $p(D|\mathbf{f}^{\mathbf{w}}(I)) \sim \mathcal{N}(\mathbf{f}^{\mathbf{w}}(I), \sigma^2)$, where σ^2 is the noise inherent in the output of the model and corresponds to aleatoric uncertainty. The model can also capture epistemic uncertainty by placing a prior distribution to the parameters $\mathbf{w}$ of the model - which is reducible by gathering more data. However, posterior $p(\mathbf{W}|\mathbf{I}, \mathbf{D}) = p(\mathbf{D}|\mathbf{I}, \mathbf{W})p(\mathbf{W})/p(\mathbf{D}|\mathbf{I})$ can not be obtained in the analytic form due to the evidence, i.e., marginal likelihood $p(\mathbf{D}|\mathbf{I})$. To approximate posterior, dropout variational inference in BNN is adopted [15]. The loss function for training the prediction model capturing both uncertainty is as follows:

$$\mathcal{L}_{BNN}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\|D(\mathbf{x}_i) - \hat{D}(\mathbf{x}_i)\|^2}{2\hat{\sigma}(\mathbf{x}_i)^2} + \frac{1}{2}\log\hat{\sigma}(\mathbf{x}_i)^2 + \frac{1-p}{N}\|\hat{\mathbf{w}}\|^2, \quad (1)$$

where i represents pixel index, p is dropout probability, and $\hat{\mathbf{w}}$ is randomly masked model parameters by dropout sampling from the approximate posterior. Model's outputs $\hat{D}(\mathbf{x})$ and $\hat{\sigma}(\mathbf{x})^2$ are obtained accordingly from $f^{\hat{\mathbf{w}}}(I)$ defined by sampled weights. In (1), the first component is residual between the target and the prediction. The residual loss is attenuated by the aleatoric uncertainty of the denominator for the erroneous labels with high uncertainty. The second regularized component prevents this uncertainty from growing infinitely. The third component is the Kullback-Leibler (KL) divergence for a variational inference to approximate the true posterior using dropout layers.

By training the model according to the loss function (1), the model can output different results of the depth prediction $\hat{D}(\mathbf{x}_i)$ and the aleatoric uncertainty $\hat{\sigma}(\mathbf{x}_i)^2$ from the same input $\mathbf{x}_i$ by the sampled weights $\hat{\mathbf{w}}$. Therefore, mean and variance of the model's output by the randomness of the model parameters are obtained as follows:

$$\mathbb{E}[D(\mathbf{x})] = \frac{1}{T}\sum_{t=1}^{T}\hat{D}_t(\mathbf{x}), \quad (2)$$

$$\mathbb{V}[D(\mathbf{x})] = \underbrace{\mathbb{E}[D(\mathbf{x})^2] - (\mathbb{E}[D(\mathbf{x})])^2}_{epistemic} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_t(\mathbf{x})^2}_{aleatoric}, \quad (3)$$

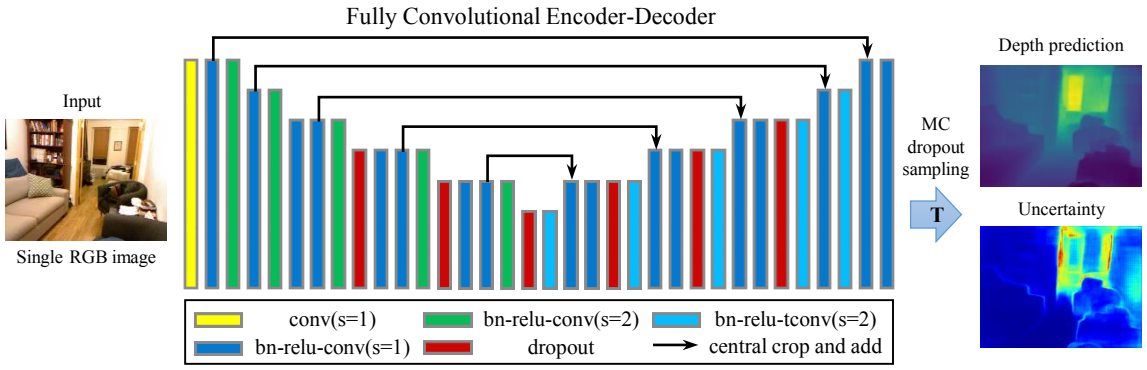

Fig. 2. **Depth and uncertainty prediction network structure.** Depth and uncertainty is predicted from a single RGB image using simple U-net based Bayesian neural network.

where $\hat{D}_t(\mathbf{x})$ and $\hat{\sigma}_t(\mathbf{x})^2$ represent t-th sampled outputs for the randomly selected weights $\hat{\mathbf{w}}_t$, and T is the total number of MC samplings in a single phase. (2) is the predictive mean of sampled outputs and (3) is the predictive variance composed of epistemic uncertainty and aleatoric uncertainty.

B. Depth Prediction Network Architecture

The proposed BNN, which is named a Bayesian U-net, has a 2D encoder-decoder structure and predicts depth from a single RGB image as an end-to-end model. The architecture of the proposed network is shown in Fig. 2. Depth regression problem resembles semantic segmentation problem because both problems require dense prediction per pixel. Therefore, the network is composed of fully convolutional layers proposed by [25] - which is mainly used in semantic segmentation. A fully convolutional network is not limited by the size of the input, and it also efficiently extracts the location information while using fewer parameters. To maintain deep structure of neural network and to transmit the detail information of the low level features to the decoder, the skip layers are connected from the encoder to the decoder like U-net [26]. Also, summation-based skip connection is used as one of the methods to overcome vanishing gradient problem which is one of U-net limitation [27]. To adjust the size of features, strides of convolutional layers is set to 2 at both downsampling and upsampling path, and center is cropped to match the size of features at skip connection. Dropout places on only half of encoder and decoder to prevent the strong regularization to the model parameters as suggested by [28]. Since MC sampling by dropout produces different T outputs for one input, we obtain the final predictions of depth and uncertainty using mean and variance of T model's outputs.

III. Probabilistic TSDF Fusion Using Uncertainty Propagation

The predicted mean d_k and the predictive variance σ_k^2 from (2) and (3) are obtained by passing the k-th input image I_k through the trained BNN. The output depth for the k-th image is assumed to follow a Gaussian distribution with the predictive mean and the predictive variance as parameters: $D_k \sim \mathcal{N}(d_k, \sigma_k^2)$. We use probabilistic TSDF fusion to reduce the noise of depth due to the uncertain prediction and to perform dense reconstruction from sequentially estimated depth and variance.

Given d_k and camera intrinsic parameters of training data, projective TSDF (or SDF) $\phi_k(\mathbf{v})$ with a voxel center $\mathbf{v} := (v_x, v_y, v_z)$ in the camera coordinate is represented as follows:

$$\phi_k(\mathbf{v}) = D_k(\mathbf{x}) - v_z, \tag{4}$$

$$\mathbf{x} = [\pi(\mathbf{v})], \tag{5}$$

where $\phi_k(\mathbf{v}) \sim \mathcal{N}(\bar{\phi}_k, \sigma_k^2)$ is a random variable of SDF for $\mathbf{v}$, $\pi(\mathbf{v})$ is the perspective projection function for $\mathbf{v}$, and $[.]$ is the nearest neighbour look up in the pixel coordinate. When the depth value follows Gaussian distribution, the TSDF, which is defined as the difference between the z-value for voxel center and the depth value, also follows the Gaussian distribution. Given observations $\phi_{1:k}(\mathbf{v})$ of k TSDF values, posterior $p(\Phi_k(\mathbf{v})|\phi_{1:k}(\mathbf{v}))$ in the global TSDF at time k is defined by Bayesian incremental learning rule as follows:

$$p(\Phi_k|\phi_{1:k}) \propto p(\phi_k|\Phi_k)p(\Phi_{k-1}|\phi_{1:k-1}). \tag{6}$$

In (6), the first component behaves like a likelihood, the second component acts like a prior distribution that constrains the posterior. In the first component, the random variable for the TSDF at time k is not affected by the global TSDF value according to the network model, thus it can be rewritten simply as $p(\phi_k(\mathbf{v}))$. Therefore, the probability distributions for mean $\bar{\Phi}_k$ and variance ν_k^2 maximizing the posterior - maximum a posteriori (MAP) inference - are arranged as:

$$\begin{aligned} p(\Phi_k(\mathbf{v})|\phi_{1:k}(\mathbf{v})) &= C \cdot \; p(\phi_k(\mathbf{v}))p(\Phi_{k-1}(\mathbf{v})|\phi_{1:k-1}(\mathbf{v})) \\ &= C \cdot \mathcal{N}(\bar{\phi}_k, \sigma_k^2) \cdot \mathcal{N}(\bar{\Phi}_{k-1}, \nu_{k-1}^2) \\ &= \mathcal{N}\left(\frac{\bar{\Phi}_{k-1}/\nu_{k-1}^2 + \bar{\phi}_k/\sigma_k^2}{1/\nu_{k-1}^2 + 1/\sigma_k^2}, \frac{1}{\nu_{k-1}^2} + \frac{1}{\sigma_k^2}\right) \\ &= \mathcal{N}(\bar{\Phi}_k, \nu_k^2). \end{aligned} \tag{7}$$

To summarize, the mean and the variance of global TSDF at time k are represented as follows using mean and variance

at $k-1$ and an observation at k:

$$\bar{\Phi}_k = \frac{\bar{\Phi}_{k-1}/\nu_{k-1}^2 + \bar{\phi}_k/\sigma_k^2}{1/\nu_{k-1}^2 + 1/\sigma_k^2}, \tag{8}$$

$$\frac{1}{\nu_k^2} = \frac{1}{\nu_{k-1}^2} + \frac{1}{\sigma_k^2}. \tag{9}$$

(8) and (9) are similar to KinectFusion [29], which updates the TSDF value using weighted running average in the original online-update manner. Whereas KinectFusion takes the weighted running average by simply letting all weights to 1, the proposed method calculates the average of new TSDF value by weighting the inverse of variance; thus, the more reliable global TSDF value can be obtained analytically. Meanwhile, both (8) and (9) yield similar results to those of [12], but in the proposed method, the likelihood $p(\phi_k(\mathbf{v}))$ is modeled not considering the sensor noise but using the uncertainty of depth prediction. Therefore, MAP inference is possible by maximizing the posterior of the global TSDF without a noise model of depth camera. Initial values of mean and variance in the global TSDF are set to $\bar{\Phi}_0 = 1$ and $\nu_0 \to \infty$ to represent free space with no information.

IV. IMPLEMENTATION DETAILS

We train our Bayesian U-net on a single NVIDIA Geforce GTX 1080Ti with 11GB of GPU memory. We use a batch size of 16 for a training and use Adam optimizer [30] with momentum parameters $\beta_1 = 0.9, \beta_2 = 0.999$. We also use the same size of MC sampling as the batch size to capture epistemic uncertainty in a single phase. The learning rate is 10^{-4} for 50 epochs, and is reduced to 10^{-5} for 10 epochs for fine-tuning. Input image and output image are resized to half size (240×320), center-cropped to 228×304 pixels for 50 epochs, and maintained in half size for 10 epochs in fine-tuning. In the experiments, the outputs of depth and uncertainty prediction are up-sampled to 480×640 of the original resolution using bilinear interpolation.

We use NYU-Depth-v2 dataset [18] to train and test our network. NYU-Depth-v2 dataset consists of 464 scenes captured with Microsoft Kinect RGB-D camera. The official train/test split consists of 249 training scenes and 215 test scenes in NYU-Depth-v2 dataset. For a training, a total of 100K RGB and depth image pairs are used from the training scenes and augmented on the fly like [10]. We sample evenly-spaced synchronized image pairs out of all training sequences. Depth image is rectified to RGB image and filled with cross bilateral filter. Also, ICL-NUIM synthetic dataset [19] is used to evaluate the performance of probabilistic TSDF fusion for 3D reconstruction. In this work, to focus on the performance of probabilistic TSDF fusion using depth and uncertainty prediction, ground truth pose is used in the tracking step. The result of reconstruction is evaluated by comparing 3D surface model of ICL-NUIM dataset.

V. EXPERIMENTAL RESULTS

In the experiments, we evaluate our method in two ways. We first evaluate our depth prediction results by comparing with other depth prediction methods. Next, we evaluate the performance of 3D reconstruction by comparing it with conventional TSDF fusion methods using 3D CAD model as the ground truth.

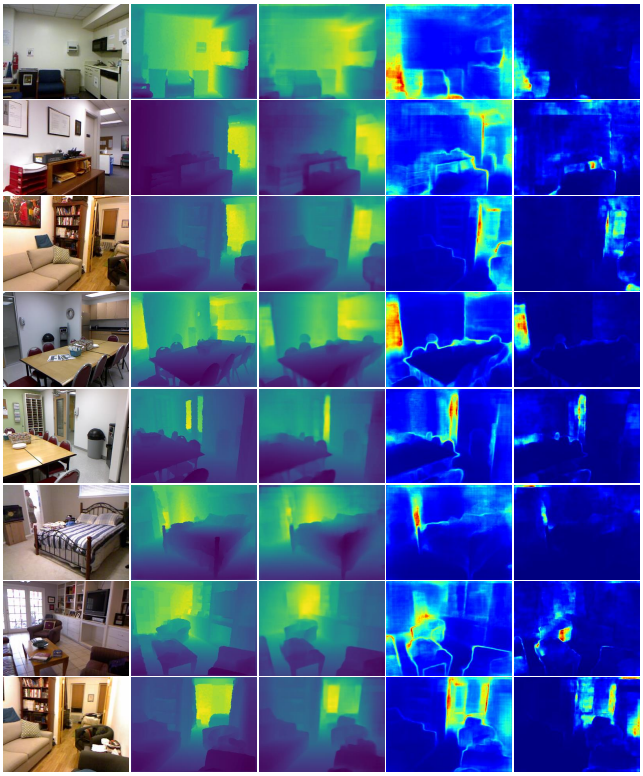

Fig. 3. **NYU-Depth-v2 prediction results.** From left to right: input RGB image, ground truth depth, depth prediction, aleatoric uncertainty, epistemic uncertainty.

A. Evaluation of Depth Prediction with Uncertainty

To evaluate the performance of the proposed method qualitatively and quantitatively, we used 654 official test images in the NYU-Depth-v2 dataset. Error and accuracy metrics for depth prediction evaluation are following previous standards [4], [5].

The results of depth prediction from an RGB image are shown in Fig. 3. As shown in Fig. 3, aleatoric and epistemic uncertainty as well as depth prediction can be obtained from a single RGB image using the proposed BNN. These results show that aleatoric uncertainty includes irreducible implicit noise for the model output, thus reflects the sensor noise model [28]; the sensor noise is greater at a larger distance and occlusion boundaries of the object. Also, aleatoric uncertainty increases on the surface of an object composed of reflective or transmissive materials, such as a mirror or glass. For instance, aleatoric uncertainty increases in the case of mirror, because it is ambiguous to predict depth given only the RGB image. To sum up, the target data for network training is gathered from a structured-light depth camera which reflects the noise model of that sensor in aleatoric uncertainty.

Meanwhile, the epistemic uncertainty can be reduced as the model gathers more data. Therefore, epistemic uncer-

TABLE I

COMPARISON OF DEPTH PREDICTION RESULTS ON NYU DEPTH V2

Methods	BNN	post-processing	Error metrics↓			Accuracy metrics↑		
			rel	rms	$\log_{10}$	δ_1	δ_2	δ_3
Eigen *et al.* [6]	✗	✗	0.215	0.907	-	61.1%	88.7%	97.1%
Li *et al.* [20]	✗	✓	0.232	0.821	0.094	62.1%	88.6%	96.8%
Liu *et al.* [21]	✗	✗	0.230	0.824	0.095	61.4%	88.3%	97.1%
Wang *et al.* [22]	✗	✓	0.220	0.745	0.094	60.5%	89.0%	97.0%
Eigen and Fergus [23]	✗	✗	0.158	0.641	-	76.9%	95.0%	98.8%
Roy and Todorovic [24]	✗	✓	0.187	0.744	0.078	-	-	-
Laina *et al.* [7]	✗	✗	0.127	0.573	0.055	81.1%	95.3%	98.8%
Kendall and Gal [17]	✓†	✗	0.110	0.506	0.045	81.7%	95.9%	98.9%
Ours	✓	✗	0.158	0.554	0.067	77.1%	93.9%	98.4%

†:MC sampling needed(It is hard to be parallelized due to GPU memory constraints)

tainty shows the limit of depth prediction that the model can estimate using the gathered data. For example, larger uncertainty is observed for humans and mirrors because it is rarely seen in the training data. Also, we can observe that epistemic uncertainty is larger for ambiguous parts where the brightness is too dark or too bright since several models give different answers for the same observation.

The comparison of the quantitative results is reported in Table I. In Table I, the proposed method, which is an end-to-end model that directly predicts the depth from a monocular image without post-processing, shows a relatively good result compared with other methods. However, the proposed method shows a lower performance than the last two end-to-end models. The model proposed by Laina *et al.* [7] shows better results, but it consumes a lot of GPU memory due to a lot of parameters (63.6×10^6), thus it is not suitable to predict the depth at every frame. Also, it cannot estimate the uncertainty directly because it is based on non-Bayesian neural network. Kendall and Gal [17] shows the best performance in terms of error and accuracy by using BNN. However, Densenet-based networks [31], [32] have problems of GPU memory constraints due to many skip connections and concatenations. Therefore, epistemic uncertainty is hardly to be captured in parallel with a single pass. Since a single prediction takes about 0.15 seconds, it is difficult to obtain epistemic uncertainty in real-time through MC sampling. On the other hand, since the proposed method is composed of lightweight network, the MC sampling can be processed in parallel, and the processing time is less than 0.1 seconds (approx. 0.09s).

B. Evaluation of 3D Reconstruction using Probabilistic TSDF Fusion

We used the ICL-NUIM dataset to compare and evaluate the proposed method using probabilistic TSDF fusion and 3D reconstruction by depth and uncertainty prediction. In particular, the ICL-NUIM dataset provides a ground truth 3D surface model to evaluate results of 3D reconstruction. Root mean square error (RMSE) metric is used to calculate the errors. To compare the results of reconstruction of probabilistic TSDF fusion, we compared it with the conventional deterministic weight options.

The comparison of the 3D reconstruction result is reported in Table II. From the row 2 to the row 4, weight varies according to SDF value. As the distance from the surface increases for the occlusion region, the weight remains constant [29] to 1 or decreases from 1 to 0 linearly [33] or exponentially [34]. For row 5 and row 6, weight changes according to the depth, not the SDF. The min-depth based weight is proposed by [11], which models the characteristics of sensor noise for larger depth. Since the weight is normalized by the smallest depth that the sensor can measure, it has the maximum weight 1 at the minimum depth. On the other hand, the minmax-depth based weight [35], [36] varies within the range interval considering not only the minimum depth but also the maximum depth that the camera can measure. Likewise, weight is normalized by 1 for the smallest depth and 0 for the largest depth. Since Kinect v1 is assumed to be a virtual RGB-D camera in the synthetic ICL-NUIM dataset, the minimum depth is set to 0.4 and the maximum depth is set to 5.0. The last two rows are evaluations of 3D reconstruction results using probabilistic fusion proposed in Sec. III. The first method uses the inverse of variance as weight and truncates it when the weight is over 1, and the second method uses the same weight, but the weight is normalized by the variance at the minimum depth considering the noise model of Kinect v1.

As reported in Table II, the proposed method (normalized uncertainty) outperforms other methods of weight option. In addition, to compare and evaluate the methods qualitatively,

TABLE II

COMPARISON OF 3D RECONSTRUCTION RESULTS ON ICL-NUIM DATASET

Weight options	kt0	kt1	kt2	kt3
Constant [29]	0.694m	1.027m	0.317m	0.988m
Linear [33]	0.711m	1.051m	0.315m	0.946m
Exponential [34]	0.701m	1.053m	0.315m	0.971m
Min-depth based [11]	**0.676m**	0.997m	0.315m	1.005m
Minmax-depth based [35], [36]	0.678m	1.022m	**0.312m**	1.013m
Truncated uncertainty	0.684m	1.027m	0.317m	0.976m
Normalized uncertainty	0.680m	**0.983m**	0.317m	**0.932m**

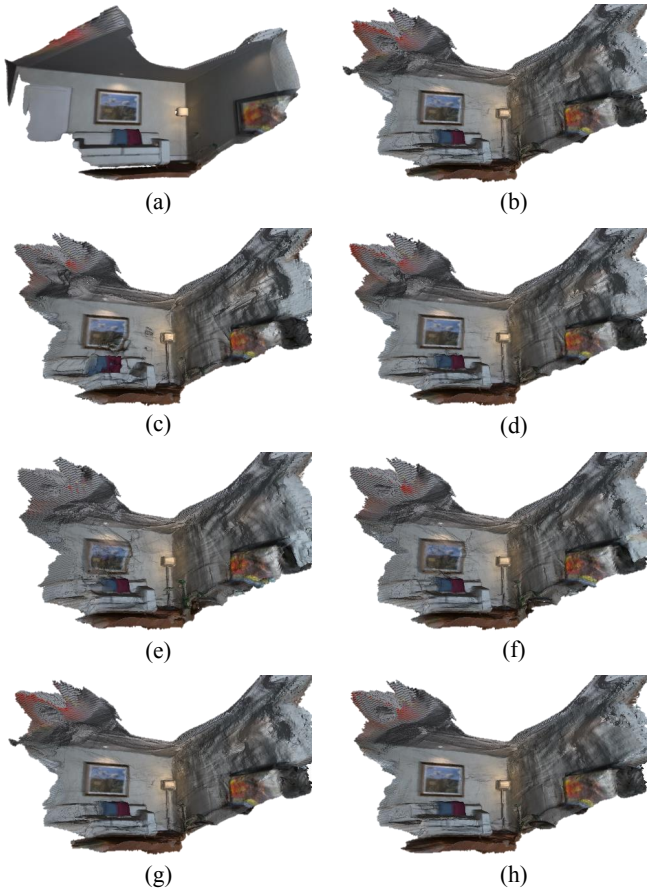

Fig. 4. **3D reconstruction results on ICL-NUIM dataset.** (*kt0* sequence). 3D meshes are extracted from the global TSDF according to the following weight options: (a) ground truth (b) constant (c) linear (d) exponential (e) min-depth based (f) minmax-depth (g) truncated uncertainty (h) normalized uncertainty

the results of 3D reconstruction are described in Fig. 4. For *kt0* sequence in ICL-NUIM dataset, 3D models of living room are reconstructed according to each weight option. As can be seen in Fig. 4, 3D reconstruction is performed using depth prediction via our proposed network. Even though some structure is missing due to the invalid depth predictions, almost scene is reconstructed well using only RGB images.

To enhance the visual comprehension, some results are enlarged in Fig. 5: ground truth, minmax-depth based weight option (the best performance on *kt0* sequence in Table II), and the proposed method using normalized uncertainty. The enlarged results show that our method outperforms other method qualitatively. Noise of depth prediction is reduced by probabilistic TSDF fusion and the surface is smoothed as can be seen in the red box in Fig. 5.

VI. CONCLUSIONS

In this paper, we proposed a dense 3D reconstruction method using RGB images in a probabilistic manner. In order to bridge the gap of the prediction models between training environment and test environment, depth was predicted with uncertainty by the proposed Bayesian neural network. We also exploited probabilistic TSDF fusion method to relax the

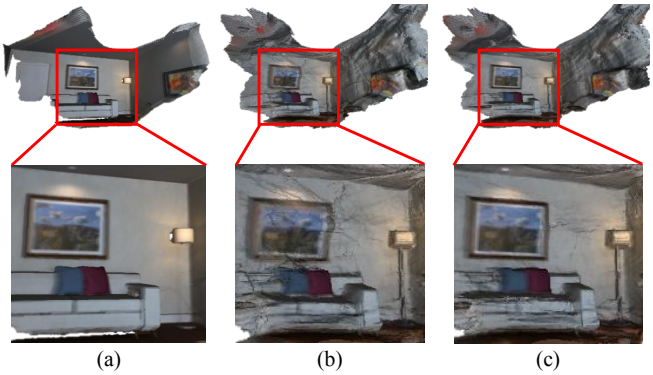

Fig. 5. **Comparison of 3D reconstruction on ICL-NUIM dataset (Enlarged).** (*kt0*). The results of 3D reconstruction are enlarged for visual comprehension. (a) ground truth (b) min-depth based (c) normalized uncertainty

accumulation noise of the depth predictions by the weighted running average of TSDF values. Probabilistic TSDF fusion was performed using the mean and the variance of depth prediction obtained from the Bayesian neural network. Through probabilistic TSDF fusion using uncertainty propagation, dense 3D reconstruction can be performed robustly against the erroneous depth prediction. In the experiment results using two official datasets, we confirmed two outcomes; the proposed neural network shows high depth prediction accuracy in real time even with low GPU memory consumption and probabilistic TSDF fusion using estimated uncertainty shows high performance in 3D reconstruction. In future works, other distribution models could be compared for probabilistic TSDF fusion using BNN, and the most suitable model would be adopted for a more robust 3D reconstruction.

REFERENCES

[1] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[2] M. O. Aqel, M. H. Marhaban, M. I. Saripan, and N. B. Ismail, "Review of visual odometry: types, approaches, challenges, and applications," *SpringerPlus*, vol. 5, no. 1, p. 1897, 2016.

[3] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: A survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, 2017.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] K. Tateno, F. Tombari, I. Laina, and N. Navab, "Cnn-slam: Real-time dense monocular slam with learned depth prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017.

[9] J. Tang, J. Folkesson, and P. Jensfelt, "Sparse2dense: From direct sparse odometry to dense 3-d reconstruction," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 530–537, 2019.

[10] F. Mal 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.

[11] C. V. Nguyen, S. Izadi, and D. Lovell, "Modeling kinect sensor noise for improved 3d reconstruction and tracking," in *2012 second international conference on 3D imaging, modeling, processing, visualization & transmission*. IEEE, 2012, pp. 524–530.

[12] V. Dietrich, D. Chen, K. M. Wurm, G. v. Wichert, and P. Ennen, "Probabilistic multi-sensor fusion based on signed distance functions," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1873–1878.

[13] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural Safety*, vol. 31, no. 2, pp. 105–112, 2009.

[14] C. M. Bishop, "Mixture density networks," Citeseer, Tech. Rep., 1994.

[15] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[16] ——, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," in *4th International Conference on Learning Representations (ICLR) workshop track*, 2016.

[17] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[18] 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.

[19] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in *2014 IEEE international conference on Robotics and automation (ICRA)*. IEEE, 2014, pp. 1524–1531.

[20] 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 computer vision and pattern recognition*, 2015, pp. 1119–1127.

[21] 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 Computer Vision and Pattern Recognition*, 2015, pp. 5162–5170.

[22] 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 Computer Vision and Pattern Recognition*, 2015, pp. 2800–2809.

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

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

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[27] T. M. Quan, D. G. Hildebrand, and W.-K. Jeong, "Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics," *arXiv preprint arXiv:1612.05360*, 2016.

[28] A. Kendall, V. Badrinarayanan, , and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[29] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*. IEEE, 2011, pp. 127–136.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[32] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19.

[33] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 303–312.

[34] E. Bylow, J. Sturm, C. Kerl, F. Kahl, and D. Cremers, "Real-time camera tracking and 3d reconstruction using signed distance functions." in *Robotics: Science and Systems*, vol. 2, 2013.

[35] H. J. Hemmat, E. Bondarev, *et al.*, "Exploring distance-aware weighting strategies for accurate reconstruction of voxel-based 3d synthetic models," in *International Conference on Multimedia Modeling*. Springer, 2014, pp. 412–423.

[36] H. J. Hemmat, E. Bondarev, G. Dubbelman, and P. H. de With, "Improved icp-based pose estimation by distance-aware 3d mapping," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 3. IEEE, 2014, pp. 360–367.