Variational Depth Networks: Uncertainty-Aware Monocular Self-supervised Depth Estimation

Georgi Dikov(✉) and Joris van Vugt

Qualcomm Technologies Netherlands B.V., Nijmegen, The Netherlands
gdikov@qti.qualcomm.com

Abstract. Using self-supervised learning, neural networks are trained to predict depth from a single image without requiring ground-truth annotations. However, they are susceptible to input ambiguities and it is therefore important to express the corresponding depth uncertainty. While there are a few truly monocular and self-supervised methods modelling uncertainty, none correlates well with errors in depth. To this end we present Variational Depth Networks (VDN): a probabilistic extension of the established monocular depth estimation framework, MonoDepth2, in which we leverage variational inference to learn a parametric, continuous distribution over depth, whose variance is interpreted as uncertainty. The utility of the obtained uncertainty is then assessed quantitatively in a 3D reconstruction task, using the ScanNet dataset, showing that the accuracy of the reconstructed 3D meshes highly correlates with the precision of the predicted distribution. Finally, we benchmark our results using 2D depth evaluation metrics on the KITTI dataset.

Keywords: Self-supervised learning · Depth estimation · Variational inference

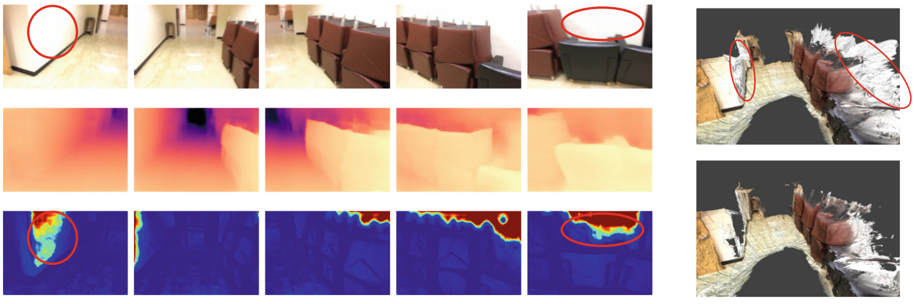
1 Introduction

Depth estimation is an important task in computer vision, since it forms the basis of many algorithms in applications such as 3D scene reconstruction [2,38,39,47,55] or autonomous driving [52,57,60] among others. Inferring depth from a single image is an inherently ill-posed problem due to a scale ambiguity: an object in an image will appear the same if it were twice as large and placed twice as far away [20]. Nevertheless, deep neural networks are able to provide reliable, dense depth estimates by learning relative object sizes from data [10]. To this end, there are two main learning paradigms: *supervised* training from

J. van Vugt—Work done while at Qualcomm Technologies Netherlands B.V.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-25085-9_3.

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13808, pp. 43–60, 2023.
https://doi.org/10.1007/978-3-031-25085-9_3



(a)

(b)

Fig. 1. (a) Top to bottom: input images from ScanNet [7] (scene 0019_00), predicted depths (far ▬ close), predicted uncertainties (low ▬ high); (b) Top: reconstructed scene from all dense depth predictions, bottom: reconstructed scene from filtered depths. Notice how the predicted uncertainty highlights regions (circled in red) for which the network would not have received meaningful error signal during self-supervised training, and is therefore susceptible to mistakes. Those can then be filtered using thresholded uncertainties as masks, leading to a sparser but more accurate scene reconstruction (Color figure online)

dense [12,49,51] or sparse [34] ground truth depth maps, e. g.obtained by a Time-of-Flight [22,43] sensor such as LiDAR [6], and *self-supervised* training which exploits 3D geometric constraints to construct an auxiliary task of photometric consistency between different views of the same scene [13,16,17,62]. The latter approach is particularly useful as it does not require ground truth depth images, and can be applied on sequences of frames taken by an ordinary, off-the-shelf monocular camera.

To reliably make use of the estimated depth in downstream tasks, a dense quantification of the uncertainty associated to the predictions is essential [28]. Consider the example given in Fig. 1 where the depth predictions for the overexposed, blank white walls are compromised (see red markings in Fig. 1a), leading to a noisy scene reconstruction as shown in the top part of Fig. 1b. To mitigate this, one can use the uncertainty maps to filter the potentially erroneous depth pixels and produce a sparser but more accurate mesh, cf.bottom of Fig. 1b. However, obtaining meaningful confidence values from a single image in a fully self-supervised learning setting is an especially challenging task, as the depth is only indirectly learnt. Consequently the majority of existing uncertainty-aware methods are either trained in a supervised fashion [4,31,35,50], assume that multiple views are available at test time [26] or model other types of uncertainty, e. g.on the photometric error [58].

The goal of our work is to extend self-supervised depth training with principled uncertainty estimation. To that end, we present Variational Depth Networks (VDN)—an entirely monocular, probabilistically motivated approach to depth uncertainty. It builds upon established self-supervised methods and leverages advancements in approximate Bayesian learning. Specifically, VDN extends MonoDepth2 [17] to model the depth as a continuous distribution, whose parameters are optimised using the framework of variational inference [18,30]. As a result, the network learns to assign high uncertainty to regions for which the depth can vary a lot without significantly increasing the photometric error, and low uncertainty otherwise. Building up on this idea, in Sect. 4 we also present a new method to quantitatively evaluate the utility of the uncertainty maps in a 3D reconstruction task using the ScanNet dataset [7], and benchmark the quality of the 2D depth predictions on the KITTI dataset [14]. In summary, our main contributions are as follows:

- We propose VDN as a novel probabilistic framework for monocular, self-supervised depth estimation, which uses approximate Bayesian inference to learn a continuous, parametric distribution over the depth. The uncertainty is then expressed as the variance of this distribution.
- We show qualitatively that the obtained uncertainty is more interpretable as it highlights regions in the image which are difficult to learn in a self-supervised setting.
- We also demonstrate that high confidence predictions are more likely to be accurate. For that, we propose an evaluation scheme based on the task of 3D scene reconstruction, where the depth uncertainty is used to filter unreliable predictions before fusion.

2 Related Work

Self-supervised Uncertainty. Self-supervised learning for monocular depth estimation was originally proposed by Zhou et al. [62]. Their core idea is that a network that predicts the depth and relative pose of a video frame can be optimised by using the photometric consistency with warped neighbouring frames as a loss function. They also include an *explainability mask* in their network to account for moving objects and non-Lambertian surfaces, which can be interpreted as a form of uncertainty estimation. Later, Godard et al. [17] consolidated several improvements into a conceptually simple method called MonoDepth2, which did not include the explainability mask since it did not have a significant impact on the accuracy of the estimated depth in practice.

Klodt and Vedaldi [31] were the first to probabilistically model the depth, pose and photometric error and use the estimated uncertainties to down-weight regions in the image that violate the colour constancy assumption made by the photometric objective function. The depth and poses are modelled through Laplacian distributions where the likelihoods of target depth and pose, obtained from a classical *Structure-from-Motion* system [36], are maximised. In contrast to their method, ours is self-contained, i. e.it does not rely on external teachers and therefore its performance is not bounded by the quality of those. In an analogous way, Yang et al. [58] also model the photometric error as a Laplacian distribution, and show that its variance can be used to improve the downstream task of visual odometry [59].

Alternatively, depth estimation can be reframed as a discrete classification problem, as proposed by Johnston and Carneiro [24], which allows for computing the variance without any additional prediction head in the network. However, their approach does not have strong guarantees on the quality of the output distribution [19] and in practice the variance appears to mostly inversely correlate with the predicted disparity except for the furthest regions in the image. On the other hand, Poggi et al. [42] present a comprehensive summary of various depth uncertainty estimation techniques for self-supervised learning and propose a combination of ensembling and self-teaching methods as an effective way to improve depth accuracy. They also propose evaluation metrics based on sparsification, which can be used to assess the quality of the predicted uncertainty. In our work we will compare to baselines from both [24] and [42].

Last, the shortcomings of photometric uncertainty estimation in the context of Multi-View Stereo [44] are addressed by Xu et al. [56] with the goal of directly improving the predicted depth. In contrast, we aim for a monocular method with interpretable uncertainty values.

Supervised Uncertainty. A fully supervised probabilistic approach is taken by Liu et al. [35], where the authors update a discrete depth probability volume (DPV) for each image, by fusing information from consecutive frames in an iterative Bayesian filtering fashion. Due to the discrete nature of the DPV, arbitrary distributions can be expressed, however to obtain an initial estimate for it, one needs to compute a cost volume from a number of frames in a video sequence. Moreover, their confidence maps show banding artefacts originating from the discrete depth representation in the cost volume.

Whereas most prior work uses a Laplacian or Gaussian distribution to model the depth and its uncertainty, ProbDepthNet from Brickwedde et al. [4] uses a Gaussian mixture model (GMM). The main benefit of GMMs is that they can represent multi-modal distributions, which can occur in foreground-background ambiguity. Walz et al. [50] propose a method for depth estimation on gated images and model the aleatoric depth uncertainty. Ke et al. [26] have the goal to improve scene reconstruction using depth uncertainty in a two-stage method: (i) predict a rough depth and uncertainty estimates using optical flow and triangulation from multiple frames; and (ii) refine the outputs of the first stage in an iterative procedure based on recurrent neural networks.

3 Methods

3.1 Background and Motivation

Fundamentals. Let $\mathcal{D} = \{I_t\}_{t=1 \ldots N}$ be a sequence of image frames and $T_{t \to s}$ the corresponding 3D camera motion from a target frame t to a source frame s. Further, let K denote the camera intrinsic matrix, projecting from 3D camera coordinates to 2D pixel coordinates $x \in \mathcal{X}$. Then, by exploiting 3D geometric constraints, one can cast the task of learning a depth map D_t for a frame I_t as a photometric consistency optimisation problem between the target and the warped source frames [13,17,62]:

$$\mathcal{L}_{\text{photo}}(I_t, D_t) = \sum_{x \in \mathcal{X}} \| I_s \langle K T_{t \to s} D_t(x) K^{-1} x \rangle - I_t(x) \|, \tag{1}$$

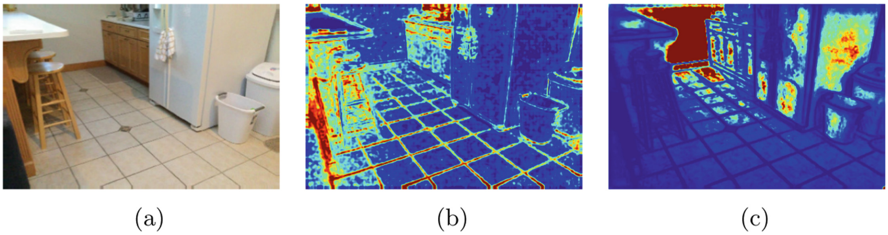


Fig. 2. (a) A sample input image from ScanNet [7] (scene 0000_00); (b) Photometric uncertainty; (c) Variational depth uncertainty; (low ▬ high)

where $I_s\langle\cdot\rangle$ stands for a (bilinear) interpolation on the source frame I_s, following the notation of [17]. For the sake of notational brevity, here and throughout the rest of the paper we omit the dependencies on K as well as on I_s and $T_{t\rightarrow s}$ in the losses.

The estimated depth D_t is usually expressed as the inverse disparity output of a deterministic convolutional neural network μ_θ, parametrised by weights θ:

$$D_t = \mu_\theta(I_t)^{-1}. \tag{2}$$

For numerical reasons, the disparity output is activated by a sigmoid non-linearity and stretched to a predefined $\left[d_{\max}^{-1}, d_{\min}^{-1}\right]$ range. In practice, the loss from Eq. (1) is also extended to account for multiple source frames (e. g.using the *minimum reprojection error* [17]) and combined with other terms such as structural similarity [53] or smoothness regularisation [16,17]. In this paper we will refer to the extended loss as $\mathcal{L}_{\text{photo}}$ and to the full model as MonoDepth2. Importantly, this will serve us as a basis framework for monocular, self-supervised depth learning upon which we will introduce a probabilistic extension in Sect. 3.2.

Uncertainty Estimation. Despite its wide-spread popularity, MonoDepth2 is not designed to account for the uncertainty associated with D_t. Following a paradigm of modelling the aleatoric uncertainty explicitly [28] one can reframe the loss from Eq. (1) into an exponential family likelihood with a learnable variance $\hat{\sigma}_\theta$:

$$p(I_t \mid D_t) \propto \frac{1}{\hat{\sigma}_\theta(I_t)} \exp\left(-\frac{\mathcal{L}_{\text{photo}}(I_t, D_t)}{\hat{\sigma}_\theta(I_t)}\right), \tag{3}$$

where we abuse the notation for the weights θ and the neural network $\hat{\sigma}$, which may share only some of its parameters with μ. At this point, it is important to clarify that $\hat{\sigma}_\theta(I_t)$, as used in Eq. (3), accounts merely for the variance in the photometric error, $\mathcal{L}_{\text{photo}}$, and not the predicted depth D_t. To give and intuitive explanation why the two uncertainties are not interchangeable, consider the following thought experiment: let all pixels in I_t and I_s have the same colour value. Then, for any predicted D_t and arbitrary $T_{t\rightarrow s}$ we have that

$I_t(x) = I_s\langle KT_{t\to s}D_t(x)K^{-1}x\rangle = I_s(x), \forall x \in \mathcal{X}$ and the likelihood from Eq. (3) is maximised with $\hat{\sigma}_\theta(I_t) \to 0$. Thus the photometric variance will collapse, while the actual depth variance is large.

In reality, this scenario can occur at large textureless surfaces, such as walls or overexposed regions close to light sources. Figures 2a and 2b show an example input and the corresponding photometric uncertainty. Notice how the network confidence is lowest in the aforementioned regions and highest on their boundaries or in high-frequency patterned areas, where small changes in D_t can substantially increase $\mathcal{L}_{\text{photo}}$. Thus, the photometric variance does not necessarily correlate with the uncertainty in the depth estimate, and in some cases it is even complementary to the latter. On the other hand, VDN is able to assign high depth variance to those regions, cf. Fig. 2c.

Despite that, the photometric uncertainty has been reported to quantitatively improve the depth estimates [42,58]. We hypothesise that this can be attributed to the effect of loss attenuation as the supervisory signal is not dominated by noise stemming from the difficult, depth sensitive areas such as non-Lambertian objects, similarly to the observations made by [28] in a supervised depth regression setup. Nevertheless, there are real-world applications, such as 3D scene reconstruction where proper depth uncertainty estimation is of greater importance, as we will show experimentally in Sect. 4.

3.2 Variational Depth Networks

Objective. In the following, we will introduce a probabilistic extension to the self-supervised depth learning pipeline, in which the variance of the predicted depth maps can be reliably estimated. Intuitively speaking, we will assume that D_t is a random variable following some conditional distribution and we will make the image warping transformation in Eq. (1) aware of the probabilistic nature of D_t. We find this intuition to fit well into the Bayesian framework of reasoning and we will leverage approximate variational inference [18,25,30] to optimise a parametric distribution over D_t.

In essence, it requires that we specify a likelihood $p(I_t \mid D_t)$, a prior $p(D_t)$ and a posterior distribution $p(D_t \mid I_t)$ to which a tractable approximation, $q_\theta(D_t \mid I_t)$ is fit. Then, using q_θ we can derive a lower bound on the marginal log-likelihood:

$$\begin{aligned}\mathbb{E}_{p_\mathcal{D}}[\log p_\theta(I_t)] &= \mathbb{E}_{p_\mathcal{D}}\left[\log \mathbb{E}_{q_\theta}\left[\frac{p(I_t \mid D_t)p(D_t)}{q_\theta(D_t \mid I_t)}\right]\right] &(4)\\ &\geq \mathbb{E}_{p_\mathcal{D},q_\theta}\left[\log \frac{p(I_t \mid D_t)p(D_t)}{q_\theta(D_t \mid I_t)}\right]. &(5)\end{aligned}$$

This can be further decomposed into a log-likelihood and a KL-divergence term, into the so-called *evidence lower bound*:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(I_t, D_t) = \mathbb{E}_{p_\mathcal{D},q_\theta}&[\log p(I_t \mid D_t)]\\ &- \mathbb{E}_{p_\mathcal{D}}[\text{KL}\left(q_\theta(D_t \mid I_t) \mid\mid p(D_t)\right)].\end{aligned} \tag{6}$$

One can show that maximising $\mathcal{L}_{\text{ELBO}}$ w. r. t.θ is equivalent to minimising $\mathbb{E}_{p_{\mathcal{D}}}[\text{KL}\left(q_\theta(D_t \mid I_t) \mid\mid p(D_t \mid I_t)\right)]$ thus closing the gap between the approximation and the underlying true posterior [18,25]. For the likelihood of VDN we choose an unnormalised density as in Eq. (3), however, throughout this work we will not model both the photometric and depth uncertainty, so as to isolate the effects of our contribution. In the subsequent sections we will specify the exact form of $q_\theta(D_t \mid I_t)$ and $p(D_t)$.

Approximate Posterior. In the context of depth estimation, one has to take into account two considerations when choosing a suitable family of variational distributions. First, it has to have a positive, bounded support over $[d_{\min}, d_{\max}]$ and, second, it has to allow for reparametrisation so that the weights θ can be learnt with backpropagation. One such candidate distribution is given by the truncated normal distribution [5], constrained to the aforementioned interval, whose location parameter is defined by the output of the neural network μ_θ and the scale by σ_θ, similarly to Eq. (3). Unlike the photometric variance $\hat{\sigma}_\theta$, σ_θ will have a direct relation to the variance of the estimated depth. For numerical reasons, however, it may be beneficial to express the approximate posterior over disparity instead of depth [17], and convert disparity samples to depth as per Eq. (2):

$$q_\theta(D_t^{-1} I_t) = \mathcal{N}_{\text{tr}}\left(D_t^{-1} \mid \mu_\theta(I_t), \sigma_\theta(I_t), d_{\max}^{-1}, d_{\min}^{-1}\right). \tag{7}$$

Backpropagating to μ_θ and σ_θ is possible through a reparametrisation using the inverse CDF function, which is readily implemented in TensorFlow [1,11] and in third-party packages [40] for Pytorch [41].

Here we assume that q_θ is a pixelwise factorised distribution and we obtain a disparity prediction using the mode, $\mu_\theta(I_t)$. Since we have defined a distribution over the disparity, it is not straightforward to obtain the mode of the transformed distribution over the depth, q_θ^{-1}. Fortunately however, for the given truncated normal parametrisation and the reciprocal transformation one can compute it analytically from the density of q_θ^{-1} using the change of variables trick, see Appendix A.2 for details:

$$\begin{aligned} \text{mode}\left(q_\theta^{-1}(D_t \mid I_t)\right) &= \min(\max(m, d_{\min}), d_{\max}), \\ \text{where} \quad m &= \frac{\sqrt{\mu_\theta(I_t)^2 + 8\sigma_\theta(I_t)^2} - \mu_\theta(I_t)}{4\sigma_\theta(I_t)^2}. \end{aligned} \tag{8}$$

Finally, to obtain the estimated pixelwise depth uncertainty, one can compute the sample variance of q_θ^{-1}.

Prior. The choice of depth prior is particularly important for us because it can adversely bias the shape of the variational posterior. To understand the reason for that, one has to compare the VDN model with a regular VAE [30]: while both models encode the input image in a latent representation, a VDN does not use

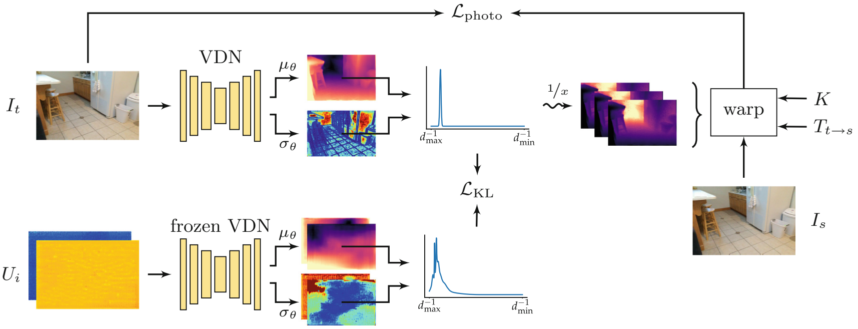


Fig. 3. Model overview of VDN with an example input from ScanNet [7] (scene 0000_00). Given a target image I_t, the subnetworks μ_θ and σ_θ predict the pixelswise location and scale parameters of the approximate posterior, resulting in a factorised distribution over disparities. Then, multiple samples are drawn and the reciprocal of each is used independently in a warping transformation of a source image I_s, assuming known intrinsics K and pose $T_{t\rightarrow s}$. The warped and interpolated source frames are used to compute the likelihood. The prior is given by the predicted location and scale parameters from a set of pseudo-inputs U_i as per [48]. The arrow $\rightsquigarrow$ denotes a sampling operation

a learnable decoder to form the likelihood but rather a fixed warping transformation. This means that a bias in the latent space, cannot be compensated for during decoding, resulting in hindered weight optimisation. For this reason we opt for a learnable prior given by the aggregated approximate posterior, which is provably the optimal prior for that task [46,48], see Appendix A.1 for details:

$$p^*(D_t) = \sum_{I_t \in \mathcal{D}} q_\theta(D_t \mid I_t) p_{\mathcal{D}}(I_t). \tag{9}$$

Unfortunately, however, the estimation of the aggregate posterior is computationally prohibitive for large, high-dimensional datasets. Therefore, we employ an approximation by Tomczak et al. [48], called *VampPrior*, where the prior is given as a mixture of the variational posteriors computed on a set of learnable pseudo inputs $\{U_i\}_{i=1\ldots k}$:

$$p(D_t) \approx \frac{1}{k} \sum_{i=1}^{k} q_\theta(D_t \mid U_i). \tag{10}$$

Earlier, we expressed the approximate posterior in disparity- rather than depth-space and consequently the prior becomes a mixture distribution over disparities too. Since the KL-divergence is invariant to continuous, invertible transformations [33] (such as the reciprocal relation of depth and disparity), one can compute KL $(q_\theta(D_t^{-1} \mid I_t) \mid\mid p(D_t^{-1}))$ instead. In summary, all of the components of VDN are presented in Fig. 3.

4 Experiments

4.1 Setup

Datasets

ScanNet. The ScanNet [7] dataset contains 1513 video sequences collected in indoor environments, annotated with 3D poses, dense depth maps and reconstructed meshes. The reason to use this dataset is to evaluate the per-image depth and uncertainty estimation and to assess the utility of uncertainty in 3D reconstruction. Consequently, we use the ground-truth poses to compute the photometric error instead of predicting them. For training we only consider every 10^{th} frame as target to reduce redundancy and for each, we find a source frame both backwards and forwards in time with a relative translation of 5–10cm and a relative rotation of at most 5°. All images are resized to 384 × 256 pixels. We use the ground-truth poses to compute the photometric error and do not train a network to predict the pose since we want to focus our analysis in this experiment on the depth and uncertainty estimates only.

KITTI. The KITTI dataset [14] is an established benchmark dataset for depth estimation research and consists of 61 sequences collected from a vehicle. Following [17], we use the Eigen split [12], resize the input images to 640 × 192 and evaluate against LiDAR ground-truth capped at 80 m. Unlike the ScanNet experiments, here the camera poses are learnt the same way as in [17] so as to allow for fair comparison.

Metrics

3D. Previous works on 3D reconstruction [3,37,45] use point-to-point distances as the basis for comparing to ground-truth meshes. They convert each mesh to a point cloud by only considering its vertices, or by sampling points on the faces, essentially discarding the surface information of the mesh. However, if a predicted point lies on the surface of the ground-truth mesh it can still incur a non-zero error since only the distance to the closest vertex is considered. To mitigate this, we propose to use a cloud-to-mesh (c → m) distance as a basis for our 3D reconstruction error computation, which is readily available in open-source software like CloudCompare [15]. Given a mesh $\mathcal{M} = (\mathcal{V}, \mathcal{F})$, where $\mathcal{V}$ denotes the vertices and $\mathcal{F}$ the faces, we compute the accuracy as the fraction of vertices for which the Euclidean distance to the closest face $f' \in \mathcal{F}'$ in another mesh $\mathcal{M}'$ is smaller than a threshold ϵ:

$$\text{acc}_{\text{c}\to\text{m}}(\mathcal{M}, \mathcal{M}') = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{1}\left[\min_{f' \in \mathcal{F}'} \text{dist}(v, f') < \epsilon\right]. \tag{11}$$

Here $\mathbb{1}[\cdot]$ denotes the indicator function. Given predicted and ground-truth meshes, $\mathcal{M}_{\text{pred}}$ and $\mathcal{M}_{\text{gt}}$ respectively, we define the *precision* as $\text{acc}_{\text{c}\to\text{m}}(\mathcal{M}_{\text{pred}},$

$\mathcal{M}_{\text{gt}}$) and the *recall* as $\text{acc}_{\text{c}\rightarrow\text{m}}(\mathcal{M}_{\text{gt}}, \mathcal{M}_{\text{pred}})$. The F-score is the harmonic mean of the precision and recall [32]. Following standard practices in 3D reconstruction literature [37,45] we use a threshold of $\epsilon = 5\,\text{cm}$ in all our evaluations.

2D. For the evaluation of the 2D predicted depth maps we compute the widely used metrics proposed by Eigen et al. [12]. Uncertainty is evaluated using sparsification curves [23] and the Area Under the Sparsification Error (AUSE) and Area Under the Random Gain (AURG) as proposed by Poggi et al. [42]. Note AURG and AUSE are computed w. r. t.another 2D depth metric and therefore comparison among different models is fair only when they perform similarly on that metric too.

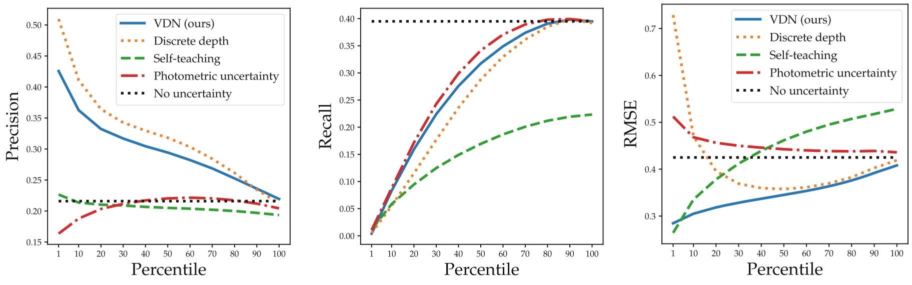
Implementation Details

Network Architectures and Training Details. Even though our model architecture strictly follows [17] there are a couple of deviations. In particular, to accommodate the prediction of the distribution location and scale parameters, we duplicate the original disparity decoder architecture and, for the scale parameter only, change the output activation to linear. To avoid numerical instability issues with the scale, we clip it to the $[10^{-6}, 3]$ interval. In all our experiments we use a ResNet-18 encoder [21], pretrained on ImageNet [9], the Adam optimiser [29] with an initial learning rate of 10^{-4} which we reduce by a factor of 10 after 30 epochs, for a total of 40 epochs. The VampPrior for our VDN models is computed as described in Sect. 3.2 with 20 pseudo-inputs, which we initialise by broadcasting a random colour value over the height and width dimensions. To estimate the loss $\mathcal{L}_{\text{ELBO}}$ from Eq. (6) the approximate posterior is sampled 10 times.

3D Reconstruction. We use the TSDF-fusion algorithm implemented in Open3D [61] to reconstruct ScanNet [7] scenes. To speed up reconstruction, we only integrate every 10^{th} frame and, during fusion, we use a sample size of 5 cm and a truncation distance of 20 cm. For evaluation we use the ground-truth meshes provided with the dataset.

4.2 ScanNet: Uncertainty-Aware Reconstruction

To evaluate the usefulness of the predicted uncertainty we use the task of 3D reconstruction on ScanNet [7] scenes. In this experiment we leverage the depth uncertainty for measurement selection by masking out pixels with uncertainty above a preselected threshold during the integration process. We compare our method to several other recently proposed depth uncertainty estimation methods, all implemented on top of the same MonoDepth2 framework. *Photometric uncertainty* refers to Eq. (3), which is used by D3VO [58] to improve visual odometry. *Self-teaching* refers to the method proposed by Poggi et al. [42], where we use the model without uncertainty as a teacher for training the student network



(a) (b) (c)

Fig. 4. ScanNet: mean reconstruction precision (a) and recall (b) as well as 2D depth RMSE (c) curves on the validation set for various filtering thresholds on the uncertainty. A monotonically decreasing precision curve indicates that the uncertainty correlates well with the errors in the depth maps used for fusion while a higher recall means that smaller portions of the geometry are being removed

in a supervised way. *Discrete depth* predicts a discrete disparity volume [24], from which continuous depth and variance can be derived. Each of these methods constitute a fair baseline as all are fully self-supervised and monocular.

Table 1. ScanNet: 2D depth, 2D uncertainty and 3D reconstruction metrics. All methods are based on the same MonoDepth2 architecture and are our own (re)implementations. ↑ and ↓ denote if higher or lower score is better

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	AUSE^a ↓	AURG^a ↑	Precision ↑	Recall ↑	F-score ↑
No uncertainty (MonoDepth2)	0.146	0.088	0.425	0.204	0.800	**0.948**	**0.985**	-	-	0.216	**0.395**	0.276
Photometric uncertainty	0.154	0.098	0.426	0.215	0.787	0.940	0.979	0.102	-0.008	0.204	**0.395**	0.266
Self-teaching [42]	0.170	0.115	0.529	0.246	0.690	0.914	0.975	**0.056**	**0.034**	0.194	0.223	0.204
Discrete depth [24]	0.147	0.086	0.419	0.202	0.796	0.946	0.984	0.091	-0.001	0.212	0.392	0.272
VDN (fixed prior)	0.148	0.093	0.416	0.211	0.797	0.944	0.981	0.083	0.006	0.217	0.392	0.276
VDN (VampPrior)	**0.144**	**0.085**	**0.402**	**0.194**	**0.801**	**0.948**	**0.985**	0.083	0.003	**0.219**	**0.395**	**0.279**
VDN (fixed prior, 10 scenes)	0.414	0.495	1.036	0.787	0.318	0.546	0.675	**0.125**	**0.094**	0.096	**0.262**	0.138
VDN (VampPrior, 10 scenes)	**0.287**	**0.238**	**0.680**	**0.348**	**0.494**	**0.797**	**0.931**	0.152	0.008	**0.103**	0.261	**0.144**

[a] Measured on Abs Rel.

Table 1 summarises the results for the standard 2D depth and 3D reconstruction metrics. First, we note that photometric uncertainty performs considerably worse than the other methods. Discrete depth performs generally on par with the No uncertainty baseline. VDN slightly outperforms all baselines on most metrics. Figure 4a shows the mean reconstruction precision when increasing the uncertainty threshold at which predictions are considered valid. We expect to see a downwards trend, as using more uncertain predictions should decrease the accuracy of the reconstructed mesh. Here, the photometric uncertainty does not show this behaviour, whereas the variational and discrete uncertainty do show it, with discrete generally having a higher precision everywhere except when using more than 90% of the pixels. Conversely, Fig. 4b shows the mean reconstruction recall where a rapid increase signifies that larger pieces of the scene geometry are being cut out. For the sake of completeness, in Fig. 4c we also a provide

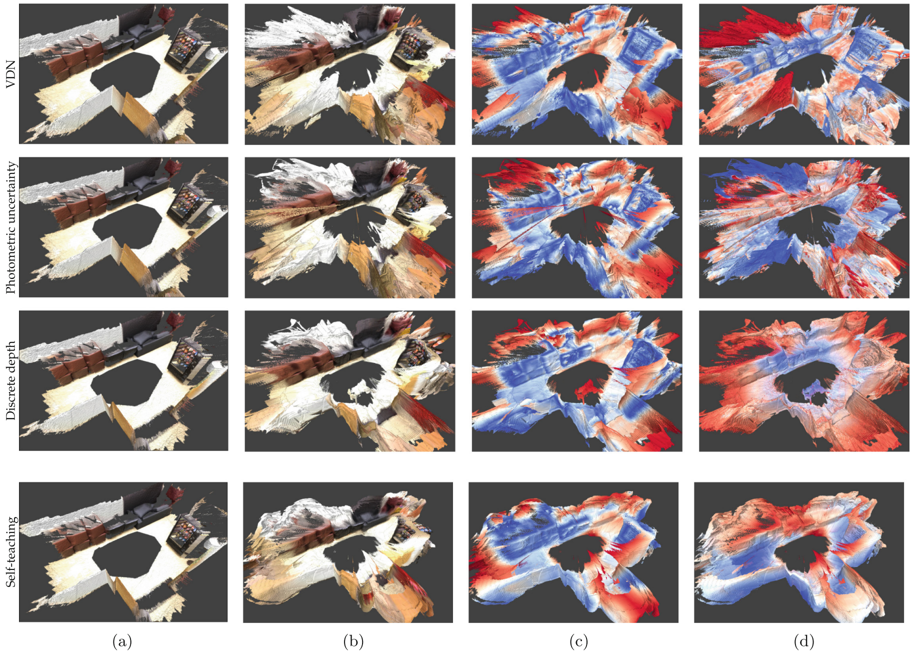


Fig. 5. (a) Meshes constructed using the ground-truth depth maps from ScanNet [7] (scene 0019_00); (b) Coloured meshes using the predicted depths; (c) Meshes from predicted depth, coloured by the cloud-to-mesh distances from the ground-truth; (d) Meshes from predicted depth, coloured by the depth uncertainty; (low ▬ high)

similar plots for the mean RMSE as measured on 2D depth images. Figures 5a and 5b show reconstructions from ground-truth and predicted depths for all uncertainty-aware baselines, and Figs. 5c and 5d depict the corresponding cloud-to-mesh distances and uncertainties. Notice how the photometric uncertainty anti-correlates with the precision, while the discrete depth merely increases the uncertainty with the distance from the camera. The output of the self-teaching model is not very interpretable either as it models the aleatoric noise in the teacher network. More qualitative examples are disclosed in Appendix B.

4.3 ScanNet: Prior Ablation Study

To investigate the adverse effects of naively specifying a prior distribution over the disparity, we compare the VampPrior against a truncated normal distribution with fixed location and scale parameters at 0.5 and 2.0 respectively, in two training scenarios: on the full training data and on a subset of 10 scenes only. The latter setup is especially interesting because it exacerbates any undesirable influence the prior might have onto the approximate posterior due to the lack of sufficient training data. While both priors are capable of regularising the spread of the variational posterior, the VampPrior shows superior results as presented

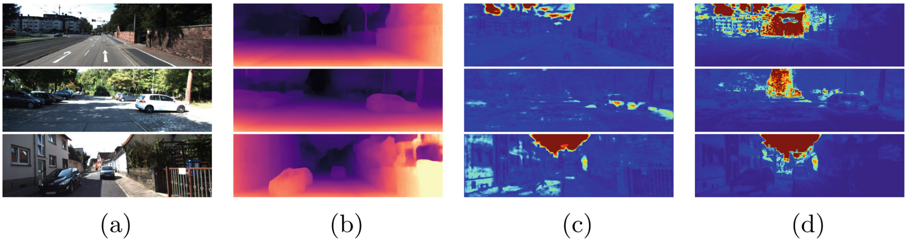


Fig. 6. (a) Sample input images from the Eigen test split [12] in KITTI [14]; (b) Predicted disparities; (c) Predicted disparity variance; (d) Estimated depth variance using 100 samples

Table 2. KITTI: 2D depth and uncertainty evaluation results on the Eigen test split [12] with raw LiDAR ground truth (80 m)

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑	Abs Rel		RMSE		$\delta < 1.25$	
								AUSE ↓	AURG ↑	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑
No uncertainty (MonoDepth2 [17])	0.115	0.882	4.791	0.190	0.879	**0.961**	0.982	-	-	-	-	-	-
Boot+Self [42][a]	**0.111**	**0.826**	**4.667**	**0.184**	**0.880**	**0.961**	**0.983**	**0.033**	**0.040**	2.124	1.857	**0.033**	**0.077**
Photometric uncertainty [42][a]	0.113	0.928	4.919	0.192	0.876	0.958	0.981	0.051	0.027	3.097	1.188	0.060	0.056
VDN (ours)	0.117	0.882	4.815	0.195	0.873	0.959	0.981	0.058	0.018	**1.942**	**2.140**	0.085	0.030

[a] The scores are taken from Tables 10 and 13 in the supplementary material of [42].

in the bottom half of Table 1. In particular, in the low data regime, it achieves significantly better scores on most metrics.

4.4 KITTI: 2D Depth Evaluation

In order to benchmark VDN on the KITTI dataset [14] against comparable prior work, we have selected as baselines the original MonoDepth2 [17], referred to as *No uncertainty*, the MonoDepth2 (*Boot+Self*) from Poggi et al. [42], which does account for depth uncertainty through self-teaching and bootstrapped ensemble learning, and the *Photometric uncertainty* baseline also presented in [42] under the name *MonoDepth2-Log*. Table 2 shows the depth and uncertainty results for the VDN and the baselines. Our model performs slightly worse than the baselines except for the RMSE-AUSE and AURG metrics, which we attribute to the increased amount of noise during training, stemming from the stochastic sampling operations. Figure 6a shows three example inputs from the test set with their corresponding disparity location and scale predicted parameters in Figs. 6b and 6c. The resulting depth uncertainty is illustrated in Fig. 6d, which highlights the depth ambiguity of the sky and distant, indistinguishable objects.

5 Conclusions

We have presented a probabilistic extension of MonoDepth2, which learns a parametric posterior distribution over depth. The method yields useful uncertainty, which correlates well with the error in the depth predictions and consequently, we have shown that one can use the uncertainty to mask out unreliable pixels and improve the precision of meshes in a 3D scene reconstruction task. Such masking, however, can come at a cost of decreased recall, resulting in sparser meshes. It is therefore a promising direction for future work to combine our method with a disparity [27] or mesh completion algorithm [8]. Other extensions of our work could combine the photometric and variational depth uncertainties, as the former is complementary to the latter, or apply VDN to multi-view, self-supervised depth estimation [54]. Finally, we note that due to the stochastic nature of our method, it is moderately demanding on computation and memory resources during training, as an additional forward-pass is needed for the VampPrior, and multiple samples are drawn from the approximate posterior to estimate the likelihood and KL-divergence terms of the loss. In addition, the depth uncertainty is computed from samples of the transformed disparity posterior. For the training and evaluation of all models we have used a single NVIDIA RTX A5000 GPU with 24 GB of memory.

Acknowledgements. We highly appreciate the constructive feedback and suggestions from our colleagues Mohsen Ghafoorian and Alex Bailo as well as the consistent support from Gerhard Reitmayr and Eduardo Esteves.

References

1. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). https://www.tensorflow.org/
2. Bloesch, M., Czarnowski, J., Clark, R., Leutenegger, S., Davison, A.J.: CodeSLAM-learning a compact, optimisable representation for dense visual slam. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2560–2568 (2018)
3. Božič, A., Palafox, P., Thies, J., Dai, A., Nießner, M.: TransformerFusion: monocular RGB scene reconstruction using transformers. arXiv preprint arXiv:2107.02191 (2021)
4. Brickwedde, F., Abraham, S., Mester, R.: Mono-SF: multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2780–2790 (2019)
5. Burkardt, J.: The truncated normal distribution. Department of Scientific Computing Website, Florida State University, pp. 1–35 (2014). https://people.sc.fsu.edu/jburkardt/presentations/truncated_normal.pdf
6. Christian, J.A., Cryan, S.: A survey of lidar technology and its use in spacecraft relative navigation. In: AIAA Guidance, Navigation, and Control (GNC) Conference, p. 4641 (2013)
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the Computer Vision and Pattern Recognition (CVPR). IEEE (2017)
8. Dai, A., Ruizhongtai Qi, C., Nießner, M.: Shape completion using 3D-encoder-predictor CNNs and shape synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5868–5877 (2017)

9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
10. Dijk, T.V., Croon, G.D.: How do neural networks see depth in single images? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2183–2191 (2019)
11. Dillon, J.V., et al.: Tensorflow distributions. arXiv preprint arXiv:1711.10604 (2017)
12. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. arXiv preprint arXiv:1406.2283 (2014)
13. Garg, R., B.G., V.K., Carneiro, G., Reid, I.: Unsupervised CNN for single view depth estimation: geometry to the rescue. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 740–756. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_45
14. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. (IJRR) **32**, 1231–1237 (2013)
15. Girardeau-Montaut, D.: Cloudcompare. France: EDF R&D Telecom ParisTech (2016). https://www.cloudcompare.org/
16. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270–279 (2017)
17. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3828–3838 (2019)
18. Graves, A.: Practical variational inference for neural networks. Adv. Neural Inf. Process. Syst. **24** (2011)
19. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
20. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004). https://doi.org/10.1017/cbo9780511811685
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
22. Horaud, R., Hansard, M., Evangelidis, G., Ménier, C.: An overview of depth cameras and range scanners based on time-of-flight technologies. Mach. Vis. Appl. **27**(7), 1005–1020 (2016). https://doi.org/10.1007/s00138-016-0784-4
23. Ilg, E., et al.: Uncertainty estimates and multi-hypotheses networks for optical flow. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 652–667 (2018)
24. Johnston, A., Carneiro, G.: Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4756–4765 (2020)
25. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Mach. Learn. **37**(2), 183–233 (1999)
26. Ke, T., Do, T., Vuong, K., Sartipi, K., Roumeliotis, S.I.: Deep multi-view depth estimation with predicted uncertainty. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 9235–9241. IEEE (2021)
27. Keltjens, B., van Dijk, T., de Croon, G.: Self-supervised monocular depth estimation of untextured indoor rotated scenes. arXiv preprint arXiv:2106.12958 (2021)

28. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977 (2017)
29. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic gradient descent. In: ICLR: International Conference on Learning Representations, pp. 1–15 (2015)
30. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014, Conference Track Proceedings (2014). http://arxiv.org/abs/1312.6114
31. Klodt, M., Vedaldi, A.: Supervising the new with the old: learning SFM from SFM. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 698–713 (2018)
32. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Trans. Graph. (ToG) **36**(4), 1–13 (2017)
33. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
34. Kuznietsov, Y., Stuckler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6647–6655 (2017)
35. Liu, C., Gu, J., Kim, K., Narasimhan, S.G., Kautz, J.: Neural RGB (R) D sensing: depth and uncertainty from a video camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10986–10995 (2019)
36. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE Trans. Rob. **31**(5), 1147–1163 (2015)
37. Murez, Z., van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: end-to-end 3D scene reconstruction from posed images. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part VII. LNCS, vol. 12352, pp. 414–431. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58571-6_25
38. Newcombe, R.A., et al.: KinectFusion: real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality, pp. 127–136. IEEE (2011)
39. Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M.: Real-time 3D reconstruction at scale using voxel hashing. ACM Trans. Graph. (ToG) **32**(6), 1–11 (2013)
40. Obukhov, A.: Truncated normal distribution in PyTorch (2020), https://github.com/toshas/torch_truncnorm
41. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
42. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3227–3237 (2020)
43. Remondino, F., Stoppa, D.: TOF Range-imaging Cameras, vol. 68121. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-27523-4
44. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 1, pp. 519–528. IEEE (2006)

45. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: NeuralRecon: real-time coherent 3D reconstruction from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15598–15607 (2021)
46. Takahashi, H., Iwata, T., Yamanaka, Y., Yamada, M., Yagi, S.: Variational autoencoder with implicit optimal priors. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5066–5073 (2019)
47. Tateno, K., Tombari, F., Laina, I., Navab, N.: CNN-SLAM: real-time dense monocular SLAM with learned depth prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6243–6252 (2017)
48. Tomczak, J., Welling, M.: VAE with a VampPrior. In: International Conference on Artificial Intelligence and Statistics, pp. 1214–1223. PMLR (2018)
49. Ummenhofer, B., et al.: DeMoN: depth and motion network for learning monocular stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5038–5047 (2017)
50. Walz, S., Gruber, T., Ritter, W., Dietmayer, K.: Uncertainty depth estimation with gated images for 3D reconstruction. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pp. 1–8. IEEE (2020)
51. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2022–2030 (2018)
52. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: bridging the gap in 3D object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8445–8453 (2019)
53. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
54. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: self-supervised multi-frame monocular depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1164–1174 (2021)
55. Whelan, T., Leutenegger, S., Salas-Moreno, R., Glocker, B., Davison, A.: ElasticFusion: dense slam without a pose graph. In: Robotics: Science and Systems (2015)
56. Xu, H., et al.: Digging into uncertainty in self-supervised multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6078–6087 (2021)
57. Xu, Y., Zhu, X., Shi, J., Zhang, G., Bao, H., Li, H.: Depth completion from sparse LiDAR data with depth-normal constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2811–2820 (2019)
58. Yang, N., Stumberg, L.v., Wang, R., Cremers, D.: D3VO: deep depth, deep pose and deep uncertainty for monocular visual odometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1281–1292 (2020)
59. Yang, N., Wang, R., Stuckler, J., Cremers, D.: Deep virtual stereo odometry: leveraging deep depth prediction for monocular direct sparse odometry. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 817–833 (2018)
60. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1983–1992 (2018)

61. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: a modern library for 3D data processing. arXiv:1801.09847 (2018)
62. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851–1858 (2017)