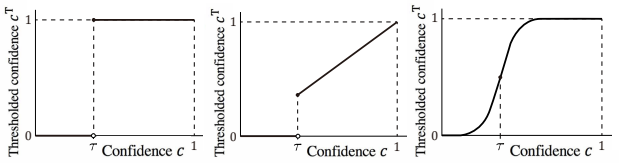
Adaptive confidence thresholding for monocular depth estimation - Supplementary material

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

In this document, we provide more comprehensive results not provided in the original manuscript due to the page limit as below. The code to reproduce our results will be publicly available soon. Note that all experiments were conducted with the supervised ThresNet.

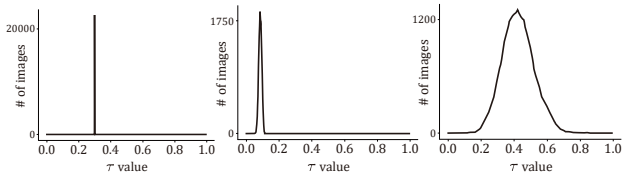
- Histogram of the learned threshold τ and qualitative evaluation of depth results computed using different thresholding methods (Section 1)
- Qualitative evaluation for monocular depth estimation with state-of-the-arts on KITTI and Cityscapes datasets (Section 2.2 and 2.3)
- Quantitative evaluation for monocular depth estimation on Cityscape dataset without fine-tuning (Section 2.4)
- Quantitative evaluation for monocular depth estimation using improved ground truth depth maps [22] on KITTI dataset (Section 2.5)
- Performance analysis according to a hyperparameter ε used in the soft-thresholding function (Section 2.6)
- Quantitative evaluation of the proposed method according to the use of DepthNet and RefineNet (Section 2.7)
- Ground truth confidence map and evaluation metric used in the confidence estimation (Section 3.1 and 3.2)
- Qualitative result for confidence estimation with state-of-the-arts on KITTI dataset (Section 3.3)
- Evaluation metric used in the uncertainty estimation (Section 4.1)
- Qualitative result for uncertainty maps on KITTI dataset (Section 4.2)

* Equal contribution. † Corresponding author.



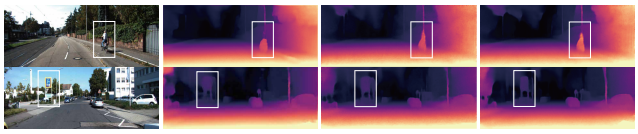
(a) (b) (c)

Figure 1. Comparison of confidence thresholding operator: (a) hard-thresholding used in [1], (b) hard-thresolding function used in [20], and (c) our soft-thresholding function. The learned threshold is used in (b) and (c), while the threshold is fixed in (a) for all training images.



(a) (b) (c)

Figure 2. Histogram of the learned threshold τ on three different thresholding methods: (a) hard-thresholding used in [1], (b) hard-thresolding function used in [20], and (c) our soft-thresholding function.

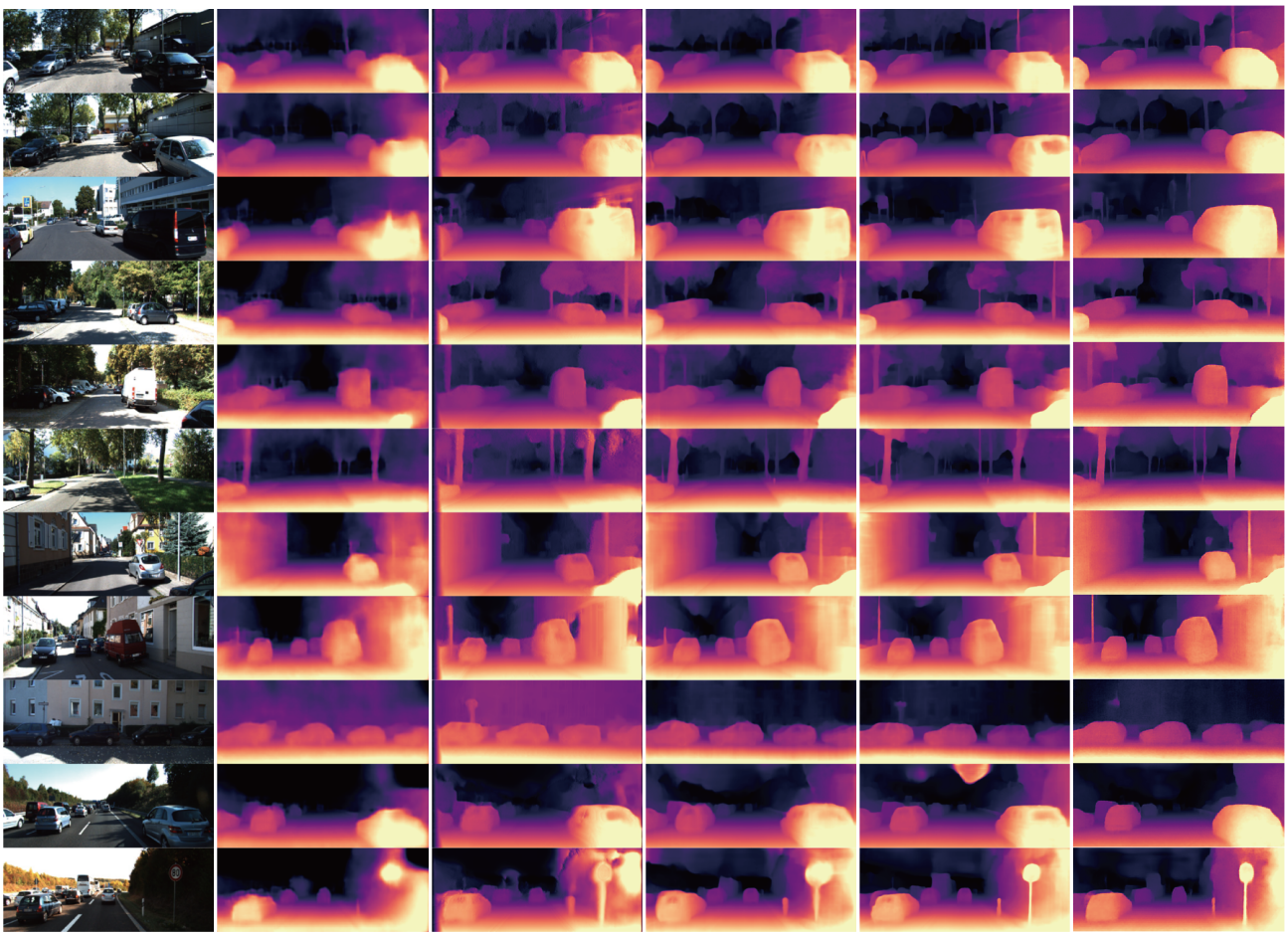


(a) (b) (c) (d)

Figure 3. Examples of gradually improved depth results of (a) input image, from (b) using the monocular depth network with fixed threshold $\tau = 0.3$ [1], (c) using the monocular depth network with learned threshold τ [20], and to (d) using the proposed method.

1. Comparison with thresholding methods

This section further highlights the effectiveness of our thresholding method by analyzing the distribution of learned threshold τ. Fig. 1 recaps the thresholding oper-



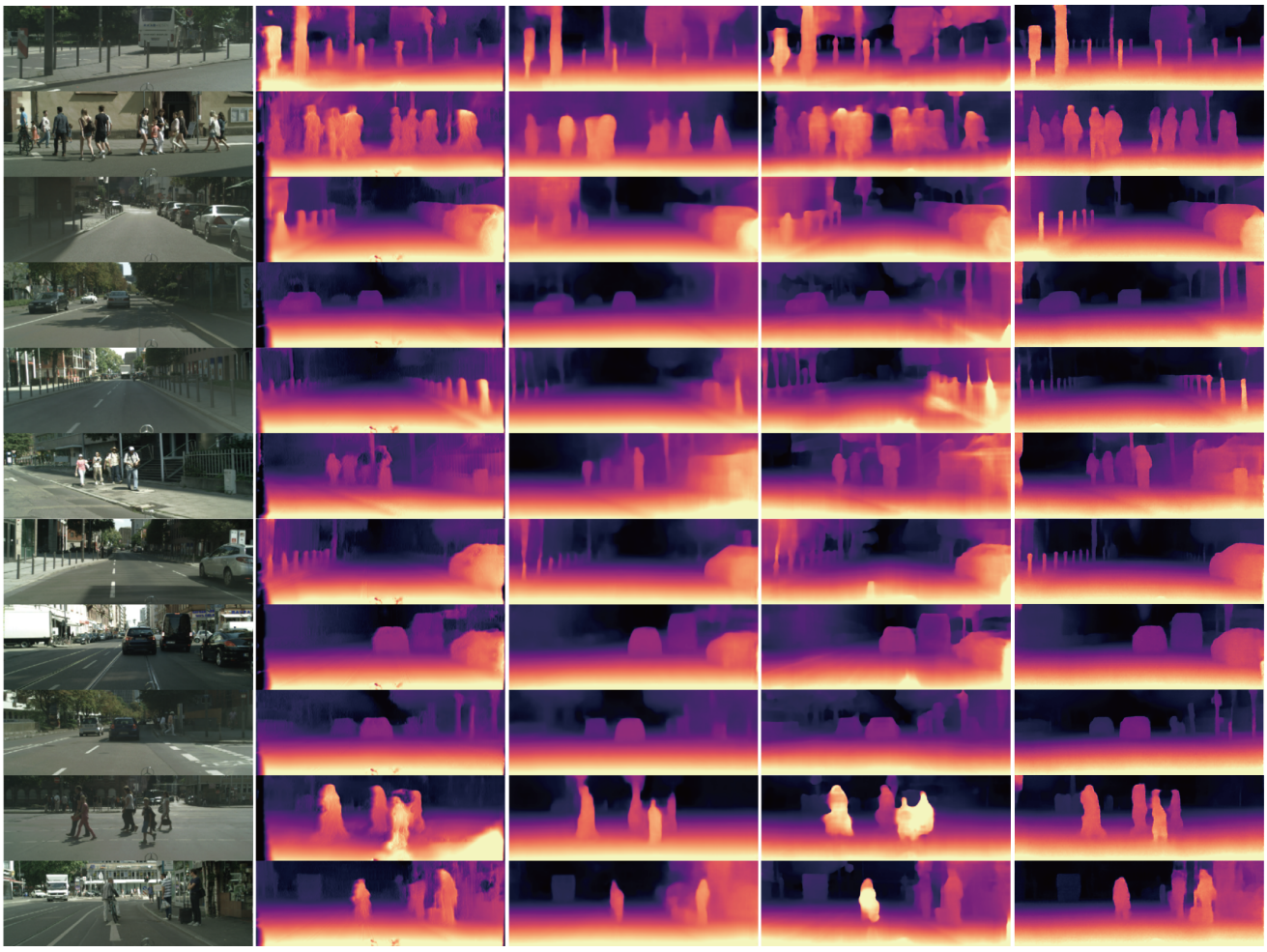
(a) (b) (c) (d) (e) (f)

Figure 4. Qualitative evaluation with existing monocular depth estimation methods on the Eigen split [3] of KITTI dataset: (a) input image, (b) Kuznietsov *et al.* [13], (c) Monodepth [5], (d) Monodepth2 [6], (e) DepthHint [24], and (f) Ours (D+R) in the submitted manuscript. Compared to other results, our method predicts instances very well without holes or distortions while recovering fine object boundaries. Additionally, our method is capable of predicting thin instances precisely.

ators (Fig. 2 of the paper) for the completeness of supplementary document. To analyze the distribution of the learned threshold τ, we plotted the histogram of τ values learned using 20k images of KITTI training dataset [4] in Fig. 2. For a fair comparison, all experiments were conducted under the same environment, including network architecture, loss function, and training data. All results were obtained without the probabilistic refinement of the RefineNet.

Cho *et al.* [1] fixed the threshold to 0.3 for all training images. Tonioni *et al.* [20] attempted to learn the threshold adaptively for each image by applying the regularization loss $-\log(1-\tau)$, but it simply prevents τ values from converging to 0 or 1 and does not take into account image characteristics that enable τ to be learned adaptively. As shown in Fig. 2 (b), τ values predicted by [20] are concentrated around specific values (0.1) with very small variance, meaning that almost similar threshold τ is used for all training images. This is the reason why the performance gain of Tonioni *et al.* [20] over Cho *et al.* [1] is relatively marginal, as reported in Tab. 4 of the original manuscript. Contrarily, our method learns image-adaptive τ values as plotted in Fig. 2 (c). Fig. 3 of the original manuscript also reports that the proposed method learned the threshold τ accordingly. The threshold τ was set low in the images where depth inference is easy, while being set high in the opposite case including saturation, low-light, and textureless region.

Fig. 3 shows the examples of gradually improved depth results according to different thresholding methods. The proposed method yields qualitatively better results, where complete instances are recovered and fine object boundaries are well preserved, than other hard thresholding methods. Also, in Tab. 4 of the original manuscript, it can be seen that the proposed method outperforms the two methods quanti-



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

Figure 5. Qualitative evaluation for depth estimation with existing methods on Cityscapes validation dataset: (a) input image, (b) [5], (c) [21], (d) [24] and (e) Ours (D+R) of the original manuscript. Similar to KITTI results, our method is remarkable at predicting fine details with no distortions at all instances and recovering thin objects that appear frequently in the Cityscapes dataset, whereas other methods often fail to predict accurate depth values at these regions.

tatively.

2. Monocular depth estimation results

This section provides more results for comparative study with state-of-the-art methods in terms of monocular depth accuracy.

2.1. Evaluation metrics

In order to evaluate the depth estimation performance, same as the original manuscript, five commonly-used evaluation metrics proposed in [3] were adopted as follows:

- Abs Rel = $\frac{1}{|\Omega|}\sum_{p\in\Omega}\frac{\left|d_p - d_p^{\mathrm{gt}}\right|}{d_p^{gt}}$
- Sq Rel = $\frac{1}{|\Omega|}\sum_{p\in\Omega}\frac{(d_p - d_p^{\mathrm{gt}})^2}{d_p^{\mathrm{gt}}}$
- RMSE = $\sqrt{\frac{1}{|\Omega|}\sum_{p\in\Omega}(d_p - d_p^{\mathrm{gt}})^2}$
- RMSE log = $\sqrt{\frac{1}{|\Omega|}\sum_{p\in\Omega}(log(d_p) - log(d_p^{\mathrm{gt}}))^2}$
- $\delta < 1.25^n = \%$ of d_p s.t. $\delta = \max(\frac{d_p}{d_p^{\mathrm{gt}}}, \frac{d_p^{\mathrm{gt}}}{d_p}) < 1.25^n$ for $n = 1, 2, 3,$

where d_p and d_p^{gt} indicate the estimated depth map and ground truth depth map at a pixel p, respectively. Ω represents a set of valid pixels.

2.2. Qualitative evaluation on KITTI

Fig. 4 shows more results on the Eigen Split [3] of KITTI dataset. We compared our results with (b) Kuznietsov *et*

Table 1. Quantitative evaluation for depth estimation with existing methods on Cityscapes validation dataset without 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	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth [5]	S	0.631	10.257	13.424	0.525	0.281	0.546	0.749
MonoResMatch [21]	S	0.241	2.149	9.064	0.296	0.570	0.891	0.966
PackNet-SfM [7]	M	0.245	2.240	8.920	0.298	0.557	0.892	0.967
Monodepth2 [6]	S	0.242	2.308	8.563	0.290	0.591	0.904	0.971
DepthHint [24]	S	**0.220**	2.008	8.363	**0.273**	0.613	0.922	<u>0.975</u>
Ours (D)	S	0.238	<u>1.983</u>	<u>8.176</u>	0.282	<u>0.629</u>	<u>0.923</u>	**0.976**
Ours (D+R)	S	<u>0.225</u>	**1.962**	**8.010**	<u>0.276</u>	**0.631**	**0.924**	**0.976**

Table 2. Quantitative evaluation for monocular depth estimation with existing methods on KITTI Eigen split dataset [3] with improved ground truth depth maps [22]. 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	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SfMLeaner [25]	M	0.176	1.532	6.129	0.244	0.758	0.921	0.971
Vid2Depth [15]	M	0.134	0.983	5.501	0.203	0.827	0.944	0.981
DDVO [23]	M	0.126	0.866	4.932	0.185	0.851	0.958	0.986
EPC++ [14]	M	0.120	0.789	4.755	0.177	0.856	0.961	0.987
Monodepth2 [6]	S	0.090	0.545	3.942	0.137	0.914	0.983	<u>0.995</u>
Uncertainty (Boot+Log) [17]	S	0.085	0.511	3.777	0.137	0.913	0.980	0.994
Uncertainty (Boot+Self) [17]	S	0.085	0.510	3.792	0.135	0.914	0.981	0.994
Uncertainty (Snap+Log) [17]	S	0.084	0.529	3.833	0.138	0.914	0.980	0.994
Uncertainty (Snap+Self) [17]	S	0.086	0.532	3.858	0.138	0.912	0.980	0.994
UnRectDepthNet [12]	M	0.081	0.414	3.412	**0.117**	0.926	**0.987**	**0.996**
PackNet-SfM [7]	M	<u>0.078</u>	0.420	3.485	0.121	**0.931**	<u>0.986</u>	**0.996**
Ours (D)	S	<u>0.078</u>	<u>0.361</u>	<u>3.223</u>	0.120	<u>0.930</u>	**0.987**	**0.996**
Ours (D+R)	S	**0.076**	**0.340**	**3.171**	<u>0.119</u>	**0.931**	**0.987**	**0.996**

al. [13], (c) Monodepth [5], (d) Monodepth2 [6], (e) DepthHint [24], and (f) Ours (D+R). Compared to other results, our method predicts instances very well without holes or distortions while recovering fine object boundaries. Additionally, our method is capable of predicting thin objects precisely.

2.3. Qualitative evaluation on Cityscapes dataset

Fig. 5 shows more qualitative results on Cityscapes dataset [2] of Fig. 5 in original manuscript. Note that it is fine-tuned on Cityscapes dataset. We compared our results with three existing methods: (b) [5], (c) [21], (d) [24] and (e) Ours (D+R) in the original manuscript. Similar to KITTI results, our method is remarkable at predicting fine details with no distortions at all instances and recovering thin objects that appear frequently in the Cityscapes dataset, whereas other methods often fail to predict accurate depth values at these regions.

2.4. Quantitative evaluation on Cityscapes dataset without fine-tuning

We also evaluated the performance of the proposed method on the Cityscapes dataset without fine-tuning. Table 1 provides the quantitative evaluation on the Cityscapes validation dataset [2], setting maximum depth to 80 meters. The performance evaluation includes Monodepth [5], MonoResMatch [21], Monodepth2 [6], DepthHint [24], PackNet-SfM [7]. Even without fine-tuning on the Cityscapes dataset, our method still outperforms state-of-the-arts approaches, and it shows that our model trained on KITTI dataset generalizes well on other dataset without bias.

2.5. Quantitative evaluation on KITTI improved ground truth depth maps

To strengthen credibility to quantitative evaluation, we also measured the monocular depth accuracy by using test frames with the improved ground truth depth maps made available in [22] for KITTI Eigen split dataset [3]. The improved ground truth maps are high quality depth maps gen-

Table 3. Quantitative depth estimation results according to ε value evaluated on KITTI Eigen Split [3] raw dataset.

ε	Abs Rel	Sqr Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
10	0.099	0.652	4.266	0.187	0.883	0.960	0.981
30	0.102	0.657	4.290	0.189	0.881	0.959	0.980
50	0.100	0.649	4.272	0.188	0.881	0.959	0.979

Table 4. Quantitative depth estimation results for three cases of the proposed method on KITTI Eigen Split [3] raw dataset.

Method	Abs Rel	Sqr Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Ours (D)	0.099	0.652	4.266	0.187	0.883	0.960	0.981
Ours (R)	0.096	0.646	4.280	0.189	0.882	0.959	0.980
Ours (D+R)	0.096	0.629	4.187	0.185	0.887	0.963	0.983

erated by accumulating LiDAR point clouds from 5 consecutive frames. Table 2 shows the quantitative evaluation with existing methods on the KITTI Eigen split dataset using the improved ground truth depth maps [22]. We compared our results with 'SfMLeaner' [25], 'Vid2Depth' [15], 'DDVO' [23], 'EPC++' [14], 'Monodepth2' [6], 'Uncertainty' [17], 'PackNet-SfM' [7], and 'UnRectDepthNet' [12]. For the training data, 'M' and 'S' indicate using a monocular video sequence and stereo images, respectively. Our method produces superior performance compared to other methods.

2.6. Choice of ε value

We set $\varepsilon = 10$ for the differentiable soft-thresholding function in (1) of the original manuscript. Table 3 shows the quantitative results according to ε on the KITTI Eigen split [3] raw dataset. Though the best accuracy was achieved with $\varepsilon = 10$, no significant change was observed depending on varying ε.

2.7. Ablation study of DepthNet and RefineNet

The evaluation of the proposed method was conducted for three cases; 'Ours (D)' trained with only the DepthNet using L_D without refining the depth map, 'Ours (R)' trained with the DepthNet and RefineNet using L_U only, and 'Ours (D+R)' trained with the DepthNet and RefineNet using all losses. Table 4 shows the quantitative results of the above three cases on KITTI Eigen Split [3] dataset. The performance gain of 'Ours (D+R)' over 'Ours (R)' supports the effectiveness of the proposed confidence learning.

3. Confidence estimation results

3.1. Generating ground-truth confidence map

To train the confidence network, the ground truth confidence map is required as supervision. Following existing confidence estimation approaches [18, 11], the ground truth confidence map c^{gt} was computed by using an absolute difference between the ground truth disparity map and the input disparity map (the pseudo ground truth disparity map in our work).

$$c_p^{\text{gt}} = \begin{cases} 1, & \text{if } |d_p - d_p^{\text{pgt}}| \leq \rho. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The threshold value ρ is set to 3 for KITTI [16] and 1 for Middlebury [19].

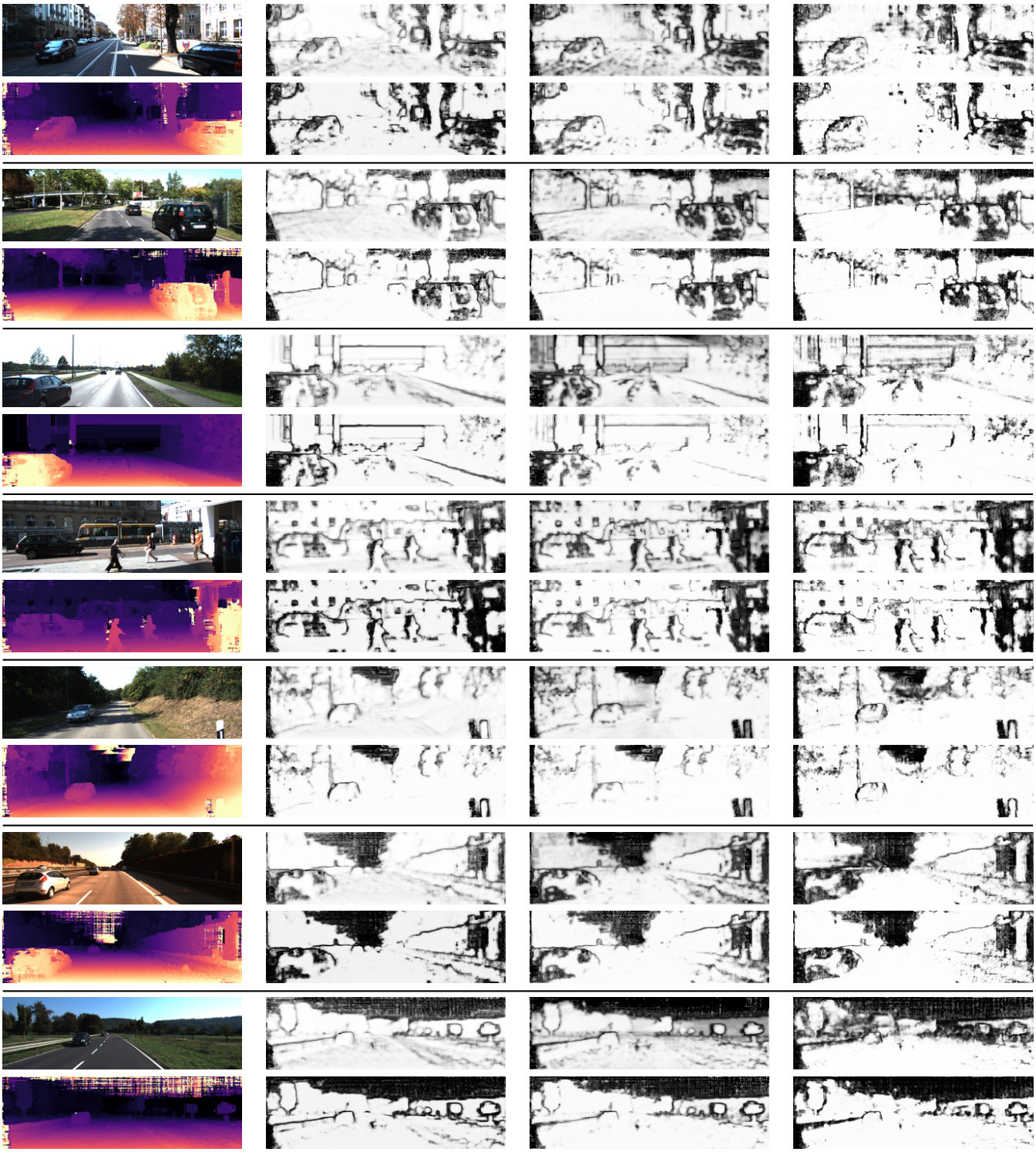
3.2. Evaluation metric

The area under the curve (AUC) [9] was used for evaluating the performance of estimated confidence maps. The receiver operating characteristic (ROC) curve is first computed by sorting disparity pixels in a decreasing order of confidence and sequentially sampling high confidence disparity pixels. It computes the error rate indicating the percentage of pixels with a difference larger than ρ from ground truth disparity. Then, AUC is computed by integral of the ROC curve. The optimal AUC is computed according to the fact that the error rate ζ is ideally 0 when sampling the first $(1-\zeta)$ pixels [9], which is equal to

$$AUC_{opt} = \int_{1-\zeta}^{1} \frac{x-(1-\zeta)}{x} dx = \zeta + (1-\zeta)\ln 1 - \zeta. \tag{2}$$

3.3. Qualitative evaluation on KITTI 2015 dataset

Fig. 6 shows more qualitative results of confidence map evaluated on KITTI 2015 dataset [16]. Input disparity maps used for confidence estimation were obtained by Census-SGM [8]. The estimated confidence maps for each input disparity map are displayed every two rows. The top and bottom of two rows indicate: (a) color image and input disparity image, (b) CCNN [18] and CCNN w/τ, (c) LAFNet* [11] and LAFNet* w/τ and (d) LAFNet and LAFNet w/τ. 'w/τ' denotes the thresholded confidence map obtained using the soft-thresholding. LAFNet* denotes the LAFNet [11] in which 3D cost volume is not used as an input. As shown in Fig. 6, the proposed thresholded confidence maps contain fewer ambiguous values than the original confidence maps.



(a) (b) (c) (d)

Figure 6. Qualitative evaluation for confidence estimation on KITTI 2015 dataset [16]: Input disparity maps used for confidence estimation were obtained by Census-SGM [8]. The estimated confidence maps for each input disparity map are displayed every two rows. The top and bottom of two rows indicate: (a) color image and input disparity image, (b) CCNN [18] and CCNN w/τ, (c) LAFNet* [11] and LAFNet* w/τ and (d) LAFNet and LAFNet w/τ. 'w/τ' denotes the proposed network using the soft-thresholding. LAFNet* denotes the LAFNet [11] in which 3D cost volume is not used as an input.

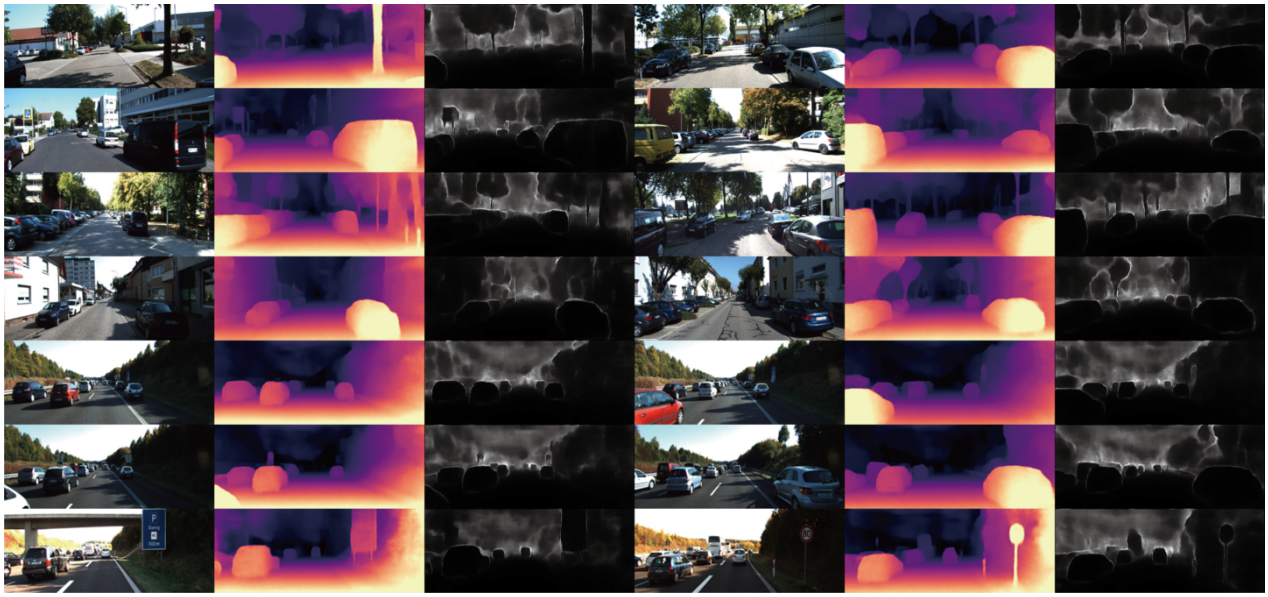


Figure 7. Qualitative result of uncertainty measure on KITTI Eigen Split [3] test dataset. From left to right, input image, estimated depth map, and uncertainty map of the estimated depth are displayed.

4. Uncertainty Estimation Details

4.1. Evaluation metric

We evaluated the performance of the uncertainty estimation used in the proposed model using the sparsification error [10]. Similar to the confidence evaluation, we first sorted disparity pixels following decreasing order of uncertainty, and iteratively extracted high uncertain disparities and provided them as inputs for computing error metrics. The ideal error ranked by the true error to the ground truth is referred to as oracle. With a sparsicifation error, we computed the Area Under the Sparsification Error curve (AUSE) and the Area Under the Random Gain (AURG) to evaluate the quality of the uncertainty map. While the AUSE is measured as the difference between the sparsification and its oracle, the AURG is obtained as subtracting the estimated sparsification curve from flat curve with a random uncertainty which is modeled as a constant.

4.2. Qualitative evaluation on KITTI dataset

Fig. 7 shows the qualitative results of uncertainty map evaluated on KITTI Eigen Split [3] test dataset. The qualitative result indicates that uncertain areas of the estimated depth map are usually located around object boundaries and sky.

References

[1] 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

[2] 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

[3] 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. 2, 3, 4, 5, 7

[4] 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. 2

[5] 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. 2, 3, 4

[6] 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. 2, 4, 5

[7] 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. 4, 5

[8] 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. 5, 6

[9] 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. 5

[10] 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. 7

[11] 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. 5, 6

[12] Varun Ravi Kumar, Senthil Yogamani, Markus Bach, Christian Witt, Stefan Milz, and Patrick Mader. Unrectdepthnet: Self-supervised monocular depth estimation using a generic framework for handling common camera distortion models. *arXiv preprint arXiv:2007.06676*, 2020. 4, 5

[13] 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, 4

[14] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ramkant Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 4, 5

[15] Reza Mahjourian, Martin Wicke, and Anelia Angelova. 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*, pages 5667–5675, 2018. 4, 5

[16] 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. 5, 6

[17] 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. 4, 5

[18] Matteo Poggi and Stefano Mattoccia. Learning from scratch a confidence measure. In *BMVC*, 2016. 5, 6

[19] 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. 5

[20] 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

[21] 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. 3, 4

[22] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 1, 4, 5

[23] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018. 4, 5

[24] 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. 2, 3, 4

[25] 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. 4, 5