Adaptive confidence thresholding for monocular depth estimation

Hyesong Choi[1]*, Hunsang Lee[2]*, Sunkyung Kim[1], Sunok Kim[3],
Seungryong Kim[4], Kwanghoon Sohn[2], Dongbo Min[1]†
[1]Ewha W. University, [2]Yonsei University, [3]Korea Aerospace University, [4]Korea University

Abstract

Self-supervised monocular depth estimation has become an appealing solution to the lack of ground truth labels, but its reconstruction loss often produces over-smoothed results across object boundaries and is incapable of handling occlusion explicitly. In this paper, we propose a new approach to leverage pseudo *ground truth depth maps of stereo images generated from self-supervised stereo matching methods. The confidence map of the pseudo ground truth depth map is estimated to mitigate performance degeneration by inaccurate pseudo depth maps. To cope with the prediction error of the confidence map itself, we also leverage the threshold network that learns the threshold dynamically conditioned on the pseudo depth maps. The pseudo depth labels filtered out by the* thresholded *confidence map are used to supervise the monocular depth network. Furthermore, we propose the probabilistic framework that refines the monocular depth map with the help of its uncertainty map through the pixel-adaptive convolution (PAC) layer. Experimental results demonstrate superior performance to state-of-the-art monocular depth estimation methods. Lastly, we exhibit that the proposed threshold learning can also be used to improve the performance of existing confidence estimation approaches.*

1. Introduction

Monocular depth estimation, which predicts a dense depth map from a single image, plays an important role in various fields such as scene understanding and autonomous driving. Early works [8, 31, 4] are based on supervised learning in which the performance depends on a huge amount of training data with ground truth depth labels.

Since establishing such a large-scale training data is very costly and labour-intensive, recent approaches rely on the self-supervised learning regime [11, 13, 32, 14, 36]. Instead of using ground truth labels for training the network, they attempt to leverage the self-supervision from a pair of stereo images or monocular video sequences, under the assumption that the geometric structure of a scene can be encoded with the reconstruction loss based on pixel-wise intensity similarities [11]. This loss function seems to be an appealing alternative to the lack of large-scale ground truth labels, but it often leads to blurry results around depth boundaries and does not consider occluded pixels [13].

Instead of relying on the self-supervised reconstruction loss across stereo images, Cho *et al.* [6] attempted to train the monocular depth estimation network through *pseudo* depth labels of the stereo images generated from pre-trained stereo matching network [34]. To mitigate performance degeneration by inaccurate pseudo depth labels, they leverage stereo confidence maps ($\in [0,1]$) indicating the reliability of the pseudo depth labels. The confidence map is truncated with a threshold [6, 46] so that depth values with low confidence are excluded. However, a fixed threshold for all training dataset still has the risk of inaccurate pseudo depth values being used in the network training [6]. The method of [46] attempted to address this issue by learning the threshold with an additional regularization term, but the performance gain is rather limited due to its hard thresholding and the implicit constraint by the regularization term.

To overcome this limitation, we propose a novel architecture that adaptively learns the threshold dynamically conditioned on the pseudo depth map. For a given inaccurate pseudo depth map, the stereo confidence map and its associated threshold are inferred in an end-to-end manner. The confidence map is then thresholded through a differential soft-thresholding operator controlled by the learned threshold. The proposed threshold learning is capable of dealing with the prediction errors of the confidence map more effectively. Note that we leverage the soft-thresholding operator to make the network differentiable. The thresholded confidence map is then used together with

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-00056) and the Mid-Career Researcher Program through the NRF of Korea (NRF-2021R1A2C2011624). S. Kim[4] was supported in part by the MSIT under the ICT Creative Consilience Program (IITP-2021-2020-0-01819).

* Equal contribution. † Corresponding author.

the pseudo depth labels for training the monocular depth estimation network. Additionally, we propose to enhance the monocular depth map in a probabilistic inference framework. Unreliable parts of the monocular depth map are identified using the uncertainty map, and these are refined through the pixel-adaptive convolution (PAC) layer [45]. Experimental results validate that the monocular depth accuracy is significantly improved by leveraging the proposed threshold learning and probabilistic depth refinement modules.

Interestingly, the threshold learning can also be beneficial to improve the performance of existing stereo confidence estimation approaches [38, 25]. The confidence map obtained from the existing approaches [38, 25] is refined through the soft-thresholding function controlled by the learned threshold. As shown in Fig. 2, the soft-thresholding function attenuates low confidence values that are less than the learned threshold τ to become as close as 0 while amplifying high confidence values to converge to 1. We validate through experiments that this process improves the prediction accuracy of the existing confidence estimation approaches. To sum up, our contributions are as follows.

- We propose a novel framework of monocular depth estimation using pseudo depth labels generated from self-supervised stereo matching methods.
- We introduce the threshold network that adaptively learns the threshold of the confidence map for better predicting the reliability of the inaccurate pseudo depth labels.
- The monocular depth map is further refined through the probabilistic refinement module based on the PAC layer.
- It is shown that the threshold network can also be used to enhance the prediction accuracy of existing confidence estimation approaches.

2. Related Work

Monocular depth estimation. Eigen *et al.* [8] initiated the monocular depth estimation through deep network that regresses a depth map with ground-truth depth information, inspiring numerous approaches based on multi-scale images [31], up-projection technique [29], motion parallax [50], ordinal regression [9], and semantic divide-and-conquer [51]. Despite remarkable performance over classical handcrafted approaches, they rely on abundant and high-quality ground-truth depth maps, which is costly to obtain.

To overcome this limitation, self-supervised learning has been introduced by leveraging other forms of supervision from stereo images and video sequences instead of ground truth depth maps. Garg *et al.* [11] used the stereo photometric reprojection. Godard *et al.* [13] further used the left-right consistency between stereo images. Zhou *et al.* [57] proposed to leverage multi-view synthesis procedure, and this idea was extended using the feature-based warping loss in [55]. To take advantages of both supervised and self-supervised learning methods, semi-supervised learning methods have also been presented. Kuznietsov *et al.* [28] directly combined supervised and unsupervised loss terms. Ji *et al.* [22] utilizes an image-depth pair discriminator with a small amount of labeled dataset, alleviating the reliance on supervision. Recently, Gonzalebello *et al.* [15] proposed mirrored exponential disparity (MED) probability volumes to handle occluded areas.

The most related to our work is the methods of Guo *et al.* [18], Cho *et al.* [6], and Tonioni *et al.* [46] in which a stereo matching knowledge is distilled to train a monocular depth network. Since the disparity map estimated by stereo matching inherently contain unreliable ones, they used stereo confidence to build a pseudo-ground-truth disparity map by thresholding the confidence. Guo *et al.* [18] used a handcrafted occlusion map sensitive to outliers. Cho *et al.* [6] used a fixed threshold empirically, but it is ineffective to use the same threshold for all images. Unlike this, Tonioni *et al.* [46] tried to learn the threshold by using an additional regularization term that allows it to be between 0 and 1, but it is also difficult to learn the appropriate threshold with the implicit constraint by the regularization term. In our method, effective threshold learning is the main contribution.

Stereo confidence estimation. In parallel with the development of predicting depth from images, stereo confidence estimation has also been actively studied. Machine learning approaches [35, 44, 26] relying on shallow classifier, e.g., random tree [1], enable one to classify correct and incorrect pixels. Recently, deep convolutional neural network (CNN)-based approaches have become a mainstream. Various methods have been proposed that use the single- or bi-modal input, e.g., disparity [38], left and right disparities [41], 3D matching cost [42], 3D matching cost and disparity [27], and disparity and color image [49, 10]. Kim *et al.* [25] proposed to make full use of the tri-modal input in conjunction with locally adaptive attention and scale networks, achieving state-of-the-art prediction accuracy. All of these techniques require ground truth depth maps and have been used to refine a depth (or disparity) map with a fixed threshold which is set empirically. Poggi et al. [37] introduced a method for learning self-supervised confidence measure with various criterions.

3. Proposed Method

Unlike recent self-supervised monocular depth estimation approaches [11, 13, 32, 14, 36], we leverage the *pseudo*

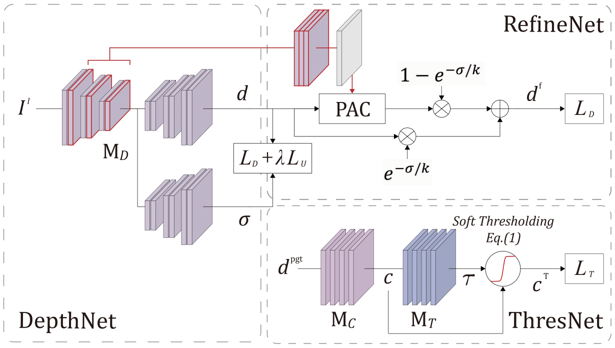


Figure 1. The proposed architecture consisting of ThresNet, DepthNet, and RefineNet. Given a pair of stereo images, the pseudo ground truth depth map d^{pgt} is precomputed using a self-supervised stereo matching network. The proposed model training begins with d^{pgt} by computing its confidence map c and the threshold τ through the ThresNet. The thresholded confidence map c^{T} is obtained using the soft-thresholding function. The DepthNet that infers the monocular depth map d and uncertainty map σ is trained by minimizing an objective defined using d^{pgt} filtered out by c^{T}. The monocular depth map d is finally refined through the probabilistic refinement module based on the pixel-adaptive convolution (PAC) layer in the RefineNet.

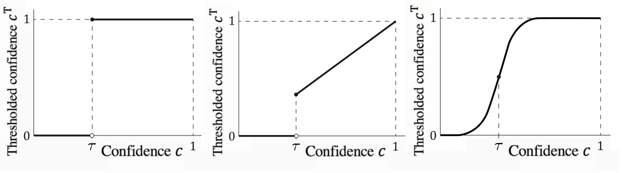
depth labels from a pair of stereo images as supervision for monocular depth estimation. Fig. 1 shows the overall procedure of the proposed method consisting of three networks, including DepthNet, RefineNet, and ThresNet.

The proposed model training begins with the pseudo depth labels d^{pgt} precomputed using the self-supervised stereo matching method [53]. Note that among various options provided in [53] for data synthesis, we adopted 'Monodepth2' [14] which is self-supervised monocular depth network. Its confidence map c is estimated by the confidence estimation module M_C, aiming at preventing the abuse of erroneous depth values in training the monocular depth network. To take into account the prediction errors of the confidence map itself, we further learn the threshold τ, truncating the confidence map, adaptively through the threshold module M_T. The thresholded confidence map c^{T} is obtained via the soft-thresholding by the learned threshold τ. This operation encourages to trust the pixel with a higher confidence value than a specific τ value. The DepthNet is trained by minimizing an objective defined using the pseudo depth labels d^{pgt} filtered out by the thresholded confidence map c^{T}. Finally, our method refines the monocular depth map d through the probabilistic refinement module based on the PAC layer [45] in the RefineNet.

3.1. Network Architecture

3.1.1 ThresNet

The ThresNet predicts the confidence map of the inaccurate pseudo depth label and its threshold in an adaptive manner and then generates the thresholded confidence map via the soft-thresholding function. For the confidence estimation network M_C, we adopted the CCNN [38] thanks to its simplicity, but more sophisticate models [38, 25, 49] can also be utilized as a backbone. The threshold network M_T consists of four convolutional layers, followed by global average pooling and 1×1 convolution.



(a) (b) (c)

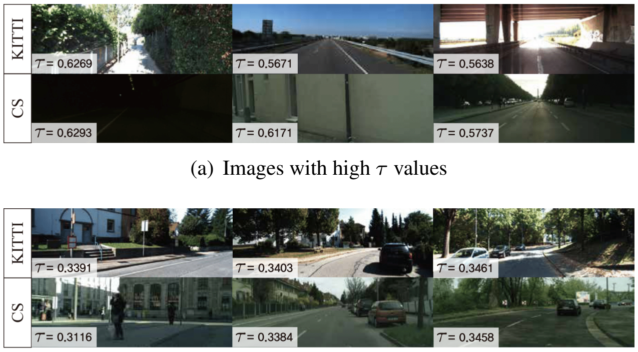
Figure 2. Comparison of confidence thresholding operator: (a) hard-thresholding used in [6], (b) hard-thresolding function used in [46], and (c) our soft-thresholding function in (1). The learned threshold is used in (b) and (c), while the threshold is fixed in (a) for all training images.

The estimated confidence map c is modulated by the threshold τ, such that a depth value with a higher confidence value than a specific τ value assumes to be trustworthy. A key issue is how to set accordingly τ which needs to vary depending on images. This threshold τ should be set low in the image where depth inference is easy while being set high in the opposite case (see Fig. 3). We approximate the thresholding operation with a smooth, differentiable function. The thresholded confidence map c^{T} is computed using the differentiable soft-thresholding function as follows:

$$c_p^{\text{T}}(\tau) = \frac{1}{1 + e^{-\varepsilon \cdot (c_p - \tau)}}, \tag{1}$$

where p represents a pixel. The slope of the thresholded confidence map c^{T} is adjusted by a hyperparameter ε, which is a positive constant. Too large ε changes the soft-thresholding function too rapidly (e.g. $\varepsilon = 90$), often making it non-differentiable. We set $\varepsilon = 10$ in experiments. The pixel-varying confidence map is transformed with the per-image threshold τ. We also investigated a pixel-varying threshold map τ_p, but its performance gain was negligible.

Fig. 2 compares the confidence thresholding functions. In Fig. 2 (a), the confidence threshold τ is fixed with a predefined value for all training images without considering image characteristics, often causing inaccurate pseudo depth values to be used during training. In Fig. 2 (b), it is learned using an additional regularization term [46], but its performance gain on the monocular depth estimation is rather limited, as reported in the original paper [46]. The proposed differential soft-thresholding function, controlled by the threshold τ dynamically conditioned on the pseudo depth map, leads to superior performance on the monocular depth estimation, when the threshold loss L_T is used



(a) Images with high τ values

(b) Images with low τ values

Figure 3. Examples of learned threshold τ by our threshold learning. CS indicates the Cityscapes dataset.

together. The ablation study of the confidence thresholding operators is provided in experiments.

Fig. 3 presents the estimation results of the ThresNet for the KITTI and Cityscape datasets [7]. The threshold τ becomes higher in images where stereo matching do not work well, and vice versa. This indicates that the ThresNet is beneficial to improving the monocular depth network by excluding unreliable pseudo depth values effectively.

3.1.2 DepthNet and RefineNet

The DepthNet and RefineNet infer and refine the monocular depth map by leveraging the pseudo depth labels, masked out by the thresholded confidence map, as supervision. The DepthNet is based on the encoder-decoder architecture [39], in which an encoder takes an image and two decoders estimate the monocular depth map d and its uncertainty map σ. The uncertainty map, indicating the variance of the predicted monocular depth map, becomes higher when the prediction is unreliable, and vice versa. The encoder network consists of the first 13 convolution layers of the VGG network [43], and the decoder is symmetrical with the encoder. The uncertainty map σ is used to refine the monocular depth map in the subsequent RefineNet.

We first upsample L feature maps (here $L = 4$) from the encoder of the DepthNet to an original resolution and concatenate them. The concatenated features are then fused by passing through 1×1 convolution, generating a guidance feature g. The estimated monocular depth map d is finally fed into the PAC layer [45] with the guidance of the feature map g. Unlike the original PAC module that directly infers refined results, we leverage the residual connection that takes into account the uncertainty map σ for predicting the refined monocular depth map d^f such that

$$d^f = e^{-\sigma/k} \cdot d + (1 - e^{-\sigma/k})d' \tag{2}$$

where d' indicates the output of the PAC layer. k is a hyperparameter to control the refinement through the PAC layer, and it was set to 1.

It should be noted that though some monocular depth estimation approaches [36, 2] have attempted to measure the uncertainty of the monocular depth estimation through deep network, our method proposes to infer the uncertainty map and use it for a subsequent refinement module. This framework can also be extended into various pixel-level labeling tasks based on the uncertainty prediction.

3.2. Loss Functions

3.2.1 Thresholding loss

The ThresNet with confidence and threshold networks can be trained in a supervised manner [38] or a self-supervised manner [37]. For the supervised training, we propose to use the sparse ground truth depth data provided by public benchmarks. For instance, we can leverage extremely sparse LiDAR depth maps of 3% density provided with a set of stereo image pairs in the KITTI dataset. The ground truth of the thresholded confidence map is generated using the sparse ground truth depth data like existing confidence estimation approaches [25] and this is used to train the ThresNet using a cross-entropy loss L_T. More details on the ground truth confidence map are provided in the supplementary material. Alternatively, the ThresNet can be trained in the self-supervised manner without using the LiDAR depth maps. Following [37], we generate the pseudo ground truth of the thresholded confidence map according to various criterions (e.g., reprojection error, disparity agreement). The loss L_T for the self-supervised training is defined as a multi-modal binary cross entropy loss of [37]. In Table 1, we compare the monocular depth accuracy when using the supervised and self-supervised ThresNets, and found the accuracy is almost similar.

In [46], the threshold is also learned to exclude depth values with low confidences when training their network. It was reported that when using the depth regression loss only, the threshold τ would converge to 1 for masking out all pixels [46]. Thus, they propose to include an additional regularization loss, $-\log(1 - \tau)$, that prevents the threshold τ from approaching 1. Though this term allows τ to be between 0 and 1, it does not guarantee to yield accurate prediction results of the threshold τ. Contrastingly, our method attempts to learn the threshold τ with the soft-thresholding function and the explicit supervision. We will verify the effectiveness of our threshold learning approach in the ablation study of Table 4.

3.2.2 Depth regression loss

A monocular depth map from the DepthNet is leveraged to compute a confidence-guided depth regression loss L_D

assisted by the thresholded confidence map c^{T} as follows:

$$L_D = \frac{1}{Z}\sum_{p\in\Omega} c_p^{\mathrm{T}}(\tau)\cdot|d_p - d_p^{\mathrm{pgt}}|, \tag{3}$$

where d and d^{pgt} indicate the predicted depth map and pseudo ground truth depth map, respectively. Ω represents a set of all pixels. The loss L_D is normalized with $Z = \sum c_p^{\mathrm{T}}(\tau)$.

Additionally, we leverage the negative log-likelihood minimization to infer the uncertainty of the network output. The predictive distribution of the network output d can be modelled as the Laplacian likelihood [24, 21, 23] as follows:

$$L_U = \frac{1}{|\Omega|}\sum_{p\in\Omega}\left(\frac{|d_p - d_p^{\mathrm{pgt}}|}{\sigma_p} + \log\sigma_p\right), \tag{4}$$

where the variance σ represents the uncertainty map of the predicted depth map. The logarithmic term $\log\sigma$ prevents σ from approaching to infinity [24]. We combine two losses L_D, taking into account the reliability of the pseudo ground truth depth map d^{pgt}, and L_U predicting the uncertainty of the predicted depth map d, such that

$$L = L_D + \lambda L_U, \tag{5}$$

where λ represents hyperparameter that balances two losses which is experimentally determined to 10^{-3}. This enables for modeling the uncertainty of the monocular depth estimation network while considering the confidence of the pseudo depth label. As shown in Fig. 1, the DepthNet that infers both the monocular depth map and uncertainty map is trained with L in (5), while the RefineNet leverages L_D in (3) as it predicts the final monocular depth map only.

3.3. Training Details

In our work, the DepthNet and RefineNet are trained simultaneously by minimizing L and L_D, while the ThresNet consisting of confidence and threshold networks is trained solely by minimizing L_T, similar to existing confidence estimation approaches [38, 35, 44, 25]. Though the whole networks can be trained end-to-end, we found through experiments that the performance gain over the separate training is relatively marginal.

It has been reported in literature [38, 49] that the confidence network trained with one dataset exhibits a good generalization capability for another dataset. In a similar context, our confidence and threshold networks trained with the KITTI dataset show satisfactory generalization capability for different datasets. Taking these into account, we transfer the knowledge learned from one dataset to another. To be specific, when only stereo image pairs are available for training (e.g. Cityscape dataset), the DepthNet and RefineNet are trained via the minimization of L and L_D, with the ThresNet being frozen with the parameters trained with the KITTI dataset. As shown in Fig. 3, the ThresNet trained with the KITTI dataset produces appropriate thresholds for both the KITTI and Cityscape datasets.

4. Extension to Confidence Estimation

The soft-thresholding attenuates low confidence values that are less than τ to become as close as 0 while amplifying high confidence values to converge to 1. It reduces the number of ambiguous pixels to determine the reliability, for which a confidence value is far from 0 or 1. We discuss how the soft-thresholding based on the threshold network can improve the prediction accuracy of existing confidence estimation approaches [38, 25]. In the ThresNet of Fig. 1, the confidence network can be replaced with the existing confidence estimation approaches. One difference is that the loss L_T (cross-entropy loss) is measured on the disparity domain, considering that the existing confidence estimation approaches are trained on the disparity domain. This formulation is model-agnostic, and any kind of existing confidence estimation approaches can be used in a plug-and-play fashion.

5. Experimental results

5.1. Implementation details

The proposed method was implemented in PyTorch framework and run Titan RTX GPU. We trained the whole networks on the learning rate of 10^{-4} and batches of 32 images resized to 192 × 480 for 30 epochs. We trained the proposed monocular depth estimation network consisting of DepthNet and RefineNet on the standard 20k stereo images provided in the KITTI dataset. We evaluate our methods on following five metrics 'RMSE', 'RMSE log', 'Abs Rel', 'Sq. Rel', and 'Accuracy', proposed in Eigen *et al.* [8].

5.2. Evaluation on monocular depth estimation

5.2.1 KITTI

In Table 1, we evaluated the monocular depth estimation performance quantitatively on the KITTI Eigen Split [8] dataset with setting maximum depth to 80 meters with Gargs crop [11]. A comprehensive evaluation was conducted with Monodepth [13], Uncertainty [36], MonoResMatch [48], Monodepth2 [14], DepthHint [52], PackNet-SfM [17], and Insta-DM [30]. For the training data, 'S' indicates using stereo images for self-supervised monocular depth estimation. 'M' represents a monocular video sequence. The evaluation of the proposed method is twofold; 'Ours (D)' trained with only the DepthNet using L_D in (3) without refining the depth map, and 'Ours (D+R)' trained with the DepthNet and RefineNet.

Table 1. Quantitative evaluation for depth estimation with existing methods on KITTI Eigen Split [8] dataset. Numbers in bold and underlined represent 1^{st} and 2^{nd} ranking, respectively. 'Ours†' is obtained using the self-supervised ThresNet [37], while 'Ours' indicates the results obtained using the supervised ThresNet.

				Lower is better				Accuracy: higher is better		
Method	Data	#p	time	Abs Rel	Sqr Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth [13]	S	56M	9.4ms	0.138	1.186	5.650	0.234	0.813	0.930	0.969
Monodepth2 [14]	S	14M	2.9ms	0.108	0.842	4.891	0.207	0.866	0.949	0.976
Uncertainty [36]	S	14M	3.6ms	0.107	0.811	4.796	0.200	0.866	0.952	0.978
MonoResMatch [48]	S	41M	8.3ms	0.111	0.867	4.714	0.199	0.864	0.954	0.979
DepthHint [52]	S	33M	6.6ms	0.102	0.762	4.602	0.189	0.880	0.960	0.981
PackNet-SfM [17]	M	122M	9.5ms	0.111	0.785	4.601	0.189	0.878	0.960	**0.982**
Insta-DM [30]	M	14M	3.0ms	0.112	0.777	4.772	0.191	0.872	0.959	**0.982**
Ours (D)	S	28M	6.8ms	0.099	0.652	4.266	0.187	0.883	0.960	0.981
Ours (D+R)	S	42M	8.2ms	**0.096**	0.627	**4.201**	0.186	**0.885**	**0.961**	**0.982**
Ours† (D)	S	28M	6.8ms	0.100	0.644	4.251	0.187	0.882	0.960	0.981
Ours† (D+R)	S	42M	8.2ms	0.098	**0.621**	4.215	**0.185**	**0.885**	**0.961**	**0.982**

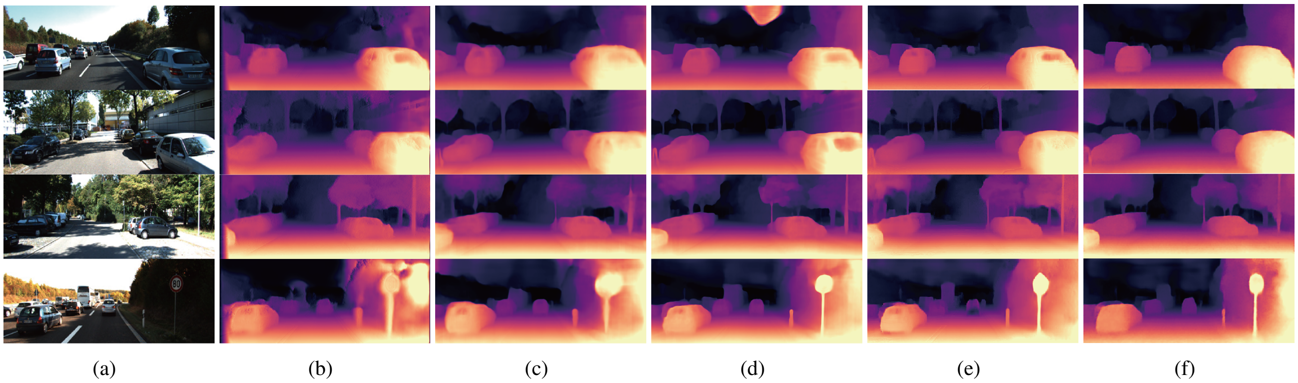


Figure 4. Qualitative evaluation with existing monocular depth estimation methods on the Eigen Split [8] of KITTI dataset: (a) input image, (b) Monodepth [13], (c) Monodepth2 [14], (d) DepthHint [52], (e) PackNet-SfM [17] and (f) Ours (D+R).

As reported in Table 1, although 'Ours (D)' leverages a rather simple encoder-decoder architecture, it achieves the superior performance over existing methods, demonstrating the effectiveness of the proposed threshold learning approach. In 'Ours (D+R)', the monocular depth accuracy was further improved by making use of the probabilistic refinement module based on the uncertainty map and the PAC layer in the RefineNet. We also evaluated the number of parameters used and an inference time, noted as '#p' and 'time', respectively. Our method uses relatively smaller or similar number of parameters compared to other methods. 'Ours†' is obtained using the self-supervised ThresNet [37], while 'Ours' indicates the results obtained using the supervised ThresNet. We found that their monocular depth accuracy is almost similar. The following results including ablation study were conducted with the supervised ThresNet. Fig. 4 shows the qualitative comparison with existing methods on the KITTI Eigen Split [8] dataset. It was shown that the proposed method recovers complete instances better while preserving fine object boundaries.

5.3. Cityscapes

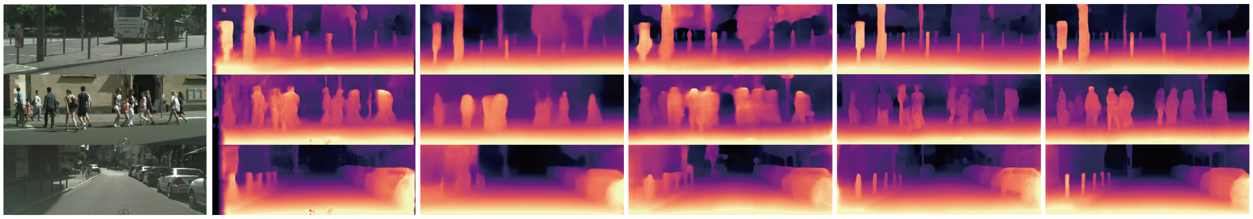
We also evaluated the performance of the proposed method on the Cityscapes dataset. The Cityscapes dataset provides only stereo images without the ground truth, and thus the ThresNet trained with the KITTI dataset was used to infer the threshold. Table 2 shows the quantitative evaluation on Cityscapes dataset [7] with the DepthNet and RefineNet fine-tuned on the Cityscapes dataset, while the ThresNet is freezed. We compared our results with Monodepth2 [14], DepthHint [52] and PackNet-SfM [17]. We set maximum depth to 80 meters with the per-image median scaling approach [57]. We used the SGM depth [19] as ground truth for the evaluation. The outstanding performance of our method supports the claim that the ThresNet trained with the KITTI dataset shows a satisfactory generalization capability for different datasets.

5.4. Evaluation on uncertainty estimation

To evaluate the performance of the uncertainty measure, we use sparsification plots used in [21]. 'AUSE' denotes the Area Under the Sparsification Error which quantifies how

Table 2. Quantitative evaluation for monocular depth estimation results on Cityscapes validation dataset with fine-tuning on Cityscapes training dataset. Numbers in bold and underlined represent 1^{st} and 2^{nd} ranking, respectively.

		Lower is better				Accuracy: higher is better		
Method	Data	Abs Rel	Sqr Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [14]	S	0.124	1.287	7.293	0.223	0.785	0.947	0.981
Struct2Depth [5]	M	0.145	1.737	7.280	0.205	0.813	0.942	0.978
DepthHint [52]	S	0.128	1.268	7.156	0.218	0.812	0.949	0.982
Gordon [16]	M	0.127	1.330	6.960	**0.195**	0.830	0.947	0.981
Ours (D)	S	0.123	1.141	6.735	0.204	0.844	0.962	0.985
Ours (D+R)	S	**0.115**	**1.125**	**6.584**	**0.195**	**0.857**	**0.963**	**0.986**



(a) (b) (c) (d) (e) (f)

Figure 5. Qualitative evaluation for depth estimation with existing methods on Cityscapes validation dataset: (a) Input image, (b) Monodepth [13], (c) MonoResMatch [48], (d) DepthHint [52], (e) PackNet-SfM [17] (f) Ours (D+R).

close the estimate is to the oracle uncertainty, which is lower the better. 'AURG' denotes the Area Under the Random Gain, which indicates how better it is compared to the case without modelling, which is higher the better. In Table 3, the uncertainty measure estimated by the proposed method was compared with 'Monodepth2-Log' of Poggi *et al.* [36], trained under the same setup as our experiments.

5.5. Ablation study

Threshold learning In Table 4, we conducted the ablation study to validate the performance improvement by the proposed threshold learning over existing threholding approaches [6, 46]. For a fair comparison, we obtained the results using the monocular depth network trained with only the DepthNet (without the uncertainty decoder), when varing thresholding functions. 'Baseline' represents the results obtained using the confidence map without thresholding. The results of [6] were obtained using the hard thresholding of Fig. 2 (a) with $\tau = 0.3$, following the setup of [6]. The performance of [6, 46] was almost similar, though the method in [46] learned the threshold τ with the thresholding function of Fig. 2 (b). We found that the regularization loss $-log(1-\tau)$ [46], used to prevent the threshold τ from approaching 1, does not generate a meaningful variant for the learned threshold due to the lack of explicit supervision for the threshold learning. 'Tonioni et al. [46] + L_T' were obtained using the thresholding function of Fig. 2 (b) and our loss L_T. The performance gain over 'Tonioni et al. [46]' demonstrates the effectiveness of L_T. 'Ours (D)' achieves a substantial performance gain, demonstrating the effectiveness of the proposed threshold learning with L_T.

Table 3. Quantitative evaluation for uncertainty estimation with the state-of-the-art method on KITTI Eigen Split [8] dataset. Numbers in bold indicate the better performance.

	Abs Rel		RMSE		$\delta \geq 1.25$	
Method	AUSE	AURG	AUSE	AURG	AUSE	AURG
Uncertainty [36]	0.022	0.036	0.938	2.402	**0.018**	0.061
Ours	**0.021**	**0.048**	**0.765**	**2.881**	0.025	**0.080**

Adaptability We also validated the effectiveness of our method when applied to different network architectures, e.g., PackNet [17]. Table 5 shows quantitative evaluation results when using our confidence threshold learning and probabilistic refinement on the PackNet architecture. 'PackNet (D)' represents the results obtained using the DepthNet only, whereas 'PackNet (D+R)' is the results using both DepthNet and RefineNet. We observed that our framework also improves the monocular depth accuracy for the PackNet architecture.

Uncertainty To evaluate the importance of using the estimated uncertainty in the RefineNet, we compared the results obtained using the proposed depth refinement of (2) and the simple depth refinement ($d^f = d + d'$) without σ in Table 6, demonstrating the effectiveness of the depth refinement based on the uncertainty map.

Pseudo ground truth depth labels So far, all experiments were conducted with the self-supervised pseudo depth maps obtained using [53]. To validate the adaptability of our framework with respect to the pseudo depth labels, we performed additional experiments with the pseudo ground truth depth maps generated by [47], which are trained with synthetic data and fine-tuned with an self-supervised recon-

Table 4. Comparison with other thresholding methods on the KITTI Eigen Split [8] dataset. We evaluated the performance with the supervised ThresNet, and L_T is a cross-entropy loss.

	τ	Abs	RMSE	$\delta < 1.25$
Baseline	×	0.108	4.552	0.869
Cho et al. [6]	fixed	0.102	4.441	0.874
Tonioni et al. [46]	learned	0.101	4.453	0.878
Tonioni et al. [46] + L_T	learned	0.100	4.390	0.879
Ours (D)	learned	0.099	4.266	0.883

Table 5. Quantitative evaluation of the results obtained by applying our threshold learning and probabilistic refinement to the PackNet-SfM architecture [17] on the KITTI Eigen Split [8] dataset.

	Abs	RMSE	$\delta < 1.25$
PackNet-SfM [17]	0.111	4.601	0.878
PackNet-SfM (D)	0.105	4.258	0.880
PackNet-SfM (D+R)	0.100	4.225	0.883

Table 6. Ablation study of the uncertainty map.

	Abs	Sqr	RMSE	RMSE log	$\delta < 1.25$
(2) w/o σ	0.099	0.661	4.298	0.188	0.881
(2)	0.096	0.627	4.201	0.186	0.885

Table 7. Quantitative evaluation when using pseudo depth labels generated by [47] on the KITTI Eigen Split [8] dataset.

	Abs	Sqr	RMSE	RMSE log	$\delta < 1.25$
Ours (D)	0.102	0.728	4.281	0.189	0.880
Ours (D+R)	0.100	0.711	4.230	0.187	0.883

struction loss with meta-learning framework. Table 7 shows that the monocular depth accuracy is still superior to state-of-the-arts monocular depth estimation approaches.

5.6. Confidence evaluation

We validated the effectiveness of the proposed threshold learning in terms of confidence prediction accuracy by applying it to two confidence estimation approaches, CCNN [38] and LAFNet [25]. We trained the two confidence estimation methods with 20 out of 194 images provided in the KITTI 2012 training dataset [12]. Note that the confidence estimation approaches [38, 25] are evaluated by training them in a supervised manner. The area under the curve (AUC) [20], which is a common metric for confidence estimation approaches, was used for an objective evaluation. Refer to the supplementary material for details on measuring AUC and optimal AUC and more results. Following confidence estimation literatures, input disparity maps used for predicting the confidence maps were obtained using two popular stereo algorithms, 'Census-SGM' [19] and 'MC-CNN' [54].

Table 8 shows objective evaluation results for 200 images of KITTI 2015 dataset [33] and 15 images of Middlebury v3 dataset [40]. 'w/τ' denotes our results using the soft-thresholding technique. LAFNet* denotes the LAFNet [25] in which 3D cost volume is not used as an input. Our approach consistently outperforms the original confidence estimation methods, demonstrating the effectiveness of the proposed threshold learning. Fig. 6 compares the confidence maps visually. While the original confidence maps contain ambiguous values for which it is difficult to determine whether the depth label is correct, our thresholded confidence map yields more distinct values that are close to 0 or 1. Such a binarization enables the estimated confidence to have similar distribution to ground truth confidence, thus improving a discriminative power.

Table 8. Performance evaluation of confidence estimation for KITTI 2015 and Middlebury v3 datasets with two popular stereo matching methods C-SGM (Census-SGM) [19] and MC-CNN [54]. AUC values are reported and the lower is the better.

	KITTI 2015	MID 2014
	C-SGM / MC-CNN	C-SGM / MC-CNN
CCNN	1.868 / 3.190	9.486 / 9.787
CCNN w/τ	1.720 / 3.525	8.314 / 9.497
LAFNet*	1.797 / 3.051	8.895 / 9.660
LAFNet* w/τ	1.687 / 3.037	8.988 / 9.456
LAFNet	1.680 / 2.903	8.884 / 9.305
LAFNet w/τ	**1.587 / 2.885**	**8.680 / 8.622**
optimal	0.737 / 2.761	3.887 / 4.985

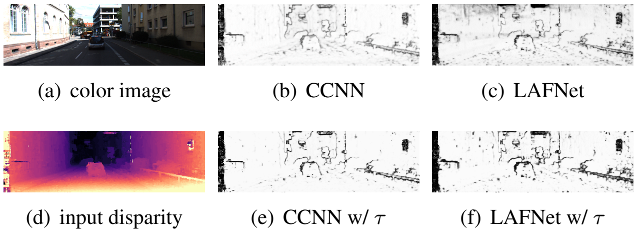

(a) color image
(b) CCNN
(c) LAFNet
(d) input disparity
(e) CCNN w/ τ
(f) LAFNet w/ τ

Figure 6. Qualitative results of confidence map on KITTI 2015 dataset using census-SGM.

6. Conclusion

In this work, we have proposed a novel framework for monocular depth estimation based on pseudo depth labels generated by self-supervised stereo matching methods. The confidence map is used to exclude erroneous depth values within the pseudo depth labels. The prediction errors in the confidence map are further suppressed by making use of the soft-thresholding based on threshold learning. Furthermore, the probabilistic refinement module enables improving the monocular depth accuracy with the help of the uncertainty map. The proposed framework has shown impressive performances over state-of-the-arts on several popular datasets. It was also shown that threshold learning can also boost the prediction accuracy of existing confidence approaches.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 2

[2] Fabian Brickwedde, Steffen Abraham, and Rudolf Mester. Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2780–2790, 2019. 4

[3] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization. *arXiv preprint arXiv:2010.07347*, 2020.

[4] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2017. 1

[5] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *CVPR Workshop on Visual Odometry and Computer Vision Applications Based on Location Cues (VOCVALC)*, 2019. 7

[6] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. A large rgb-d dataset for semi-supervised monocular depth estimation. *arXiv preprint arXiv:1904.10230*, 2019. 1, 2, 3, 7, 8

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 4, 6

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1, 2, 5, 6, 7, 8

[9] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 2

[10] Zehua Fu and Mohsen Ardabilian Fard. Learning confidence measures by multi-modal convolutional neural networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1321–1330. IEEE, 2018. 2

[11] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016. 1, 2, 5

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 8

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 1, 2, 5, 6, 7

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3838, 2019. 1, 2, 3, 5, 6, 7

[15] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *Advances in Neural Information Processing Systems*, 33:12626–12637, 2020. 2

[16] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8977–8986, 2019. 7

[17] 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*, pages 2485–2494, 2020. 5, 6, 7, 8

[18] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018. 2

[19] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005. 6, 8

[20] Xiaoyan Hu and Philippos Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2121–2133, 2012. 8

[21] 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)*, pages 652–667, 2018. 5, 6

[22] Rongrong Ji, Ke Li, Yan Wang, Xiaoshuai Sun, Feng Guo, Xiaowei Guo, Yongjian Wu, Feiyue Huang, and Jiebo Luo. Semi-supervised adversarial monocular depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017. 5

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 5

[25] Sunok Kim, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Laf-net: Locally adaptive fusion networks for stereo confidence estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 205–214, 2019. 2, 3, 4, 5, 8

[26] Sunok Kim, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Feature augmentation for learning confidence measure in stereo matching. *IEEE Transactions on Image Processing*, 26(12):6019–6033, 2017. 2

[27] Sunok Kim, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Unified confidence estimation networks for robust stereo matching. *IEEE Transactions on Image Processing*, 28(3):1299–1313, 2018. 2

[28] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6647–6655, 2017. 2

[29] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 2

[30] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 5, 6

[31] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015. 1, 2

[32] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018. 1, 2

[33] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 8

[34] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 887–895, 2017. 1

[35] Min-Gyu Park and Kuk-Jin Yoon. Leveraging stereo matching with learning-based confidence measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 101–109, 2015. 2, 5

[36] 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*, pages 3227–3237, 2020. 1, 2, 4, 5, 6, 7

[37] Matteo Poggi, Filippo Aleotti, Fabio Tosi, Giulio Zaccaroni, and Stefano Mattoccia. Self-adapting confidence estimation for stereo. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 715–733. Springer, 2020. 2, 4, 6

[38] Matteo Poggi and Stefano Mattoccia. Learning from scratch a confidence measure. In *BMVC*, 2016. 2, 3, 4, 5, 8

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[40] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014. 8

[41] Akihito Seki and Marc Pollefeys. Patch based confidence prediction for dense disparity map. In *BMVC*, volume 2, page 4, 2016. 2

[42] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2017. 2

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[44] Aristotle Spyropoulos and Philippos Mordohai. Correctness prediction, accuracy improvement and generalization of stereo matching using supervised learning. *International Journal of Computer Vision*, 118(3):300–318, 2016. 2, 5

[45] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019. 2, 3, 4

[46] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised domain adaptation for depth prediction from images. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2, 3, 4, 7, 8

[47] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9661–9670, 2019. 7, 8

[48] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019. 5, 6, 7

[49] Fabio Tosi, Matteo Poggi, Antonio Benincasa, and Stefano Mattoccia. Beyond local reasoning for stereo confidence estimation with deep learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 319–334, 2018. 2, 3, 5

[50] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047, 2017. 2

[51] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 541–550, 2020. 2

[52] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2162–2171, 2019. 5, 6, 7

[53] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. In *European Conference on Computer Vision*, pages 722–740. Springer, 2020. 3, 7

[54] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *The journal of machine learning research*, 17(1):2287–2318, 2016. 8

[55] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018. 2

[56] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *European Conference on Computer Vision*, pages 420–439. Springer, 2020.

[57] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 2, 6