electronics

MDPI

Article

The Constraints between Edge Depth and Uncertainty for Monocular Depth Estimation

Shouying Wu [1], Wei Li [2,*], Binbin Liang [2] and Guoxin Huang [1]

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China; shouyingwu531@gmail.com (S.W.); huangguoxin@stu.scu.edu.cn (G.H.)
[2] School of Aeronautics and Astronautics, Sichuan University, Chengdu 610065, China; sculiang@126.com
* Correspondence: li.wei@scu.edu.cn

Abstract: The self-supervised monocular depth estimation paradigm has become an important branch of computer vision depth-estimation tasks. However, the depth estimation problem arising from object edge depth pulling or occlusion is still unsolved. The grayscale discontinuity of object edges leads to a relatively high depth uncertainty of pixels in these regions. We improve the geometric edge prediction results by taking uncertainty into account in the depth-estimation task. To this end, we explore how uncertainty affects this task and propose a new self-supervised monocular depth estimation technique based on multi-scale uncertainty. In addition, we introduce a teacher–student architecture in models and investigate the impact of different teacher networks on the depth and uncertainty results. We evaluate the performance of our paradigm in detail on the standard KITTI dataset. The experimental results show that the accuracy of our method increased from 87.7% to 88.2%, the *AbsRel* error rate decreased from 0.115 to 0.11, the *SqRel* error rate decreased from 0.903 to 0.822, and the *RMSE* error rate decreased from 4.863 to 4.686 compared with the benchmark Monodepth2. Our approach has a positive impact on the problem of texture replication or inaccurate object boundaries, producing sharper and smoother depth images.

Keywords: monocular depth estimation; self-supervised method; uncertainty estimation

Citation: Wu, S.; Li, W.; Liang, B.; Huang, G. The Constraints between Edge Depth and Uncertainty for Monocular Depth Estimation. *Electronics* **2021**, *10*, 3153. https://doi.org/10.3390/electronics10243153

Academic Editor: Byung Cheol Song

Received: 10 November 2021
Accepted: 14 December 2021
Published: 17 December 2021

1. Introduction

Monocular depth estimation refers to the ability to learn a dense depth map at the pixel level from the video stream. It is a fundamental challenge in the field of computer vision with potential applications in robotics, autonomous driving, 3D reconstruction, and medical imaging [1–4]. How to predict a high quality dense depth map remains a problem to be solved. As the edges of objects in the image are prone to noise, bleeding, feature shifts, and surface field curvature changes can lead to distortion of the depth around the object. Even with high-precision camera equipment, these factors are inevitably introduced in the process of acquiring image data.

The exploration of edge depth has been around since long before deep learning became prevalent. By edge depth, we mean the depth of the edges of objects in the depth map. In the field of free point-of-view television technology, to improve the depth characteristics of object edges, Liu et al. [5] classified the pixels of the depth map and proposed a modified smoothness function to improve the accuracy of object edge depth values. However, monocular depth estimation methods have paid little attention to occlusion and image detail distortion. Related research is reflected in edge-aware depth estimation techniques.

To solve the problem of image edge detail distortion,Chou et al. [6] investigated the effect of the quality of the depth map on the synthetic focusing performance and proposes a synthetic focusing paradigm that integrates RGB images and depth information. In the task of reconstructing depth from raw video data using neural networks, Yang et al. [7] proposed using surface normal vectors to constrain the estimated depth. They constructed a depth-normal consistency that perceived the edges of objects inside the image.

Additional work includes estimating the depth map from the light field by sparse depth edges and gradients [8] and extracting RGB image edges and depth edges for alignment to improve the estimation accuracy [9].

Currently, most of the work related to improving edge depth requires the introduction of an additional network, e.g., semantic segmentation [10–12], edge map detection networks [13–15], or optical flow [16]. We found that research on uncertainty, which has only recently entered the limelight, can also improve the quality of edge depth and without learning other complex networks. Uncertainty is defined into two categories, epistemic and aleatoric [17].

The former can be used to understand examples that are different from those inside a training set, such as new scenes or new targets under which the model will predict the wrong depths with high probability, and such wrong depth results need to be detected. The latter can correctly learn the uncertainty (confidence) of the depth at the edge of the object, which is exactly what we require. The monocular depth estimation is mainly optimized for photometric error of the pixels, with little attention to the depth of geometric edges. This strategy of encouraging depth continuity can lead to misalignment of the edge depth in certain situations, e.g., when there is low texture, low luminance, or occlusion. To this end, we investigated the depth uncertainty of the monocular depth-estimation task.

In summary, to obtain an edge-aware depth estimation model, we introduce a depth uncertainty strategy, and use a framework in migration learning that allows our network to learn better quality models on monocular video sequences. Our main contributions are summarized as follows:

1. We study the impact of multi-scale uncertainty on self-supervised monocular depth estimation and find that it yields more edge-depth uncertainty.
2. We analyze the effect of different teacher–student combination strategies on the uncertainty of self-supervised monocular depth estimation.
3. We propose a paradigm for self-supervised monocular depth estimation based on a teacher–student framework and combined with multi-scale uncertainty. Our method effectively improves the overall performance of depth estimation.

We provide detailed experimental results on the KITTI [18] dataset. Our qualitative results show that our method obtains smoother results on the edges of people, cars, or road signs compared to previous work, which blends some of the edges with the background. The depth uncertainty map correctly represents the uncertainty of geometric edges and can restrict the learning of depth to edge pixels with large uncertainty.

2. Related Work

In this section, we review the relevant paradigms for self-supervised monocular depth estimation and how uncertainty estimation techniques are used on depth-estimation tasks.

2.1. Self-Supervised Methods

Before the prevalence of self-supervised methods, supervised methods were mainly used. Learning-based methods use the relationship between color images and their depth to fit prediction models, such as combining nonparametric scene sampling [19] or local prediction [20]. Later, end-to-end supervised models emerged [21,22]. The learning-based approach is the optimal representation of optical flow and stereo estimation.

However, we already know that it is difficult to obtain ground truth data in different real-world environments. How to predict the depth without labeled data is the goal pursued by all. To overcome this problem, self-supervised methods using image reconstruction have become a popular research topic, and they contain two classes of monocular video sequences and stereo pairs.

The paradigm of stereo pairs is to model the geometric properties between stereo image pairs as depth information, which is obtained by projection transformations between images [23]. Such an approach allows training the network based on the loss of photometric error between the actual image and the projected image. Later, model architectures for pose

prediction networks and depth prediction networks between frames of video sequences were developed [24].

Garg et al. [25] used L2 loss as the photometric error loss while producing ambiguity in the prediction results. Godard et al. [23] adopted a combination of L1 loss as well as SSIM loss to train the model named Monodepth, which was combined with post-processing operations to obtain more accurate depth accuracy. Then, they proposed the Monodepth2 [26] model, which solved the depth blurring problem in the object occlusion region to an extent by minimizing the reprojection loss for each pixel. Some specific network structures [27–30] or improved loss functions [31] have also appeared to optimize models.

There are also hybrid methods that use both stereo pair data and video sequence image frames [32,33]. Other refinement strategies, such as [34–36]. Some recent approaches used relatively bulky architectures to improve the depth quality [37], which indicates a higher memory cost as well as time.

2.2. Uncertainty Estimation

Uncertainty in decision making is crucial in real-world practical applications of computer vision, as it prevents overconfident and wrong decisions. Before the prevalence of neural networks, uncertainty was studied regarding stereo matching and optical flow problems. Stereo matching performs deep learning by inferring the differences in the network feature maps and estimating the difference maps [38]. Confidence inference for optical flow mainly consists of two types of posterior and model-inherent inferences. The first is by analyzing the uncertainty fraction of the optical flow field [39], and the second is by using the minimization energy module [40].

Recently, depth-estimation tasks have used uncertainty to improve model performance by adding confidence to the model output [41]. Depth estimation for 3D reconstruction can also introduce uncertainty in depth to improve the accuracy and robustness of learning [42]. To address depth estimation in regions without illumination, [43] extends the Gated2Depth framework by adding uncertainty to help filter the depth of these regions.

Our work focuses on solving the problem of inaccurate geometric edge depth estimation in complex scenes. Similar to our objective, [42] proposed a new photometric loss function using an uncertainty-based monocular depth estimation method to solve the problem of edge pixel pull due to object movement. They mainly rely on the proposed loss to constrain the pull of moving object depth and do not fully utilize the role of uncertainty in this problem—instead, only to evaluate the reliability of the output results.

In contrast, we consider the distribution properties of uncertainty at the geometric edges of images, which can be used to constrain the learning of object edge depth. Moreover, we find that it is not only the pulling problem of moving objects but also the case of occlusion where uncertainty makes a great contribution. By constraining the depth with uncertainty, it is possible to distinguish the depth of object in front from the depth of the obscured object behind.

Previous methods have predicted the depths of two objects with overlap to be the same depth. Figure 1 shows the depth estimation with edge awareness. The pulling phenomenon of moving character edges is not effectively addressed in monodepth2, while our method solves this problem by limiting the depth value of pixels with larger uncertainty in edge depth.

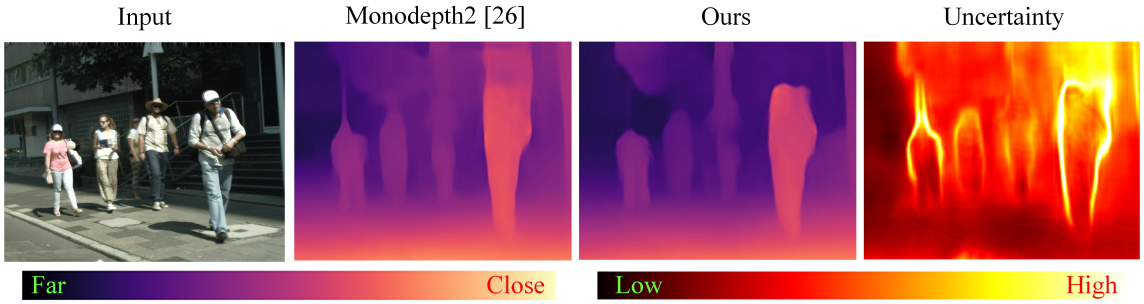

Figure 1. Depth map with edge-aware. The previous method (monodepth2) is prone to blurring on object boundaries. In contrast, our proposed method can predict smoother, sharper, and high-quality depth maps by considering the depth uncertainty.

3. Method

In this section, we first introduce the basic concepts of self-supervised depth estimation, then introduce techniques for uncertainty in depth-estimation tasks, and finally introduce the teacher–student frameworks and multi-scale uncertainty.

The overview diagram of our method is shown in Figure 2. To decouple the depth and pose networks when modeling uncertainty, we introduce the teacher–student framework. The teacher uses self-supervised monocular depth estimation, which incorporates a depth and pose network using a video sequence as input, where t′ represents adjacent frames. To investigate the effect of model uncertainty on depth estimation, we design different strategies, where BaseT (baseline as teacher), DropT (dropout as teacher), and BagT (bagging as teacher) denote different teacher networks. The parameter N denotes the number of depth and pose networks used by the strategies. The student includes only the depth network and incorporates the uncertainty distribution of depth.

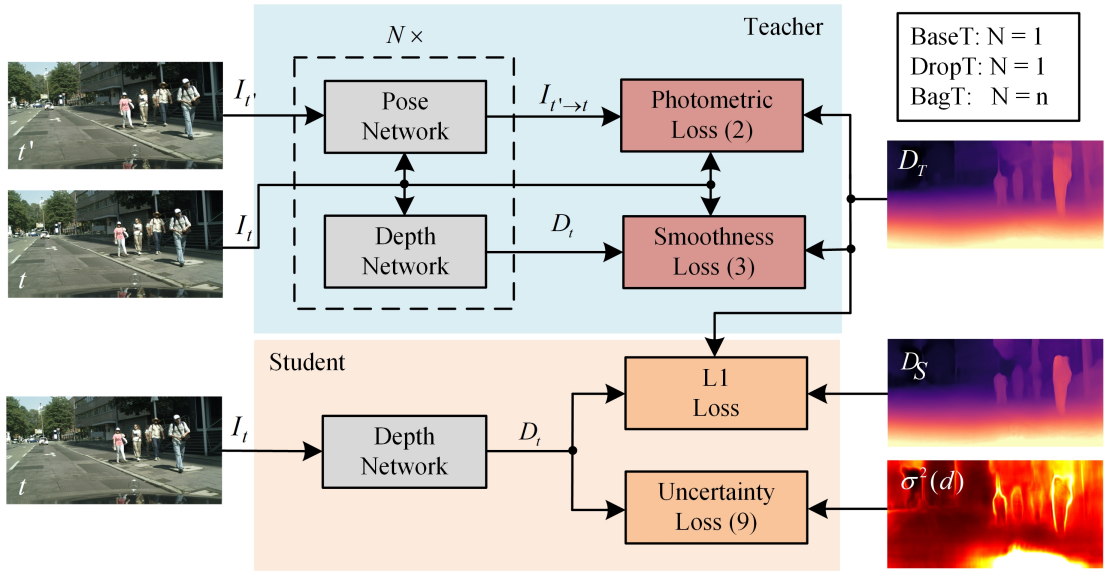

Figure 2. Overview of our proposed framework for joint uncertainty and depth prediction. The blue background corresponds to teacher network, and the gray blocks indicate encoders and decoders. The value of N in figure indicates the number of depth and pose networks taking the value of 1 for baseline or dropout as the teacher network, and N for bagging as the number of sub-networks. The light orange background corresponds to the student network. The red and orange blocks are the loss functions used, and the numbers inside represent the corresponding equations.

3.1. Self-Supervised Depth Estimation

The definition of self-supervised monocular estimation is to estimate the pixel-level depth value of frame sequences without ground truth. The geometric constraints between

multiple frames are used as the supervision signal. Specifically, the image frame at discrete time t' is warped to the previous frame at time t in the training process:

$$I_{t'\to t} \sim KT_{t\to t'}D_t(p_t)K^{-1}I_{t'}, \tag{1}$$

where K is the known intrinsic matrix of the camera, and $T_{t\to t'}$ indicates spatial transformation between image I_t and $I_{t'}$. The pixel in the image is expressed as p_n, and $D_t(p_t)$ represents the depth value of all pixels in the image at t. Therefore, we can obtain depth D_t and the spatial transformation $T_{t\to t'}$ through a depth and pose network. Generally, the more popular choice for the loss function $\mathcal{L}_{ss}$ is the weighted sum between the Structured Similarity Index Measure $(SSIM)$ [44] and $L1$:

$$\mathcal{L}_{ss} = \mathcal{F}(I_t, I_{t'\to t}) = \alpha\frac{1 - SSIM(I_t, I_{t'\to t})}{2} + (1-\alpha)\|I_t - I_{t'\to t}\|_1. \tag{2}$$

Here, α is commonly set to 0.85 [26]. In addition, adjacent locations are encouraged to have similar depth values, using edge-aware smoothness loss on the mean-normalized inverse depth $\hat{\rho}_t$, and $\hat{\rho}_t = \frac{\rho_t}{\bar{\rho}_t}$. The loss function is defined by $\mathcal{L}_{sm}$:

$$\mathcal{L}_{sm} = |\partial_h\hat{\rho}_t|e^{-|\partial_h I_t|} + |\partial_w\hat{\rho}_t|e^{-|\partial_w I_t|}, \tag{3}$$

where ∂_h and ∂_w, respectively, denote the one-dimensional difference quotient of pixels in the image height and width directions. This encourages adjacent pixels to have contiguous depths, causing the network to ignore the depth of edge pixels during training.

3.2. Uncertainty Estimation

The uncertainty of the depth-estimation task comes from two aspects, uncertainty of the model's own ability to learn the data and self-induced uncertainty of the data. The former is due to limitation of the amount of data learned by models, resulting in a large uncertainty in the predicted depth on unlearned scenarios, which can be reduced by expanding the dataset. We still take the model uncertainty into account because we believe the training set is usually collected in similar scenes, e.g., street scenes or indoor scenes.

Although the sample size is large, the sample variety is not complete, and the model still has a probability to encounter unlearned objects. The latter comes from the noise caused by acquisition devices during data collection and the uncertainty caused by complex scenes, such as low brightness, unclear textures, and blurred edges due to relative motion or occlusion of objects. This cannot be reduced by expanding the dataset, but serves as a limiting signal for network to help us solve the edge depth problem associated with depth pulling or occlusion.

3.2.1. Uncertainty in Depth Models

The uncertainty of the model (also known as the epistemic uncertainty) can be calculated by measuring the variance between multiple network instances. One of typical methods is Monte Carlo Dropout [45]. In this way, multiple network instances are sampled from the weight distribution of a single model to estimate the variance. The connections between network layers are randomly discarded with probability. Dropout is turned on during the test, and different network instances of this model can be obtained every time the sample is taken. As follows, the mean $\mu(d)$ and variance $\sigma^2(d)$ can be calculated by performing N forward inferences:

$$\mu(d) = \frac{1}{N}\sum_{i=1}^{N} d_i, \tag{4}$$

$$u_{mod}(d) = \sigma^2(d) = \frac{1}{N}\sum_{i=1}^{N}(d_i - \mu(d))^2. \tag{5}$$

Here, the variance calculated by N sampling is defined as the model uncertainty u_{mod}.

A similar sampling approach can also be used to compute the uncertainty of the depth model, namely bagging (also called bootstrap aggregation) [46]. By training different instance models on random subsets of the training set, we can compute the mean $\mu(d)$ and variance $\sigma^2(d)$ of different depth outputs. This approach requires training N independent sub-networks and passing each network forward in computing variance.

3.2.2. Uncertainty from Depth Distribution

Unlike the model uncertainty presented above, uncertainty introduced by data is unexplained (also called aleatoric). Monocular depth estimation predicts the depth of the same object by multiple viewpoints, which is based on the assumption of grayscale invariance. Clearly, real-world objects must have different grayscale values at their boundaries due to light intensity, surface field curvature, or complex structures, which contradicts the assumption. To encode these uncertainties, we learn its prediction model, whose predicted values are a function of the depth network weights and inputs.

One of popular strategies is to train a network to infer the uncertainty of the depth distribution $p(d^*|I, D)$ of parameters φ by minimizing the negative log-likelihood:

$$\log p(d^*|I, D) = \sum_{i=1}^{N} \log p(d^*|\varphi(I, w_i)), \tag{6}$$

where w is the network weights. The predictive distribution can be modeled as Laplacian in the case of L1 loss computation with respect to d^* [47]. In the self-supervised monocular estimation task, due to the lack of ground truth d^*, the depth data uncertainty u_{dat} can be modeled by photometric matching [48], implying minimization of the following loss function:

$$\mathcal{L}_{uncert} = \frac{\min \mathcal{F}(I_t, I_{t'\to t})}{u_{dat}} + \log u_{dat}. \tag{7}$$

The variance is trained to be in logarithmic form to avoid zero variance. An extra logarithmic term in the formula then restricts the pixels to make infinite predictions.

3.2.3. Multi-Scale Depth Uncertainty

The monocular depth estimation network uses multi-scale feature maps to prevent the gradient from entering the minimum. This idea comes from the work of Lin et al. [49]. The depth uncertainty is modeled on depth weights. We believe that different scale depths need to be constrained by the uncertainty of the corresponding scale. Specifically, the output of the decoder is given additional extra intermediate outputs. Each layer contains one 3×3 convolution, which can be estimated at lower scales. This effectively prevents the training from falling into a local minimum. The existing models mostly use multi-scale for image reconstruction and depth estimation. Our total loss is a combination of the losses of every single scale.

In low-to-medium resolution images, large low-texture regions are not easy to learn. Since monocular depth estimation is predicted based on grayscale values, similar grayscale values between objects or backgrounds cause the network to tend to predict continuous depth. Figure 3 shows a visual example of depth uncertainty estimation. The color of pedestrian clothes in the top left image is similar to the background, and the color of vehicles in the bottom left image is similar to the low textured trees in the background.

These regions are predicted with strong depth uncertainty, which helps constrain the network to make continuous depth predictions on discontinuous boundaries. The multi-scale uncertainty in the figure predicts more uncertainty details (compare the uncertainty intensity of pedestrians and trees in the background), which is helpful for structurally complex objects.

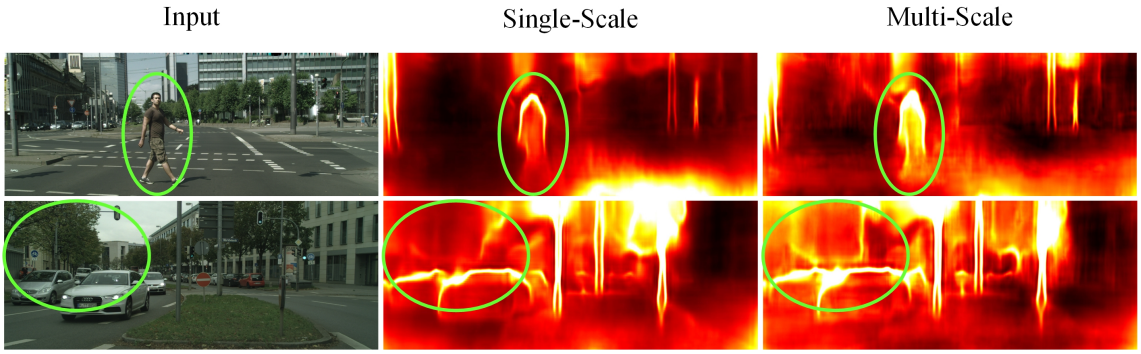

Figure 3. Depth uncertainty estimation cases. Multi-scale uncertainty obtain more depth uncertainty details than single-scale, especially in the geometric blur caused by moving targets (top) or relative motion between targets and tree in the background (bottom).

3.3. Teacher-Student Frameworks

In order to decouple the depth and pose when modeling depth uncertainty, we first train a self-supervised monocular depth estimation network and then use a depth network to mimic it. This teacher–student framework is a type of transfer learning, which can learn smaller models in the same field. Generally, the teacher structure is a complex deep neural network, while the student is lightweight and simple model. Poggi et al. [50] improved the performance of depth estimation results by incorporating this architecture into a network.

We used three teacher–student combinations to investigate the effects on modeling depth uncertainty. The teacher models contained self-supervised monocular depth estimation, dropout layer networks, and bagging strategy models, respectively. The student models mimic the depth distribution from the teacher, which is equivalent to supervised learning, with the supervised signal coming from the output of teacher. Then, we can model the depth uncertainty on the student. Specifically, we train a teacher instance to obtain an output d_T. Assuming the use of $L1$ loss, the depth uncertainty can be modeled as $\mathcal{L}_{TS}$:

$$\mathcal{L}_{TS} = \frac{\|\mu(d_S) - d_T\|_1}{\sigma(d_S)} + \log \sigma(d_S), \tag{8}$$

To avoid zero values in the denominator, let $u_S = \log \sigma(d_S)$, and the loss function can be transformed into:

$$\mathcal{L}_{TS} = \exp(-u_S)\|\mu(d_S) - d_T\|_1 + u_S, \tag{9}$$

The $\mu(d_S)$ and u_S represent the depth mean and variance of student outputs. When the depth of the prediction result cannot imitate the teacher depth well, the value of L1 becomes larger, and the network will increase the uncertainty of pixels in order to minimize the loss.

4. Experiments

In this section, we validate the effectiveness of using uncertainty for self-supervised depth estimation and research different networks of teachers.

4.1. Training Details and Metrics

At first, we describe the relevant details about training and the metrics used in the evaluation model.

4.1.1. Details on the Learning Procedure

In our experiment, the training process uses a monocular sequence, and the open source model *Monodepth*2 [26] is chosen as the baseline model. Most of our protocols follow the setting of [26], the input and output image size is 192 × 640, trained for 20 epochs using Adam, and the batch size is set to 12 (due to memory limitations, we use gradient ac-

cumulation technology). Moreover, the ImageNet [51] pre-training is used at the beginning of the encoder. In addition, we turn on dropout during training and testing, and only use it in the decoder. Regarding our methods, we set the hyperparameter N to 8 and randomly extract 25% of the training set for each bagging network.

4.1.2. Depth Metrics

In order to compare the performance of depth networks, we report the following seven standard performance criteria. They are the Absolute Relative Error ($AbsRel$), Squared Relative Difference ($SqRel$), Root Mean Squared Error ($RMSE$), Root Mean Squared Logarithmic Error ($RMSE$ log), and three accuracy metrics (threshold $\delta < 1.25^k, k \in 1, 2, 3$). For a detailed description, please refer to the reader to [52] for further description of these metrics.

4.1.3. Uncertainty Metrics

In addition to the above metrics to evaluate the performance of our depth estimation model, we measure both Area Under the Sparsification Error ($AUSE$) and Area Under the Random Gain ($AURG$) as measures to evaluate the quality of uncertainty prediction. They come from the sparsity of specified parameters—the so-called sparsification plots [47]. This graph shows similarity between the estimated uncertainty and true error. Generally the descending sorting curve of true error versus ground truth is called oracle sparsification, difference between sparsification and oracle is called $AUSE$.

We used the method described in [50] to measure the area under the sparsification error curve as first indicator. Specifically, for the uncertainty of given error,pixel level is sorted in descending order, we extracted 2% of the pixel subset each time, and the remaining pixel error curve can be drawn. If uncertainty is correctly coded, the error curve will be reduced. Therefore, the lower the value of this indicator, the better. In addition, subtracting the estimated sparsification from random uncertainty is the area $AURG$, the higher the better. We follow [50] and set three error metrics, namely $AbsRel$, $RMSE$, and $\delta \geqslant 1.25$.

4.1.4. Dataset

We used the KITTI dataset [18], which is an outdoor scene image data captured by vehicle with depth estimation sensor while driving on city street, it contains 61 scenes. We used Eigen et al.'s [53] data split and follow Zhou et al.'s [24] of pretreatment to remove the static frame. This outcome 39,810 monocular triplets for training and 4424 for validation. We set the principal point of camera as the center of the image and use the same intrinsic function for all images, taking the average of all focal lengths as the final focal length. During evaluation, we set the maximum depth to 80 m according to standard practice in [23] and use the per-image median ground truth scaling introduced by [24] to report the results.

4.2. Monocular Depth Estimation with Uncertainty

Here, we compare the monocular estimation methods published previously, demonstrate the effectiveness of our approach by depth estimation evaluation. Then, we conduct ablation experiments on multiple model architectures with depth uncertainty for different KITTI benchmarks.

4.2.1. KITTI Eigen Split

We compared the best combined results of our model with those of some state-of-the-art depth estimation models, all of which are monocular methods. The quantitative results are shown in Table 1, all on KITTI using Eigen split. The best results for each evaluation metric are shown in bold. As can be seen, our model is the best result for each metric, which significantly improves the performance compared to the baseline model monodepth2 with the error rate metrics decreased by 4.35%($AbsRel$), 8.97%($SqRel$), 3.64%($RMSE$), and 3.11%($RMSElog$), respectively, and the accuracy rate increased by 0.57%($\delta < 1.25 \uparrow$). The qualitative results are shown in Figure 4.

Table 1. Quantitative results. Comparison of our method with existing methods using the Eigen split on KITTI [54]. The best results in each metric are shown in bold.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Zhou [24]	0.183	1.595	6.709	0.27	0.734	0.902	0.959
Yang [55]	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian [31]	0.163	1.24	6.22	0.25	0.762	0.916	0.968
DF-Net [56]	0.15	1.124	5.507	0.223	0.806	0.933	0.973
Bian [57]	0.137	1.089	5.439	0.217	0.83	0.942	0.975
Chen [58]	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Tosi [59]	0.111	0.867	4.714	0.199	0.864	0.954	0.979
Godard [26]	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Klingner [60]	0.113	0.835	4.693	0.191	0.879	**0.961**	0.981
Poggi [50]	0.112	0.861	4.762	0.189	0.88	**0.961**	**0.982**
Ours	**0.11**	**0.822**	**4.686**	**0.187**	**0.882**	**0.961**	**0.982**

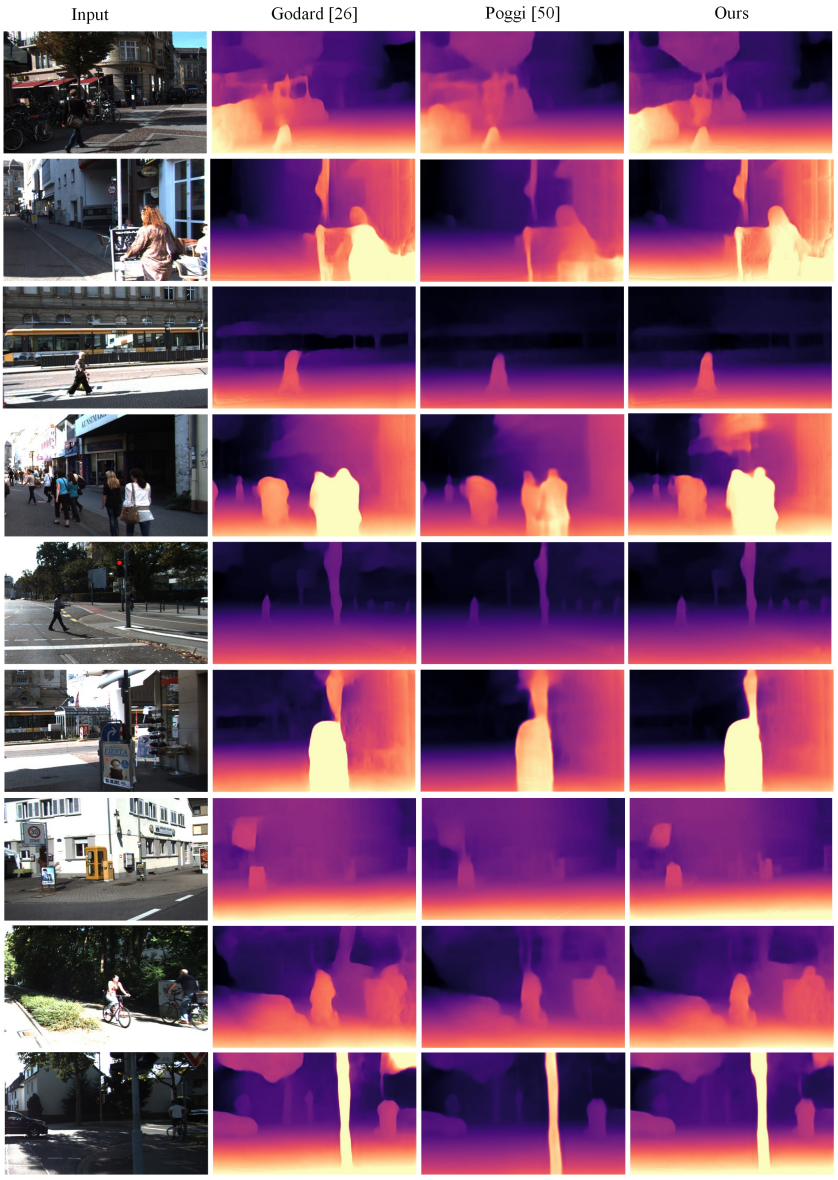

Figure 4. Qualitative results on the KITTI Eigen split. Our model produces depth maps with sharper object boundaries in most cases, which reflects the superior quantitative results in Table 1.

To further elaborate the findings of the experimental results, we selected representative results to explain the effect of uncertainty on depth estimation. Figure 5 contains two typical scenes, occlusion and a low-texture region. Object occlusions (with a front-to-back position relationship) are typically predicted to have the same depth values. In the first image, monodepth2 predicts the backward vehicle and forward vehicle as having similar depth values.

Our method correctly detects the depth of the vehicle in front. The second image is a case of predicting depth of low-texture region with dark background and low saturation of the road sign color. Monodepth2 produced a depth pulling phenomenon at the edges of road sign, and the pixel size of predicted depth region is larger than the original pixel size in input image. In contrast, our method produces better results.

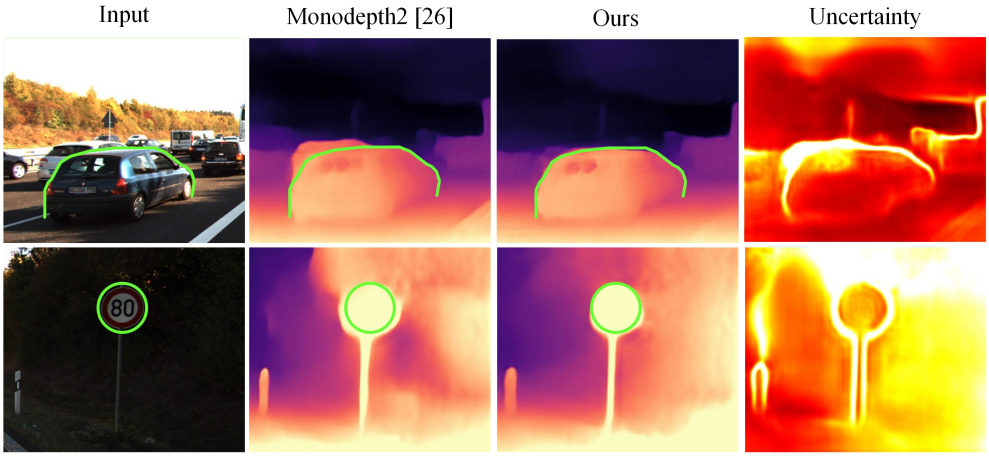

Figure 5. The case of depth estimation with edge optimization. In the case of object occlusion (white and black car in the above figure), monodepth2 predicts both as similar depths, and the silvery green in the figure is the front car outline. Our method correctly predicts the depth of the front car. For the object in the low-texture region (the road sign in the figure below), monodepth2 predicts the depth area pixel size not only larger than the pixel size in the original image but also a little blurred at the depth edge (the depth difference with the background is not obvious). Our method produces better results both in terms of pixel size and smoothness of edges.

4.2.2. Ablation Study

In order to study depth uncertainty and how our strategy contributes to monocular depth estimation training, we implemented two methods of depth uncertainty (aleatoric and epistemic) with different teacher–student strategies for ablation research. Table 2 reports our final results, where BaseT, BagT, and DropT represent the use of baseline, bagging, and dropout networks as the teacher models. In addition, A, E, and S indicate the use of aleatoric, epistemic, and single-scale uncertainty, respectively (multi-scale uncertainty is used by default). The T+S represents that this method uses teacher–student framework, the Alea and Epis are shorthand for aleatoric and epistemic.

The Bagging method that generates model uncertainty obtained the best results for the RMSE metric. While combining Bagging as a teacher model and the two uncertainty methods BagT+AE obtained the best results for the Sq Rel metric. This indicates that the model can learn richer information when trained with the combination of uncertainty and can reduce the error between the predicted and true results.

Our baseline uses Monodepth2, which can already produce high accuracy by itself, and the networks using the baseline as the teacher model both improve in accuracy compared to the baseline. Showing that the teacher–student strategy allows the final network to learn a better distribution of weights. For uncertainty evaluation, the results show that the best combined result is the BaseT+A, while the method using single-scale uncertainty (BaseT+A+S) obtains suboptimal results. For the scale-dependent experiments, we only perform on the strategy with the optimal results (BaseT+A).

We report visualization view of depth uncertainty under different strategies as shown in Figure 6. It is clearly seen that the approach using the teacher–student strategy yields clearer uncertainty results, and the approach using the baseline as the teacher model is superior. These observations are consistent with the quantitative results Tables 2 and 3.

Table 2. Ablation. The results of our different strategies for monocular depth evaluation using Eigen split at KITTI. The best results in each metric are shown in bold.

Method	T+S	Alea	Epis	Multi-Scale	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta <$ 1.25 ↑	$\delta <$ 1.25^2 ↑	$\delta <$ 1.25^3 ↑
Baseline					0.113	0.87	4.756	0.19	0.879	0.96	0.981
Baseline+A		√		√	0.113	0.865	4.76	0.189	0.88	**0.961**	0.982
Bagging			√	√	0.113	0.79	**4.607**	0.189	0.876	**0.961**	**0.983**
Dropout			√	√	0.124	0.835	4.74	0.194	0.852	0.955	**0.983**
BaseT+A+S	√	√			0.111	0.823	4.676	0.188	**0.882**	**0.961**	0.982
BaseT+A	√	√		√	**0.11**	0.822	4.686	**0.187**	**0.882**	**0.961**	0.982
BaseT+AE	√	√	√	√	0.111	0.844	4.706	0.188	0.881	**0.961**	0.982
BagT+AE	√	√	√	√	0.113	**0.786**	4.628	0.189	0.875	**0.961**	**0.983**
DropT+AE	√	√	√	√	0.119	0.787	4.627	0.189	0.861	0.957	**0.983**

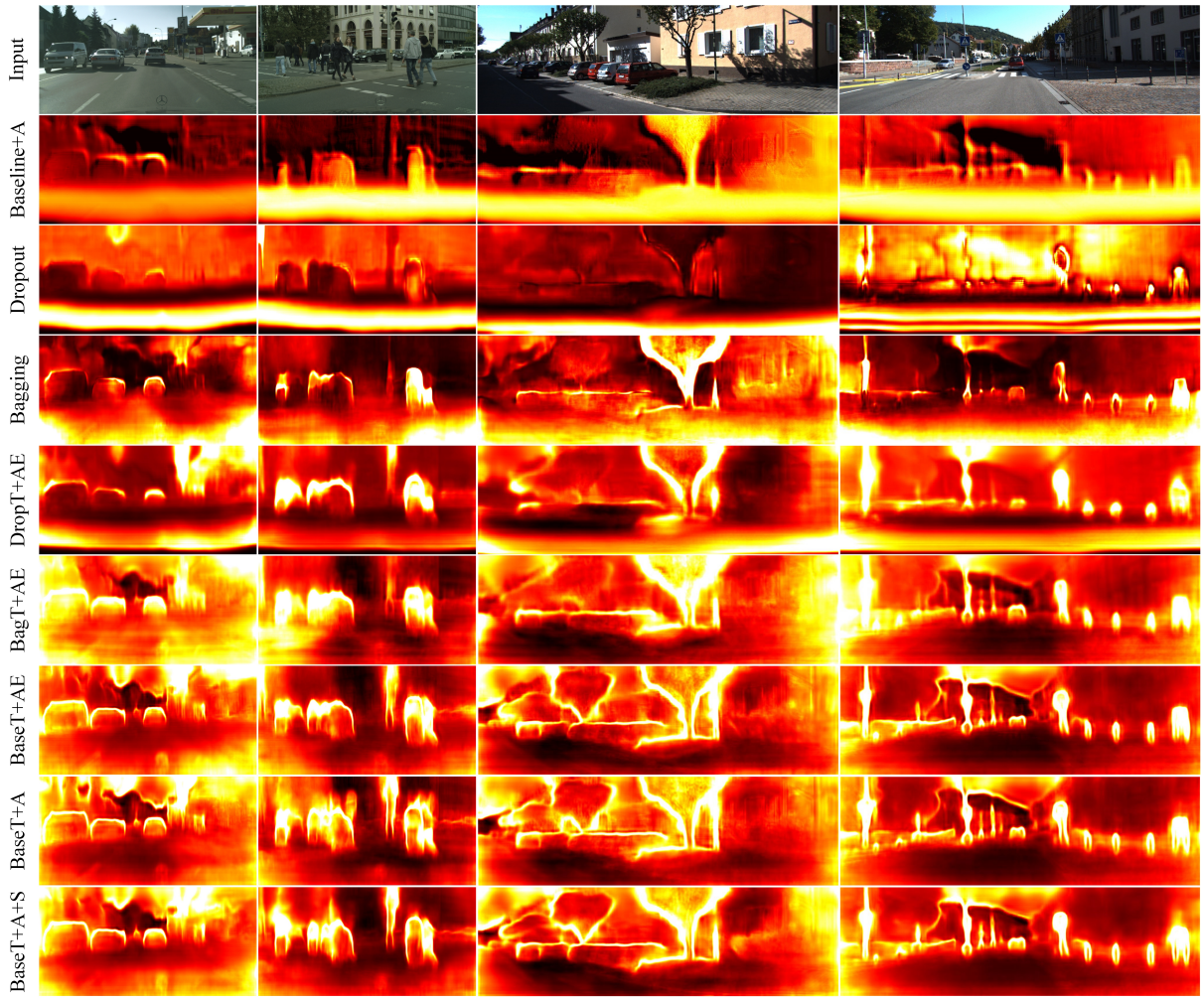

Figure 6. Visualization view of depth uncertainty under different strategies.

Table 3. The results of our different strategies for depth uncertainty evaluation. The best results in each metric are shown in bold.

Method	Abs Rel		RMSE		$\delta \geqslant 1.25$	
	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑
Baseline+A	0.063	0.01	3.676	0.383	0.09	0.02
Bagging	0.063	0.01	3.631	0.275	0.094	0.019
Dropout	0.075	0.005	3.582	0.395	0.122	0.01
BaseT+A+S	**0.034**	**0.039**	2.175	1.833	**0.034**	**0.075**
BaseT+A	**0.034**	**0.039**	**2.168**	**1.871**	0.035	**0.075**
BaseT+AE	0.036	0.038	2.274	1.79	0.037	0.074
BagT+AE	0.037	0.036	2.507	1.428	0.041	0.073
DropT+AE	0.064	0.014	2.488	1.406	0.095	0.03

4.2.3. KITTI New Benchmark

We performed quantitative experiments on the KITTI new benchmark [61], a dataset with 93,000 depth annotated images, all from the original KITTI dataset. This new dataset can be used for training and evaluation of depth upsampling techniques and depth prediction. It is the first dataset with high quality depth annotations for this case. Tables 4 and 5 show the quantitative results of our different strategies on the KITTI new benchmark.

Consistent with the previous analysis, baseline as the teacher model with the addition of multi-scale uncertainty method BaseT+A obtains the best results for combined evaluation metrics. Regarding the baseline as a teacher model approach, the uncertainty evaluation metric has multiple identical metric results. These methods all have the same teacher–student model architecture, only with different uncertainty methods, proving that the main strategy affecting the overall performance of the model is the teacher–student frameworks. In addition, increasing epistemic uncertainty reduces accuracy.

Table 4. Quantitative results of our different strategies for depth evaluation on KITTI new benchmark [61]. The best results in each metric are shown in bold.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Baseline	0.089	0.525	3.867	0.135	0.916	0.983	0.995
Baseline+A	0.088	0.508	3.835	0.133	0.918	0.983	0.995
Bagging	0.091	0.5	3.806	0.136	0.912	0.983	0.995
Dropout	0.099	0.553	4.064	0.147	0.895	0.978	0.995
BaseT+A+S	0.087	0.497	**3.79**	0.132	0.919	**0.984**	**0.996**
BaseT+A	**0.086**	**0.488**	3.792	**0.131**	**0.92**	**0.984**	0.995
BaseT+AE	0.087	0.508	3.813	0.132	0.919	**0.984**	0.995
BagT+AE	0.091	0.501	3.838	0.136	0.912	0.982	0.995
DropT+AE	0.094	0.497	3.862	0.14	0.905	0.981	0.995

Table 5. Quantitative results of our different strategies for uncertainty evaluation on KITTI new benchmark [61]. The best results in each metric are shown in bold.

Method	Abs Rel		RMSE		$\delta \geqslant 1.25$	
	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑
Baseline+A	0.049	0.007	3.016	0.259	0.067	0.01
Bagging	0.051	0.007	3.047	0.184	0.071	0.011
Dropout	0.06	0.004	3.177	0.262	0.09	0.006
BaseT+A+S	**0.029**	**0.027**	1.926	1.333	**0.028**	**0.047**
BaseT+A	**0.029**	**0.027**	**1.916**	**1.362**	0.029	**0.047**
BaseT+AE	0.03	**0.027**	2.019	1.287	0.03	**0.047**
BagT+AE	0.033	0.026	2.35	0.921	0.037	0.045
DropT+AE	0.05	0.01	2.316	0.955	0.069	0.019

5. Conclusions

In this paper, we proposed an improved model for self-supervised monocular depth estimation. To study the influence of the uncertainty of different scales on depth, the results revealed that multi-scale obtained more depth uncertainty details compared with single-scale, which is helpful to solve the problem of depth boundary blur caused by occlusion and low-texture. In addition, we studied the effects of different teacher–student paradigms (using different networks as teachers) on the monocular depth performance.

This architecture not only decoupled the original depth and pose networks but also used a simple model to learn the weight distribution of the teacher network. The student model showed better comprehensive performance, which can improve the depth accuracy of the model. Our results show that the network combining multi-scale depth uncertainty with the optimal teacher–student architecture achieved the best results. The qualitative results show that our proposed network approach can generate high quality depth maps with clarity and details.

Author Contributions: Conceptualization, S.W. and W.L.; methodology, S.W.; software, G.H.; validation, S.W., G.H. and B.L.; formal analysis, W.L.; investigation, S.W.; resources, S.W.; data curation, S.W.; writing—original draft preparation, S.W.; writing—review and editing, S.W., W.L. and B.L.; visualization, S.W.; supervision, S.W., W.L., B.L. and G.H.; project administration, B.L.; funding acquisition, W.L. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was co-supported by the Key *R&D* project of Sichuan Province, China (No.22ZDYF3720), the funding of CAFUC (No.MZ2022KF10), and the funding from Sichuan University (Nos. 2021SCUVS005 and 2020SCUNG205).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hao, Y.; Li, J.; Meng, F.; Zhang, P.; Ciuti, G.; Dario, P.; Huang, Q. Photometric Stereo-Based Depth Map Reconstruction for Monocular Capsule Endoscopy. *Sensors* **2020**, *20*, 5403. [CrossRef] [PubMed]
2. Urban, D.; Caplier, A. Time- and Resource-Efficient Time-to-Collision Forecasting for Indoor Pedestrian Obstacles Avoidance. *J. Imaging* **2021**, *7*, 61. [CrossRef] [PubMed]
3. Hwang, S.J.; Park, S.J.; Kim, G.M.; Baek, J.H. Unsupervised Monocular Depth Estimation for Colonoscope System Using Feedback Network. *Sensors* **2021**, *21*, 2691. [CrossRef] [PubMed]
4. Jia, Q.; Chang, L.; Qiang, B.; Zhang, S.; Xie, W.; Yang, X.; Sun, Y.; Yang, M. Real-Time 3D Reconstruction Method Based on Monocular Vision. *Sensors* **2021**, *21*, 5909. [CrossRef] [PubMed]
5. Liu, X.; Chang, Y.; Li, Z.; Yuan, H. A depth estimation method for edge precision improvement of depth map. In Proceedings of the 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering, Chengdu, China, 12–13 June 2010; Volume 3, pp. 422–425.
6. Chou, H.Y.; Shih, K.T.; Chen, H. Occlusion-and-edge-aware depth estimation from stereo images for synthetic refocusing. In Proceedings of the 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
7. Yang, Z.; Wang, P.; Xu, W.; Zhao, L.; Nevatia, R. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
8. Khan, N.; Kim, M.H.; Tompkin, J. Edge-aware Bidirectional Diffusion for Dense Depth Estimation from Light Fields. *arXiv* **2021**, arXiv:2107.02967.
9. Li, Z.; Zhu, X.; Yu, H.; Zhang, Q.; Jiang, Y. Edge-Aware Monocular Dense Depth Estimation with Morphology. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milano, Italy, 10–15 January 2021; pp. 2935–2942.
10. Palafox, P.R.; Betz, J.; Nobis, F.; Riedl, K.; Lienkamp, M. SemanticDepth: Fusing Semantic Segmentation and Monocular Depth Estimation for Enabling Autonomous Driving in Roads without Lane Lines. *Sensors* **2019**, *19*, 3224. [CrossRef] [PubMed]
11. Kwak, D.; Lee, S. A Novel Method for Estimating Monocular Depth Using Cycle GAN and Segmentation. *Sensors* **2020**, *20*, 2567. [CrossRef]
12. Wang, R.; Zou, J.; Wen, J.Z. SFA-MDEN: Semantic-Feature-Aided Monocular Depth Estimation Network Using Dual Branches. *Sensors* **2021**, *21*, 5476. [CrossRef] [PubMed]
13. Song, X.; Zhao, X.; Hu, H.; Fang, L. Edgestereo: A context integrated residual pyramid network for stereo matching. In Proceedings of the Asian Conference on Computer Vision, Singapore, 20–23 May 2018; pp. 20–35.
14. Xiong, L.; Wen, Y.; Huang, Y.; Zhao, J.; Tian, W. Joint Unsupervised Learning of Depth, Pose, Ground Normal Vector and Ground Segmentation by a Monocular Camera Sensor. *Sensors* **2020**, *20*, 3737. [CrossRef] [PubMed]

15. Richter, S.; Wang, Y.; Beck, J.; Wirges, S.; Stiller, C. Semantic Evidential Grid Mapping Using Monocular and Stereo Cameras. *Sensors* **2021**, *21*, 3380. [CrossRef] [PubMed]
16. Han, L.; Huang, X.; Shi, Z.; Zheng, S. Depth Estimation from Light Field Geometry Using Convolutional Neural Networks. *Sensors* **2021**, *21*, 6061. [CrossRef] [PubMed]
17. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv* **2017**, arXiv:1703.04977.
18. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
19. Karsch, K.; Liu, C.; Kang, S.B. Depth extraction from video using non-parametric sampling. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 775–788.
20. Saxena, A.; Sun, M.; Ng, A.Y. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 824–840. [CrossRef] [PubMed]
21. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 239–248.
22. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2002–2011.
23. 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, Honolulu, HI, USA, 21–27 July 2017; pp. 270–279.
24. 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, Honolulu, HI, USA, 21–27 July 2017; pp. 1851–1858.
25. Garg, R.; Bg, V.K.; Carneiro, G.; Reid, I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 740–756.
26. 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, Seoul, Korea, 27 October–2 November 2019; pp. 3828–3838.
27. Gao, H.; Liu, X.; Qu, M.; Huang, S. PDANet: Self-Supervised Monocular Depth Estimation Using Perceptual and Data Augmentation Consistency. *Appl. Sci.* **2021**, *11*, 5383. [CrossRef]
28. Zhu, Z.; Ma, Y.; Zhao, R.; Liu, E.; Zeng, S.; Yi, J.; Ding, J. Improve the Estimation of Monocular Vision 6-DOF Pose Based on the Fusion of Camera and Laser Rangefinder. *Remote Sens.* **2021**, *13*, 3709. [CrossRef]
29. Fan, C.; Yin, Z.; Xu, F.; Chai, A.; Zhang, F. Joint Soft–Hard Attention for Self-Supervised Monocular Depth Estimation. *Sensors* **2021**, *21*, 6956. [CrossRef]
30. Jung, G.; Won, Y.Y.; Yoon, S.M. Computational Large Field-of-View RGB-D Integral Imaging System. *Sensors* **2021**, *21*, 7407. [CrossRef] [PubMed]
31. Mahjourian, R.; Wicke, M.; Angelova, A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5667–5675.
32. Zhan, H.; Garg, R.; Weerasekera, C.S.; Li, K.; Agarwal, H.; Reid, I. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 340–349.
33. Chen, Z.; Ye, X.; Yang, W.; Xu, Z.; Tan, X.; Zou, Z.; Ding, E.; Zhang, X.; Huang, L. Revealing the Reciprocal Relations Between Self-Supervised Stereo and Monocular Depth Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 15529–15538.
34. Cheng, J.; Wang, Z.; Zhou, H.; Li, L.; Yao, J. DM-SLAM: A Feature-Based SLAM System for Rigid Dynamic Scenes. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 202. [CrossRef]
35. Zhang, X.; Zhang, L.; Lewis, F.L.; Pei, H. Non-Uniform Discretization-based Ordinal Regression for Monocular Depth Estimation of an Indoor Drone. *Electronics* **2020**, *9*, 1767. [CrossRef]
36. Liu, P.; Zhang, Z.; Meng, Z.; Gao, N. Monocular Depth Estimation with Joint Attention Feature Distillation and Wavelet-Based Loss Function. *Sensors* **2021**, *21*, 54. [CrossRef]
37. Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; Gaidon, A. 3d packing for self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2485–2494.
38. Kim, S.; Kim, S.; Min, D.; Sohn, K. Laf-net: Locally adaptive fusion networks for stereo confidence estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 205–214.
39. Mac Aodha, O.; Humayun, A.; Pollefeys, M.; Brostow, G.J. Learning a confidence measure for optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1107–1120. [CrossRef] [PubMed]
40. Wannenwetsch, A.S.; Keuper, M.; Roth, S. Probflow: Joint optical flow and uncertainty estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1173–1182.
41. Pu, C.; Song, R.; Tylecek, R.; Li, N.; Fisher, R.B. SDF-MAN: Semi-Supervised Disparity Fusion with Multi-Scale Adversarial Networks. *Remote Sens.* **2019**, *11*, 487. [CrossRef]

42. Song, C.; Qi, C.; Song, S.; Xiao, F. Unsupervised Monocular Depth Estimation Method Based on Uncertainty Analysis and Retinex Algorithm. *Sensors* **2020**, *20*, 5389. [CrossRef]
43. Walz, S.; Gruber, T.; Ritter, W.; Dietmayer, K. Uncertainty depth estimation with gated images for 3D reconstruction. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Virtual, 20–23 September 2020; pp. 1–8.
44. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
45. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
46. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv* **2016**, arXiv:1612.01474.
47. Ilg, E.; Cicek, O.; Galesso, S.; Klein, A.; Makansi, O.; Hutter, F.; Brox, T. Uncertainty estimates and multi-hypotheses networks for optical flow. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 August 2018; pp. 652–667.
48. Klodt, M.; Vedaldi, A. Supervising the new with the old: learning sfm from sfm. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 August 2018; pp. 698–713.
49. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, Honolulu, HI, USA, 21–27 July 2017; pp. 2117–2125.
50. 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, Seattle, WA, USA, 16–18 June 2020; pp. 3227–3237.
51. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
52. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *arXiv* **2014**, arXiv:1406.2283.
53. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
54. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
55. Yang, Z.; Wang, P.; Xu, W.; Zhao, L.; Nevatia, R. Unsupervised learning of geometry with edge-aware depth-normal consistency. *arXiv* **2017**, arXiv:1711.03665.
56. Zou, Y.; Luo, Z.; Huang, J.B. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 August 2018; pp. 36–53.
57. Bian, J.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.M.; Reid, I. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 35–45.
58. Chen, P.Y.; Liu, A.H.; Liu, Y.C.; Wang, Y.C.F. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019; pp. 2624–2632.
59. Tosi, F.; Aleotti, F.; Poggi, M.; Mattoccia, S. Learning monocular depth estimation infusing traditional stereo knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–19 June 2019; pp. 9799–9809.
60. Klingner, M.; Termöhlen, J.A.; Mikolajczyk, J.; Fingscheidt, T. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 582–600.
61. Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; Geiger, A. Sparsity invariant cnns. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 11–20.