MonoProb: self-supervised monocular depth estimation with interpretable incertainty

Rémi Marsal, Florian Chabot, Angélique Loesch, William Grolleau, Hichem Sahbi

MonoProb: Self-Supervised Monocular Depth Estimation with Interpretable Uncertainty

Rémi MARSAL[*†] Florian CHABOT[*] Angelique LOESCH[*] William GROLLEAU[*]
Hichem SAHBI[†]
[*]Université Paris-Saclay, CEA, LIST, F-91120, Palaiseau, France
firstname.lastname@cea.fr
[†]Sorbonne University, CNRS, LIP6 F-75005, Paris, France
firstname.lastname@lip6.fr

Abstract

Self-supervised monocular depth estimation methods aim to be used in critical applications such as autonomous vehicles for environment analysis. To circumvent the potential imperfections of these approaches, a quantification of the prediction confidence is crucial to guide decision-making systems that rely on depth estimation. In this paper, we propose MonoProb, a new unsupervised monocular depth estimation method that returns an interpretable uncertainty, which means that the uncertainty reflects the expected error of the network in its depth predictions. We rethink the stereo or the structure-from-motion paradigms used to train unsupervised monocular depth models as a probabilistic problem. Within a single forward pass inference, this model provides a depth prediction and a measure of its confidence, without increasing the inference time. We then improve the performance on depth and uncertainty with a novel self-distillation loss for which a student is supervised by a pseudo ground truth that is a probability distribution on depth output by a teacher. To quantify the performance of our models we design new metrics that, unlike traditional ones, measure the absolute performance of uncertainty predictions. Our experiments highlight enhancements achieved by our method on standard depth and uncertainty metrics as well as on our tailored metrics. https://github.com/CEA-LIST/MonoProb

1. Introduction

Advances in deep learning in the field of computer vision have led to breakthroughs in depth estimation [7, 24]. This task is crucial for applications like autonomous driving, as it provides an analysis of the environment that can inform about the presence of obstacles, for instance. Therefore, it must be sufficiently reliable to be used for decision-making. In particular, the expected common criteria are high performance, fast inference, and the ability to quantify the confidence in the model predictions. However, these requirements often conflict with the inherent limitations of traditional deep learning methods. Indeed, these approaches require extensive labeled datasets, incurring substantial additional costs. Furthermore, they are black-box systems, yielding predictions without reliability clue. Considering supervised learning for monocular depth estimation, labeled data can be provided by recording videos of multiple scenes while performing a synchronized lidar acquisition at the same time [2, 26]. An alternative consists in training a model on synthetic data [13] which nevertheless introduces a domain gap between the training set and real use-case data.

The challenge of acquiring labeled data can be mitigated through unsupervised training strategies, where the objective is to minimize the reconstruction error between source and target images captured from a different perspective [9, 10] or at a different instant [39] within the same scene. The lack of confidence can be addressed by ensembling methods [19, 29] that combine the results of multiple inferences from one or more models to obtain a variance for each prediction. However, the aforementioned approaches incur computational overhead during both training and inference, particularly in the context of bootstrap ensembles. Alternatively, predictive methods [19, 22, 29] return complementary outputs in addition to the depth map to quantify the uncertainty. Since they require only one inference per frame, they are more attractive for real-time applications.

In the case of supervised learning, a simple and direct uncertainty estimation approach is to model depth with a probability distribution [19]. The estimator returns the parameters of this distribution instead of scalars and the traditional distance to the ground truth as loss function is replaced by the likelihood maximization of the predicted depth distri-

bution. Notably, this method offers the advantage of providing interpretable uncertainty values within a single inference that is the depth variance. In this paper, we consider that an uncertainty is interpretable if it gives an estimate of the expected error between the predicted depth and the ground truth. However, this technique depends on the availability of ground truth data. In the context of unsupervised learning, several works [22, 29] adapt this approach by modeling the image reconstruction by a probability distribution and maximizing the likelihood of the reconstruction. By leveraging the weight of pixels where the supervisory signal from the reconstruction loss is unreliable, they achieve performance enhancements. Still, this uncertainty is difficult to use in practical applications due to its complex link with depth that limits its interpretability.

In this paper, (1) we first present MonoProb, an unsupervised depth estimation method that provides an interpretable confidence measure of its predictions. This approach extends the likelihood maximization strategy to unsupervised learning. Unlike [22, 29] that only model the reconstruction by a probability distribution, we express this probabilistic reconstruction with respect to a probability distribution over depth. As a result, this technique provides an interpretable and reliable uncertainty relative to depth in a single inference, enabling informed subsequent decision-making in real-time. This uncertainty is the standard deviation (STD) of a predicted depth distribution. Furthermore, we demonstrate the ability of MonoProb to improve performance on depth estimation. (2) Second, [29] estimates uncertainty using a self-distillation method where pseudo ground truth scalar depth maps from a teacher model supervise a student model. We enhance the quality of our uncertainty predictions by adapting this approach to handle probability distributions as pseudo ground truth provided by the teacher. (3) Finally, we propose two new metrics tailored for evaluating the quality of interpretable uncertainty. To compare different types of uncertainties, specifically non-interpretable ones, [17, 29] introduced metrics that are invariant to an increasing bijection on uncertainty. They measure a relative uncertainty within an image by indicating the effectiveness of uncertainty in sorting image pixels by order of performance on a given depth metric. Conversely, our metrics are designed by considering the ability of the network to predict its absolute performance.

2. Related works

2.1. Unsupervised monocular depth estimation

When no ground truth is available, image reconstruction stands as a widely popular pretext task. It enables the model to learn depth by linking it to the apparent motion of pixels characterizing the same object between two images of a scene in a unique way. This technique also requires the camera's intrinsic calibration and the relative position of the camera between two views of a scene. Pioneering works use either stereo images [9, 10] or image sequences [39] for this purpose, then [11, 30, 38] combine them both at training. Subsequent works improve performance by guiding the model training with traditional non-learning based methods [22, 35], using learned features [32], enhancing network architecture design, incorporating semantic guidance [3, 14, 21] or regularizing training with self-distillation [1, 27, 28].

2.2. Uncertainty for supervised learning

The study of uncertainty estimation for depth estimation task in deep learning has gained significant attention in recent years. These methods can be categorized into two families: ensembling and predictive methods. Ensembling methods assume that neural network weights follow a probability distribution. Uncertainty is then given by the variance of multiple predictions obtained by sampling several model weights. Bayesian neural networks [4, 25, 36] explicitly assign a probability distribution to each individual weight. Monte Carlo dropout [8] consists in randomly switching off network weights. Bootstrapped ensembles [23] involve training several models on different subsets of the training dataset, while in Snapshot Ensembles [16] store multiple models from a single training. On the other hand, predictive methods [19] aim at returning direct quantification of the uncertainty in their outputs. These methods are less computationally intensive, requiring only a single inference step, although they require an adjustment of the loss function. The outputs belong to a family of distributions whose parameters are estimated by a neural network. The loss function to be minimized is the negative log-likelihood (NLL) of the actual ground truth, making it incompatible for direct use in unsupervised learning scenarios. In our work, we adapt this approach for the unsupervised learning of a probabilistic depth distribution.

2.3. Uncertainty for unsupervised depth estimation

Early unsupervised monocular depth estimation methods include uncertainty reconstruction. The explainability mask of Zhou et al.'s method [39] encodes the network's confidence in its ability to reconstruct each pixel. Since it is used to weight each pixel of the reconstruction loss differently, it needs a regularization term to prevent a trivial solution. [22,37], inspired by [19], remove the regulation term and incorporate the uncertainty map in a likelihood maximization-like problem of reconstruction where the reconstruction error is modeled by a Laplacian distribution. Thus, the predicted variance quantifies how likely the reconstruction loss is to be minimized. However, this is not exactly equivalent to uncertainty on depth because on photorealistic datasets [2, 26, 31], even a perfectly estimated depth map can produce reconstruction failures due to phenomena like bright-

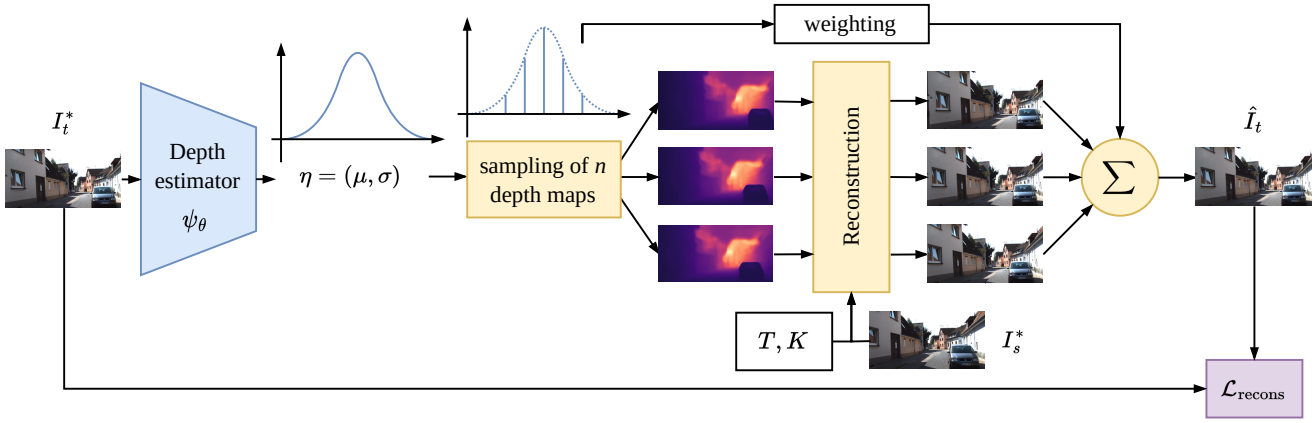

Figure 1. Our depth estimator takes the target image I_t^* as input and returns a map of the parameters $\eta = (\mu, \sigma)$ of a multi-variate depth distribution D. A sampling yields n depth maps used to reconstruct n times I_t^*. These reconstructions also involve a source image I_s^*, the intrinsic camera calibration K, and the camera motion T between I_s^* and I_t^*. They are then weighted according to the parameters η of the distribution D and averaged. This gives the final reconstruction $\hat{I}_t$, which is compared to the original I_t^* in the $\mathcal{L}_{\text{recons}}$ loss.

ness changes, non-Lambertian surfaces or highly detailed areas. Nevertheless, it can serve as a useful tool for approximating depth uncertainty as shown by Poggi et al. [29]. They provide a broad synthesis of techniques that can be used to estimate uncertainty, ranging from empirical approaches such as dropout or ensembling methods to predictive approaches such as likelihood maximization of the reconstruction loss or of a pseudo ground truth, or the prediction of the reconstruction loss. They also combine empirical and predictive approaches. Several works [12, 18]. reformulate depth estimation into a classification problem of discrete disparities that can be used to estimate a variance. Dikov et al. [6] propose the first predictive approach of an interpretable uncertainty that is depth STD. The reconstruction loss is reformulated as a reconstruction likelihood expressed with respect to the depth distribution thanks to Bayes' theorem. Instead, we chose the law of total probability to make the depth distribution appear in the reconstruction likelihood. This enables to avoid learning pseudo-inputs as in [6] to get a prior distribution on depth. We demonstrate that this improves both depth and uncertainty performance.

3. Unsupervised depth estimation

Consider a collection of image pairs denoted as $\mathcal{I} = \{(I_s^*, I_t^*)_i\}_i$ where each image I_j^*, $j \in \{s, t\}$ in a pair abides by $I_j^* \in \mathbb{R}^{H \times W \times C}$ and H, W, C respectively stand for the height, the width and the number of channels. We assume each image of a pair represents a different point of view of a same scene which verifies the brightness consistency assumption and has no uniformly textured surface. The change of point of view can result from a small movement of the camera relative to the scene (e.g. stereo images) or from a different timestamp (as in structure-from-motion methods [34]). The brightness consistency assumption means that objects in a scene keep a constant brightness between two different close views. Hence, pixel motion is sufficient to explain the transformation to be applied to an image named the source image I_s^* of an image pair in $\mathcal{I}$ to obtain the other one called the target image I_t^*. In a static scene, the new location p_s in I_s^* of a pixel p_t from I_t^* can be expressed as a function of the depth D_t of the target image, the camera motion $T_{t \to s}$ from the target to the source image and the camera calibration K with the formula:

$$p_s = K T_{t \to s} D_t(p_t) K^{-1} p_t. \tag{1}$$

For each pixel p_t, sampling in the source image at location p_s results in a new image that is a reconstruction of the target image I_t^*. Thus, minimizing the reconstruction error provides a relevant pretext task for the unsupervised learning of depth, given camera motion and calibration.

4. Method

4.1. Reconstruction loss with probabilistic depth

We consider a depth uncertainty to be interpretable if its value directly informs about the expected depth error. To provide depth with an interpretable uncertainty, we choose to model depth by a distribution D whose mean μ_D is used as depth prediction and STD σ_D as uncertainty prediction. The unsupervised learning framework described in Sec. 3 is re-designed into a probabilistic paradigm. For this purpose, the image reconstruction is modeled by a distribution that is expressed as a function of the depth distribution D by applying the law of total probability. Since this formulation requires the costly computation of an integral over depth, we propose several approximations to make it

Sup	Resolution	#Trn	Abs Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Abs Rel		RMSE		$\delta < 1.25$		ARU↓	RMSU↓
						AUSE ↓	AURG↑	AUSE ↓	AURG↑	AUSE ↓	AURG↑		
M	[11]	1	0.090	3.942	**0.914**	-	-	-	-	-	-	-	-
M	[29]-Repr	1	0.092	3.936	0.912	0.051	0.008	2.972	0.381	0.069	0.013	-	-
M	[29]-Log	1	0.091	4.052	0.910	0.039	0.020	2.562	0.916	0.044	0.038	-	-
M	**Ours**	1	**0.089**	**3.852**	**0.914**	**0.031**	**0.026**	**0.719**	**2.560**	**0.030**	**0.050**	**0.064**	**2.912**
M	[29]-Self	2	**0.087**	3.826	**0.920**	0.030	0.026	2.009	1.266	0.030	0.045	0.074	3.730
M	**Ours-self**	2	**0.087**	**3.762**	0.919	**0.022**	**0.034**	**0.326**	**2.880**	**0.014**	**0.061**	**0.066**	**2.969**
S	[11]	1	0.085	3.942	0.912	-	-	-	-	-	-	-	-
S	[29]-Repr	1	0.085	3.873	0.913	0.040	0.017	2.275	1.074	0.050	0.030	-	-
S	[29]-Log	1	0.085	3.860	0.915	**0.022**	**0.036**	0.938	2.402	**0.018**	**0.061**	-	-
S	**Ours**	1	**0.084**	**3.834**	**0.916**	0.023	0.033	**0.661**	**2.655**	0.023	0.055	**0.075**	**3.540**
S	[29]-Self	2	**0.084**	3.835	**0.915**	0.022	0.035	1.679	1.642	0.022	0.056	0.083	3.686
S	**Ours-self**	2	**0.084**	**3.792**	0.914	**0.018**	**0.038**	**0.349**	**2.924**	**0.019**	**0.060**	**0.072**	**3.068**
MS	[11]	1	0.084	**3.739**	**0.918**	-	-	-	-	-	-	-	-
MS	[29]-Repr	1	0.084	3.828	0.913	0.046	0.010	2.662	0.635	0.062	0.018	-	-
MS	[29]-Log	1	**0.083**	3.790	0.916	0.028	**0.029**	1.714	1.562	**0.028**	**0.050**	-	-
MS	**Ours**	1	0.084	3.806	0.915	**0.027**	**0.029**	**0.840**	**2.436**	0.029	0.049	**0.077**	**3.573**
MS	[29]-Self	2	0.083	3.682	**0.919**	0.022	0.033	1.654	1.515	0.023	0.052	0.083	3.686
MS	**Ours-self**	2	**0.082**	**3.667**	**0.919**	**0.016**	**0.039**	**0.293**	**2.859**	**0.014**	**0.061**	**0.078**	**3.528**

Table 1. Results of monocular only (M), stereo only (S) and monocular and stereo (MS) trainings of our MonoProb with and without self-distillation compared to other methods.

tractable. This leads to a reformulation of the reconstruction loss. Our method is illustrated in Fig. 1.

Probabilistic model. The reconstruction of the target image is modeled by a distribution I_t. Using the law of total probability, the likelihood of the reconstruction conditionally on I_s^*, $I \mapsto \mathrm{p}_{I_t}(I|I_s^*)$ can be expressed with respect to depth distribution D defined on $\mathcal{D}$ so that:

$$\mathrm{p}_{I_t}(I \mid I_s^*) = \int_{\mathcal{D}} \mathrm{p}_{I_t}(I \mid d, I_s^*)\, \mathrm{p}_D(d)\, \mathrm{d}d. \tag{2}$$

Given $\mathrm{p}_{I_t}(I|d, I_s^*)$, maximizing the likelihood $\mathrm{p}_{I_t}(I|I_s^*)$ with respect to p_D for $I = I_t^*$ provides an estimator of D.

Neural network estimator. The depth distribution D is assumed to belong to a family of distributions $\mathcal{L}_\eta$ with unknown parameters η. We introduce ψ_θ a function with learnable parameters θ that returns the η parameters of D given an image pair $(I_t^*, I_s^*) \in \mathcal{I}$. The likelihood of the reconstruction becomes:

$$\mathrm{p}_{I_t}(I \mid I_s^*, \theta) = \int_{\mathcal{D}} \mathrm{p}_{I_t}(I \mid d, I_s^*) \mathrm{p}_D(d \mid \eta) \mathrm{d}d, \tag{3}$$

with $\eta = \psi_\theta(I_t^*)$. Thus, ψ_θ is an estimator of D by minimizing the negative log-likelihood of the reconstruction with respect to the parameters θ when $I = I_t^*$. Let $\mathrm{recons}(.,.)$ be a reconstruction function that takes as input a punctual estimate of depth and an image, and returns an image so that $\mathrm{recons}(D^*, I_s^*) = I_t^*$ and $\mathrm{err}(.,.)$ a distance between two images, we define $\mathrm{p}_{I_t}(I \mid d, I_s^*) = \frac{1}{\lambda} \exp(-\mathrm{err}(\mathrm{recons}(d, I_s^*), I))$. The parameter λ aims at enforcing the upper bound of the cumulative distribution function relative to $\mathrm{p}_{I_t}(I|d, I_s^*)$ to be equal to 1, thus $\lambda = \int_{\mathcal{D}} \exp(-\mathrm{err}(\mathrm{recons}(d, I_s^*), I))\mathrm{d}d$. The choice for $\mathrm{p}_{I_t}(I \mid d, I_s^*)$ enables to recover [11]'s reconstruction error when computing the negative log-likelihood of $\mathrm{p}_{I_t}(I \mid I_s^*)$ with a punctual distribution as D distribution.

Approximations of the reconstruction likelihood. In this paper, ψ_θ is a neural network that returns a map η of the parameters of the distribution $D \in \mathcal{D} = \mathbb{R}^{H \times W \times 1}$ of depth. To make the prediction of η tractable with usual state-of-the-art neural networks, ψ_θ outputs are restricted to the parameters of the $H \times W \times 1$ marginal distributions of D. This means that $\mathrm{p}_D(d|\eta)$ in Eq. (3) is partially unknown. A solution to this problem is to minimize an approximation to the upper bound of the negative log-likelihood of the reconstruction rather than the negative log-likelihood itself. Thus, using the convex property of the $x \mapsto -\log(x)$ function, the Jensen inequality can be applied as follows:

$$\begin{aligned} -\log \mathrm{p}_{I_t}(I \mid I_s^*, \theta) &= -\log \int_{\mathcal{D}} \mathrm{p}_{I_t}(I \mid d, I_s^*) \mathrm{p}_D(d \mid \eta) \mathrm{d}d \\ &= -\log \mathbb{E}_D[\mathrm{p}_{I_t}(I \mid D, I_s^*) \mid \eta] \\ &\leq \mathbb{E}_D[-\log \mathrm{p}_{I_t}(I \mid D, I_s^*) \mid \eta]. \end{aligned} \tag{4}$$

Then this upper bound is approximated:

$$\begin{aligned} \mathbb{E}_D[-\log \mathrm{p}_{I_t}(I|D, I_s^*)|\eta] &= \mathbb{E}_D[\mathrm{err}(\mathrm{recons}(D, I_s^*), I)|\eta] + c \\ &\approx \mathrm{err}(\mathbb{E}_D[\mathrm{recons}(D, I_s^*)|\eta], I) + c. \end{aligned} \tag{5}$$

where $c = -\log(\lambda)$ is a constant. The case of equality occurs when the depth estimator ψ_θ returns a punctual distribution, i.e. with a null scale parameter. This means that the model is absolutely certain about its prediction. This is a behavior that ψ_θ would tend towards if it had an infinite capacity and was trained with an infinite dataset verifying the brightness consistency assumption and without any

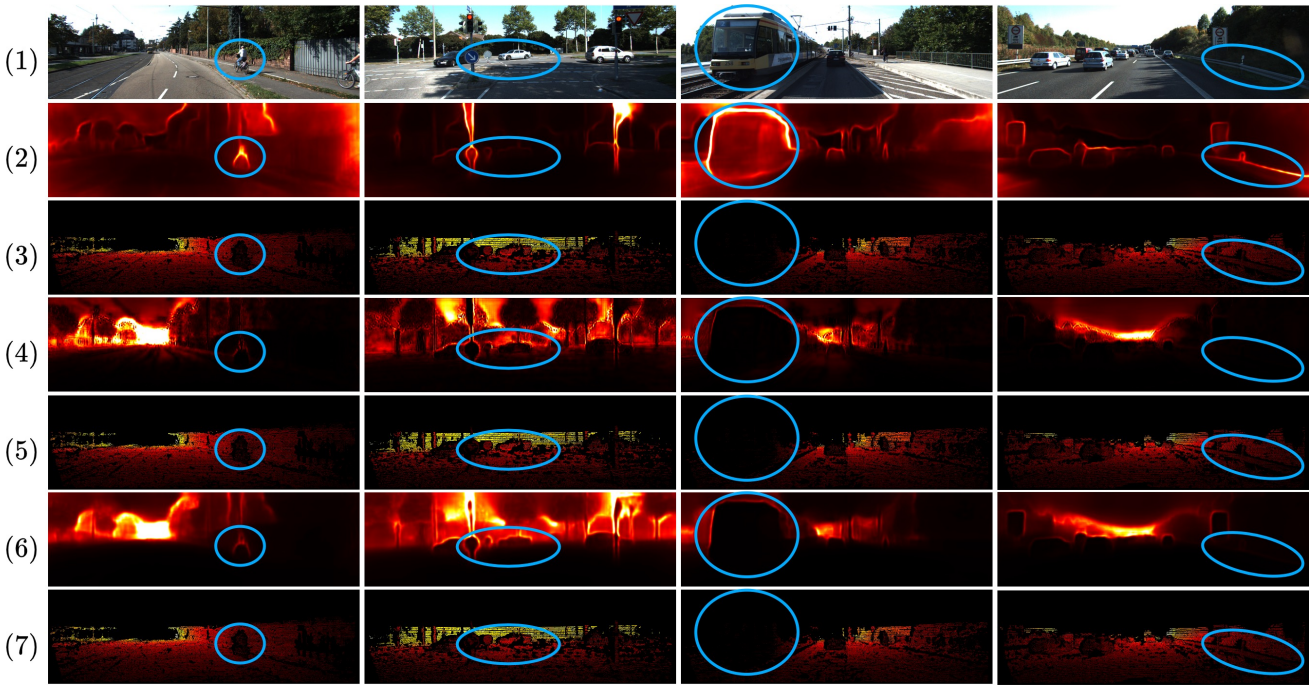

Figure 2. Qualitative results of trainings on KITTI monocular videos. (1) Input image, (2) Uncertainty map from [29]-Self, (3) Depth error map from [29]-Self, (4) Uncertainty map from our MonoProb without self-distillation, (5) Depth error map from our MonoProb without self-distillation, (6) Uncertainty map from our MonoProb with self-distillation, (7) Depth error map from our self-distilled MonoProb.

Methods	Abs Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Abs Rel		RMSE		$\delta < 1.25$	
				AUSE ↓	AURG↑	AUSE ↓	AURG↑	AUSE ↓	AURG↑
VDN [6]	0.117	4.815	0.873	0.058	0.018	1.942	2.140	0.085	0.030
Ours	**0.114**	**4.772**	**0.875**	**0.044**	**0.030**	**1.625**	**2.437**	**0.060**	**0.054**

Table 2. Comparison with VDN [6] on KITTI with the raw ground truth.

uniformly textured surface. In this situation, the model is always capable of finding the unique depth prediction that minimizes the reconstruction error. In our experiments, we assume that the datasets are sufficiently large, that their images respect the aforementioned ideal properties and that the neural network architecture is wide enough to apply the approximation in Eq. (5).

Finally, it is sufficient to have only the marginal values of the multivariate distribution D to compute the expectation in Eq. (5) since the recons$(.,.)$ function only applies pixel-wise operations (see proof in the supplementary material).

Sampling strategy. The expectation in Eq. (5) is approximated with a sampling strategy. The samples cannot be random because this would not allow computing the gradients of the distribution parameters that are later used in the back-propagation algorithm to update the parameters θ of ψ_θ. Instead, for each marginal distribution of D, a set of n samples $\mathcal{S}_\eta$ is defined so as to accurately represent the predicted distribution (more details in the supplementary material). At the end of the sampling stage, n depth maps of size HW are obtained, from which n reconstructions are computed. These reconstructions are weighted according to the sample used and summed. Thus, the final reconstruction loss $\mathcal{L}_{\text{recons}}$ is:

$$\mathcal{L}_{\text{recons}}(\theta) = \text{err}\left(\frac{\sum_{d\in\mathcal{S}_\eta} \text{p}_D(d|\eta)\text{recons}(d, I_s^*)}{\sum_{d\in\mathcal{S}_\eta} \text{p}_D(d|\eta)}, I_t^*\right). \tag{6}$$

4.2. Self-distillation training

[29] is the first unsupervised monocular depth estimation method that requires only one inference to provide an interpretable uncertainty. It consists in training an unsupervised monocular depth estimation *teacher* model without uncertainty, then using its predictions as pseudo ground truth in the supervised training of a *student* model. This *student* model returns the parameters of a probability distribution of the depth D_s. The loss function is the negative log-likelihood. We redesign this loss in case the *teacher* already returns a distribution D_t. We train the *student* so that its predicted distribution matches the frozen

Sup	Resolution	Self-dist	Abs Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Abs Rel		RMSE		$\delta < 1.25$		ARU↓	RMSU↓
						AUSE ↓	AURG↑	AUSE ↓	AURG↑	AUSE ↓	AURG↑		
M	640 × 192		0.084	3.621	0.920	0.025	0.030	0.744	2.365	0.025	0.048	0.074	3.405
M	640 × 192	✓	**0.082**	3.570	0.926	0.022	0.031	**0.315**	**2.728**	0.015	0.054	0.063	2.870
M	1024 × 320		0.087	3.653	0.922	0.028	0.030	0.543	2.574	0.022	0.052	0.064	**2.720**
M	1024 × 320	✓	0.083	**3.481**	**0.928**	**0.020**	**0.034**	0.324	2.640	**0.011**	**0.056**	**0.061**	2.751
S	640 × 192		0.080	3.653	0.920	0.023	0.030	0.687	2.467	0.022	0.051	0.072	3.404
S	640 × 192	✓	0.079	3.606	0.922	**0.015**	**0.037**	0.303	**2.811**	0.013	**0.058**	0.075	3.451
S	1024 × 320		0.074	3.361	0.931	0.022	0.028	0.597	2.304	0.021	0.043	**0.065**	**3.061**
S	1024 × 320	✓	**0.073**	**3.315**	**0.933**	**0.015**	0.034	**0.277**	2.584	**0.012**	0.051	0.069	3.172
MS	640 × 192		0.080	3.486	0.924	**0.015**	**0.038**	0.288	**2.701**	**0.013**	**0.058**	0.076	3.336
MS	640 × 192	✓	0.079	3.456	0.925	0.025	0.028	0.602	2.358	0.024	0.045	0.071	3.175
MS	1024 × 320		0.076	3.338	0.927	0.023	0.027	0.571	2.296	0.022	0.046	**0.067**	**3.057**
MS	1024 × 320	✓	**0.075**	**3.241**	**0.929**	0.016	0.033	**0.275**	2.499	**0.013**	0.052	0.071	3.109

Table 3. Results of the M, S and MS MonoProb methods with the Resnet50 architecture and two different resolutions with and without self-distillation. These demonstrate the ability of our MonoProb method to work with other architectures and high-resolution images.

Dataset	Method	Abs Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Abs Rel		RMSE		$\delta < 1.25$		ARU↓	RMSU↓
					AUSE ↓	AURG↑	AUSE ↓	AURG↑	AUSE ↓	AURG↑		
Make3D	[11]	**0.322**	7.417	-	-	-	-	-	-	-	-	-
	[29]-Self	0.334	6.840	0.514	0.173	0.029	4.954	0.065	0.251	0.036	0.286	6.500
	Ours	0.333	6.729	0.514	0.124	0.079	1.966	2.977	0.231	0.060	0.271	5.705
	Ours-self	0.327	**6.687**	**0.521**	**0.112**	**0.087**	**1.583**	**3.335**	**0.219**	**0.070**	**0.267**	**5.781**
Nuscenes	[11]	0.226	10.043	0.653	-	-	-	-	-	-	-	-
	[29]-Self	0.220	9.765	0.662	0.127	0.009	7.446	0.575	0.224	0.017	0.197	9.518
	Ours	0.224	9.757	0.666	0.081	0.053	3.148	4.905	0.128	0.118	0.162	8.165
	Ours-self	**0.219**	**9.559**	**0.670**	**0.0720**	**0.060**	**1.644**	**6.218**	**0.099**	**0.141**	**0.175**	**8.498**

Table 4. Evaluation on the Make3D and Nuscenes datasets of our M models trained on KITTI with monocular images only shows that our method generalizes well to other datasets both for the depth and the uncertainty.

teacher distribution. For this purpose, our loss function is the Kullback-Leibler divergence between the *teacher* and the *student* distributions. Following the results in Tab. 7, we choose the Gaussian distribution as the family of distributions for the *teacher*: $D_t = \mathcal{N}(\mu_t, \sigma_t)$ and the *student*: $D_s = \mathcal{N}(\mu_s, \sigma_s)$. Thus, the self-distillation loss is $\mathcal{L}_{\text{self}}$:

$$\mathcal{L}_{\text{self}}(\theta) = D_{KL}(D_s \| D_t) = \log \frac{\sigma_t}{\sigma_s} + \frac{\sigma_s^2 + (\mu_s - \mu_t)^2}{2\sigma_t^2}. \tag{7}$$

5. Experiments

5.1. Absolute uncertainty metrics

The depth metrics are those of [29], in particular the absolute relative error (Abs Rel), the root mean square error (RMSE), and the amount of inliners ($\delta < 1.25$). We use the STD of the predicted distributions as a pixel-wise uncertainty estimate. Following [29], we compute the Area Under the Sparsification Error (AUSE) and the Area Under the Random Gain (AURG), which are relative metrics within an image. They measure the ability of the uncertainty estimation to sort pixels in order of descending error ϵ for a given depth error metric in an image. AUSE compares the sparsification curve obtained by this sorting to the perfect sparsification curve obtained by sorting pixels directly based on their error. AURG quantifies the improvement of the sparsification curve of the predicted uncertainty relative to a random sparsification curve where no uncertainty modeling is performed. More details about these metrics can be found in [29].

These metrics are designed to compare any kind of uncertainty metrics even if they are not directly related to depth but rather to the reconstruction quality for instance, as in [22, 29]. However, they suffer from some drawbacks: they cannot provide an absolute measure of uncertainty and they are not consistent from one image to another. Indeed, since they are based on sorting uncertainties within an image, these metrics are invariant within a growing bijection on the predicted uncertainties. Likewise, a given pair of uncertainty and corresponding error, will not have the same contribution to the global uncertainty metric depending on the performance of the other pixels of the same image.

Therefore, we introduce two new uncertainty metrics that measure the ability to anticipate the true depth error. Thus, we propose the Absolute Relative Uncertainty (ARU), which is relative to the ground truth depth, and the Root Mean Square Uncertainty Error (RMSU), which is not. Let D^*, $\hat{D}$ and U be respectively the ground truth depth map, the predicted depth map and the predicted uncertainty map for a single image:

$$\begin{aligned} \text{ARU} &= \|(U - |\hat{D} - D^*|) \oslash D^*\|_1 / HW \\ \text{RMSU} &= \sqrt{\|(U - |\hat{D} - D^*|)^2\|_1 / HW}, \end{aligned} \tag{8}$$

where $|.|$, $\oslash$ and $\|.\|_1$ denote the element-wise absolute value, the element-wise division and the ℓ_1-norm, respec-

Uncertainty prediction	Abs Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Abs Rel		RMSE		$\delta < 1.25$		ARU↓	RMSU↓
				AUSE ↓	AURG↑	AUSE ↓	AURG↑	AUSE ↓	AURG↑		
σ	0.418	12.047	0.321	0.226	0.003	9.087	0.045	0.346	0.006	0.418	12.047
α	**0.089**	**3.852**	**0.914**	**0.031**	**0.026**	**0.719**	**2.560**	**0.030**	**0.050**	**0.064**	**2.912**

Table 5. The ablation on how to predict uncertainty in unsupervised training with probabilistic image reconstruction shows that directly predicting σ is unable to train the model contrary to predicting α and then deducing $\sigma = \alpha \times \mu$.

Num samples	Abs Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Abs Rel		RMSE		$\delta < 1.25$		ARU↓	RMSU↓
				AUSE ↓	AURG↑	AUSE ↓	AURG↑	AUSE ↓	AURG↑		
5	0.092	3.895	0.911	0.033	0.025	0.729	**2.581**	0.033	0.049	0.066	2.968
9	**0.089**	**3.852**	**0.914**	**0.031**	0.026	**0.719**	2.560	**0.030**	0.050	**0.064**	**2.912**
13	0.092	3.907	0.912	0.032	**0.027**	0.753	2.573	**0.030**	**0.051**	0.066	2.952

Table 6. The ablation study on the number of samples provides an optimal value of 9 samples.

tively. These metrics are then averaged over all the ground truth depth maps of the dataset. By construction, these metrics are only suitable for quantifying uncertainty that is directly related to the depth estimation error.

5.2. Implementation details

We implement our method on top of Monodepth2 [11], an unsupervised monocular depth estimation approach that is trained with either monocular videos only (M), pairs of stereo image pairs only (S), or both monocular videos and stereo images (MS). When monocular videos are used for training (M or MS), an additional *pose* network is jointly trained with the depth network to provide the camera motion T. In contrast, when stereo images are used the camera motion is given by the translation between the two stereo cameras. Monodepth2 is also designed to handle occlusions and ignore dynamic objects that would disrupt the training. Thus, following [11], we finetune a Resnet18 network [15] pretrained on the ImageNet dataset [5] for 20 epochs with the Adam optimizer [20]. We use batches of 12 images resized to 192×640 and augmented as in [11]. We choose a Gaussian distribution $\mathcal{N}(\mu, \sigma)$ as the family of distributions for depth, so our depth network must return a two-channel map, one for each parameter. On one of them, we apply a sigmoid activation function to predict a disparity from which the mean μ of the depth distribution is deduced following [11]. When training with our probabilistic reconstruction loss, a sigmoid is also applied to the second channel so that it outputs a scalar $\alpha \in [0, 1]$ that weights the mean depth μ to obtain the STD $\sigma = \alpha \times \mu$. Instead, at the self-distillation stage, the network directly returns the σ by applying an exponential activation function as in [29]. The number of samples is set to 9. The learning rate is set to 10^{-4} and dropped to 10^{-5} over the last 5 epochs. We carry out experiments on the KITTI dataset [26], which is composed of 42K images extracted from 61 driving videos. Following common practice, we employ the popular Eigen split [7]. At evaluation, we cap depth to 80 meters and use the improved ground truth provided by [33]. The uncertainty is the STD of the predicted distributions.

5.3. Results on KITTI

We conducted experiments on the three depth estimation paradigms of [11]: M, S and MS. In Tab. 1, we compare the performance obtained with our baseline [11] that does not provide uncertainty and with predictive methods implemented in [29] (i.e. methods that directly predict an uncertainty quantification within a single inference). In accordance with Sec. 5.1, we cannot compute our new metrics (ARU and RMSU) on [29]-Repr and [29]-Log. Indeed, these methods respectively provide as uncertainty quantification, an estimate of the reprojection error and an estimate of the scale parameter of the distribution of the reconstruction error. For fair comparisons, we report the number of trainings (#Trn). In the following, we refer to our MonoProb method without self-distillation as raw-MonoProb and to our MonoProb with self-distillation as self-distilled-MonoProb. Qualitative results in Fig. 2 show that [29]-Self tends to underestimate the uncertainty of distant objects and background, and to overestimate the uncertainty of object edges compared to our methods (more qualitative results on KITTI including depth maps are available in the supplementary material).

First, let's describe the results of single training methods. For the M training, raw-MonoProb performs slightly better on depth and significantly improves performance on uncertainty metrics. For the S training, the same behavior is observed for depth metrics: raw-MonoProb is a bit more efficient on depth than the other methods that require a single training. In terms of uncertainty, our raw-MonoProb results tend to be equivalent to [29]-Log. For the MS training, the depth performance of our raw-MonoProb is slightly below [11] and very close to other methods that require a single training to predict an uncertainty. The results on uncertainty metrics show that our method is either better than or very close to other comparable methods. For self-distilled methods, we observe the same behavior regardless of the training paradigm (M, S, or MS). Our self-distilled-MonoProb performs similarly to [29]-Self (trained with self-distillation on scalar pseudo ground truth) on depth metrics but the uncer-

Family of distributions	Abs Rel ↓	RMSE ↓	$\delta < 1.25$ ↑	Abs Rel		RMSE		$\delta < 1.25$		ARU↓	RMSU↓
				AUSE ↓	AURG↑	AUSE ↓	AURG↑	AUSE ↓	AURG↑		
Laplace	0.091	3.913	**0.914**	0.032	**0.027**	0.857	2.481	**0.030**	**0.050**	0.099	3.816
Normal	**0.089**	**3.852**	**0.914**	**0.031**	0.026	**0.719**	**2.560**	**0.030**	**0.050**	**0.064**	**2.912**

Table 7. The ablation of the family of distributions for D shows that the normal distribution performs better than the Laplace distribution.

Methods	Abs Rel↓	RMSE↓	$\delta < 1.25$ ↑	Abs Rel		RMSE		$\delta < 1.25$		ARU↓	RMSU↓
				AUSE↓	AURG↑	AUSE↓	AURG↑	AUSE↓	AURG↑		
NLL	0.088	3.785	0.917	**0.022**	0.034	0.330	2.897	0.016	**0.061**	**0.066**	**2.992**
KL-Div	**0.087**	**3.781**	**0.919**	**0.022**	**0.035**	**0.322**	**2.904**	**0.015**	**0.060**	**0.066**	2.997

Table 8. Comparing the performance of self-distillation with negative log-likelihood (NLL) where the pseudo ground truth is a depth map, and the Kullback-Leibler divergence (KL-Div) where the pseudo ground truth is a map of depth distributions, highlights a slight improvement with the Kullback-Leibler divergence.

tainties of self-distilled-MonoProb are significantly better than those of [29]-Self. Interestingly we show that our raw-MonoProb outperforms [29]-Self on our new metrics ARU and RMSU. We also show in the supplementary material that our MonoProb has similar or better performance than uncertainty methods that require more than one inference.

We evaluate our M model on KITTI with the raw ground truth for comparison with the VDN approach [6] in Tab. 2. This shows a higher effectiveness of our method for both depth estimation and uncertainty quantification. Finally, we provide results of our method with the Resnet50 architecture [15] and high-resolution images in Tab. 3. This shows that MonoProb also works on different architectures and image resolutions as the results are better than those of Tab. 1.

5.4. Generalization to other datasets

We also evaluate the generalization ability of our method on two other datasets of outdoor and urban scenes: Make3D [31] and Nuscenes [2]. The results on both datasets (see Tab. 4 and supplementary for qualitative results) show that our method without self-distillation is equivalent to the self-distilled approach of [29] in terms of depth performance and better at estimating uncertainty. Our self-distilled models are more accurate than other methods in both depth and uncertainty estimation. This shows that our method has a good generalization ability on images from different datasets.

5.5. Ablation

To justify the impact of our contributions and the interest of our implementation choices, we conduct an ablation study including the computation of the STD σ, the family distributions for depth, the number of samples and the self-distillation loss. First, we show the benefit of predicting α before deducing $\sigma = \alpha \times \mu$ instead of directly predicting σ when training with our probabilistic reconstruction loss. Experiments in Tab. 5 highlight that a direct prediction of σ prevents the network from learning anything, leading to very low performance compared to our strategy of considering the σ as a fraction of the mean depth μ. Studying the number of samples yields an optimal value of 9 in Tab. 6. The ablation in Tab. 7 highlights that normal distributions perform better overall than Laplace distributions as the family of distributions for depth. Finally, the experiments in Tab. 8 show the influence of our self-distillation loss: fitting two distributions with the Kullback-Leibler divergence outperforms minimizing the negative log-likelihood. As a matter of reproducibility, we provide the expression of the negative log-likelihood loss $\mathcal{L}_{\text{NLL}}$ we use in this ablation. For a fair comparison, we assume the student follows a normal distribution $D_s = \mathcal{N}(\mu_s, \sigma_s)$:

$$\mathcal{L}_{\text{NLL}}(\theta) = (\mu_s - \mu_t)^2/(2\sigma_s^2) + \log(\sigma_s), \tag{9}$$

μ_t being the mean of the teacher's depth distribution used as pseudo ground truth.

6. Conclusion

In this paper, we propose MonoProb an unsupervised method for training monocular depth estimation networks that returns an interpretable uncertainty within a single inference. This uncertainty, which anticipates the prediction errors, is the STD of the depth distribution returned by the depth estimation network. We also introduce a new self-distilled loss to train another network using a depth distribution as pseudo ground truth. Finally, we design new metrics that are better suited to measure the performance of interpretable uncertainty, i.e., uncertainty that is a direct anticipation of depth prediction errors. Through extensive experiments, we highlight that MonoProb improves performance relative to other unsupervised depth estimation methods that provide a quantification of uncertainty. We also demonstrate the cross-domain generalization ability of our method, that it works on different neural network architectures and on high-resolution images.

Acknowledgment This work was performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011013778) and the FactoryIA supercomputer, financially supported by the Ile-De-France Regional Council. Thanks to Riccardo Finotello for his helpful discussions.

References

[1] Juan Luis Gonzalez Bello and Munchurl Kim. Self-supervised deep monocular depth estimation with ambiguity boosting. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):9131–9149, 2021. 2

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 8

[3] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[4] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of ICML*, 2014. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 7

[6] Georgi Dikov and Joris van Vugt. Variational depth networks: Uncertainty-aware monocular self-supervised depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3, 5, 8

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 2014. 1, 7

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of ICML*, 2016. 2

[9] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 4, 6, 7

[12] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *Advances in Neural Information Processing Systems*, 2020. 3

[13] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[14] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 8

[16] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *ICLR*, 2017. 2

[17] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[18] Adrian Johnston and Gustavo Carneiro. 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 (CVPR)*, 2020. 3

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 2017. 1, 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 7

[21] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[22] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 6

[23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 2017. 2

[24] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[25] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. 2

[26] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 7

[27] Rui Peng, Ronggang Wang, Yawen Lai, Luyang Tang, and Yangang Cai. Excavating the potential capacity of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[28] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[29] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[30] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *Proceedings of the IEEE international conference on 3D Vision (3DV)*, 2018. 2

[31] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Learning 3-d scene structure from a single still image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2007. 2, 8

[32] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[33] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *Proceedings of the IEEE international conference on 3D Vision (3DV)*, 2017. 7

[34] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017. 3

[35] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[36] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of ICML*, 2011. 2

[37] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. 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 (CVPR)*, 2020. 2

[38] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[39] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2