– Supplementary Material – These Maps Are Made by Propagation: Adapting Deep Stereo Networks to Road Scenarios with Decisive Disparity Diffusion

Chuang-Wei Liu, Yikang Zhang, Qijun Chen, *Senior Member, IEEE*, Ioannis Pitas, *Life Fellow, IEEE*, and Rui Fan, *Senior Member, IEEE*

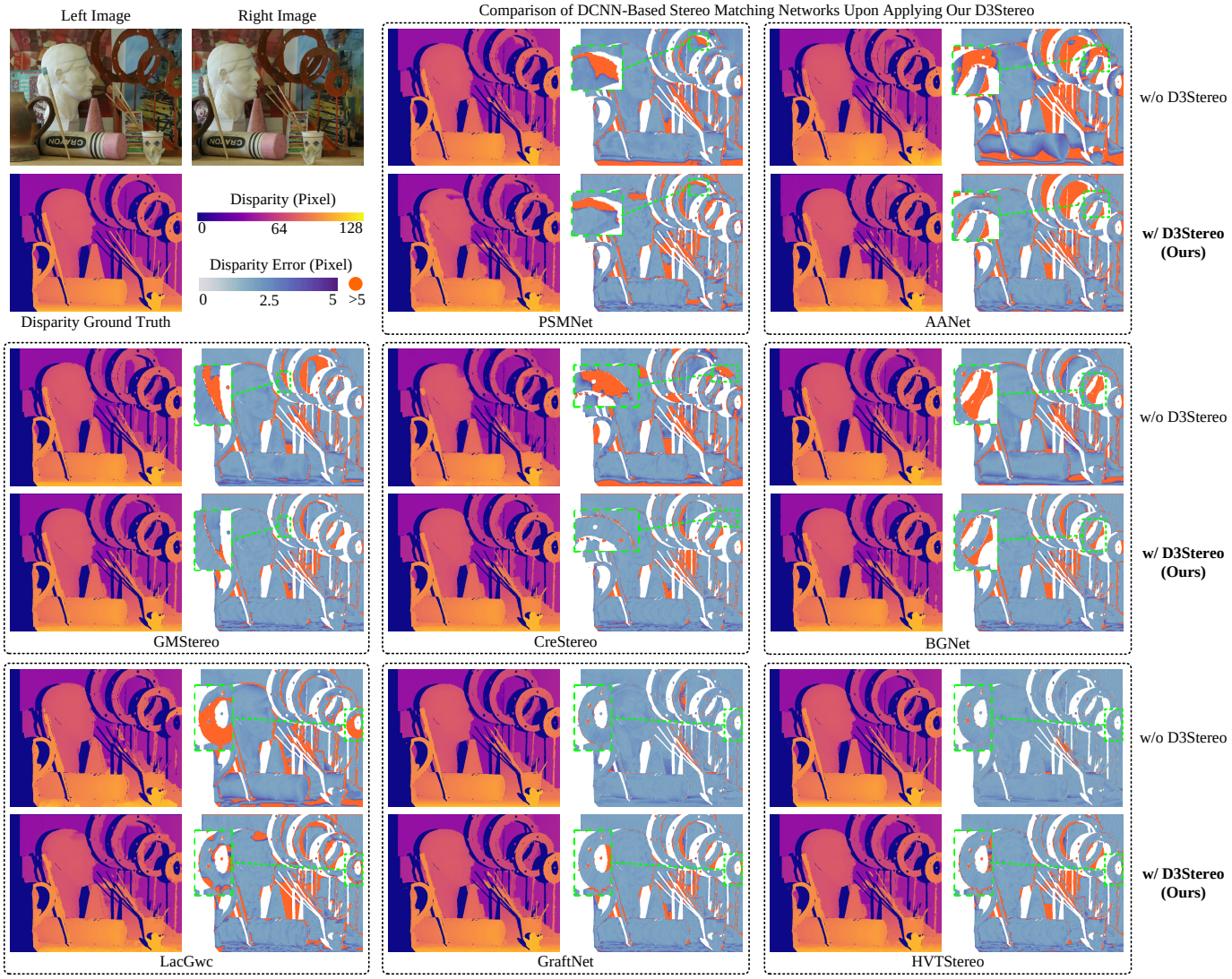

Fig. 1. Examples of disparity estimation results on the Middlebury dataset. Significantly improved regions are shown with green dashed boxes.

I. Performance Evaluation of Stereo Matching Networks Fine-tuned on the UDTIRI-Stereo Dataset

To further demonstrate the significance of our proposed Decisive Disparity Diffusion (D3Stereo) and UDTIRI-Stereo dataset, we fine-tune the eight stereo matching networks on the UDTIRI-Stereo dataset and then evaluate them on both the UDTIRI-Stereo and Stereo-Road datasets. Specifically, scenes of tidy road surfaces with mild sunlight are used for evaluation due to their optimal weather conditions, while other scenes are used for model fine-tuning. The quantitative results on UDTIRI-Stereo and Stereo-Road datasets are presented in Tables I and II, respectively. It can be observed that all deep convolutional neural networks (DCNNs) exhibit remarkable stereo matching accuracy on the UDTIRI-Stereo validation

TABLE I
COMPARISONS OF FINE-TUNED SOTA STEREO MATCHING NETWORKS WITHOUT AND WITH OUR PROPOSED D3STEREO STRATEGY APPLIED ON THE UDTIRI-STEREO VALIDATION SUBSET.

Method	PEP (%) ↓		EPE (pixel) ↓	PSNR (dB) ↑	MSE ↓	SSIM ↑
	δ=0.5	δ=1				
PSMNet [1]	6.21	2.35	0.34	33.98	40.72	0.949
PSMNet+D3Stereo (Ours)	**1.96**	**1.35**	**0.25**	**34.81**	**35.19**	**0.951**
AANet [2]	**5.10**	2.04	**0.30**	**34.62**	**37.75**	**0.951**
AANet+D3Stereo (Ours)	7.78	**1.07**	0.40	34.39	39.28	0.949
BGNet [3]	4.11	1.03	0.22	35.03	35.33	0.954
BGNet+D3Stereo (Ours)	**0.85**	**0.48**	**0.16**	**35.21**	**34.33**	**0.956**
LacGwc [4]	**1.30**	**0.04**	**0.15**	**35.75**	**28.95**	**0.959**
LacGwc+D3Stereo (Ours)	1.67	1.14	0.28	34.77	30.96	0.954
GMStereo [5]	4.15	**0.19**	**0.18**	**35.26**	**29.41**	**0.957**
GMStereo+D3Stereo (Ours)	**1.02**	0.64	0.26	33.68	33.38	0.951
CreStereo [6]	**0.18**	**0.03**	**0.12**	**36.01**	**27.67**	**0.961**
CreStereo+D3Stereo (Ours)	3.41	1.46	0.31	32.76	36.66	0.954
GraftNet [7]	2.39	**0.08**	**0.11**	**36.17**	**27.92**	**0.961**
GraftNet+D3Stereo (Ours)	**0.66**	0.08	0.13	35.96	28.11	0.959
HVTStereo [8]	**1.12**	**0.06**	**0.17**	**35.76**	**29.01**	**0.958**
HVTStereo+D3Stereo (Ours)	2.15	0.27	0.18	35.49	29.97	0.958

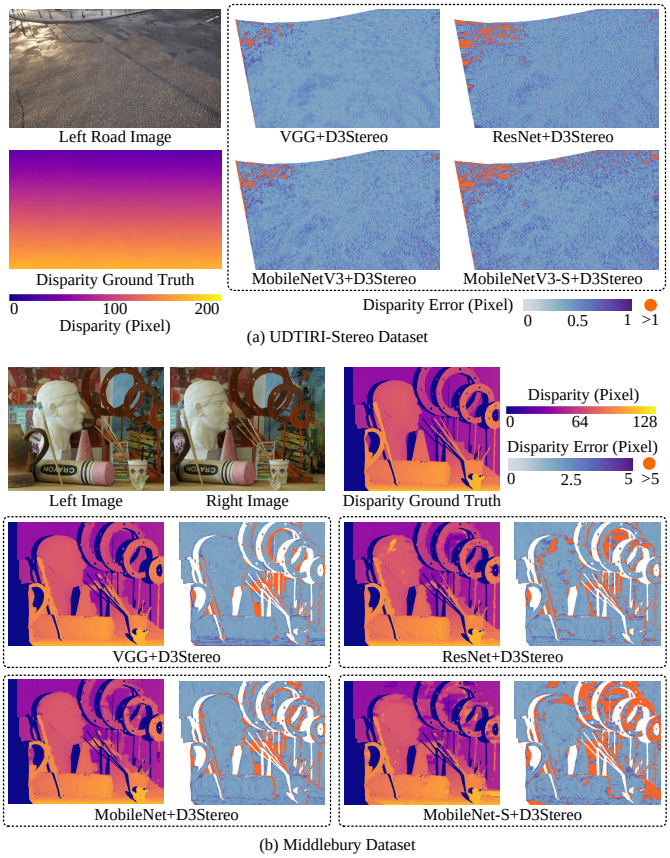

Fig. 2. Examples of disparity estimation results yielded by applying our proposed D3Stereo strategy on DCNN backbones pre-trained for image classification.

subset after the model fine-tuning. Specifically, the fine-tuned LacGwc, CreStereo, and HVTStereo surpass their combinations with D3Stereo strategy across all metrics. On the Stereo-Road dataset, the fine-tuned LacGwc, GMStereo, CreStereo, and GraftNet achieve higher peak signal-to-noise ratio (PSNR) compared with their combinations with D3Stereo strategy. Additionally, applying our D3Stereo strategy to all stereo matching DCNNs leads to improved structure similarity index measure (SSIM). In general, stereo matching networks fine-tuned on our synthetic UDTIRI-Stereo dataset demonstrate improved stereo matching accuracy on the real-world Stereo-Road dataset, and D3Stereo strategy achieves satisfactory road surface 3D reconstruction performance without the need for any model fine-tuning.

TABLE II
COMPARISONS OF FINE-TUNED SOTA STEREO MATCHING NETWORKS WITHOUT AND WITH OUR PROPOSED D3STEREO STRATEGY APPLIED ON THE STEREO-ROAD DATASET. THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

Method	PSNR (dB) ↑	MSE ↓	SSIM ↑
PSMNet [1]	31.61	59.28	0.932
PSMNet+D3Stereo (Ours)	**31.77**	**55.10**	**0.940**
AANet [2]	31.11	68.36	0.917
AANet+D3Stereo (Ours)	**31.35**	**60.49**	**0.931**
BGNet [3]	31.63	56.89	0.933
BGNet+D3Stereo (Ours)	**31.65**	**56.18**	**0.938**
LacGwc [4]	**31.91**	**54.51**	0.932
LacGwc+D3Stereo (Ours)	31.74	55.33	**0.940**
GMStereo [5]	**32.01**	**53.56**	0.932
GMStereo+D3Stereo (Ours)	31.70	56.14	**0.939**
CreStereo [6]	**31.88**	**55.15**	0.931
CreStereo+D3Stereo (Ours)	31.62	56.98	**0.936**
GraftNet [7]	**31.81**	56.11	0.929
GraftNet+D3Stereo (Ours)	31.70	**55.48**	**0.939**
HVTStereo [8]	31.55	57.08	0.935
HVTStereo+D3Stereo (Ours)	**31.73**	**56.22**	**0.940**

TABLE III
COMPARISONS OF D3STEREO WITH ADDITIONAL COST VOLUMES CONSTRUCTED BASED ON NCC AND VGG FEATURE SIMILARITY AT FULL IMAGE RESOLUTION. THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

Cost calculation	Sizes of the selected feature maps of VGG	UDTIRI-Stereo dataset		Middlebury dataset	
		EPE (pixel)↓	Runtime (s)↓	EPE (pixel)↓	Runtime (s)↓
Feature similarity	[(H,W,64), ($\frac{H}{2}$, $\frac{W}{2}$,128), ($\frac{H}{4}$, $\frac{W}{4}$,256), ($\frac{H}{8}$, $\frac{W}{8}$,512)]	**0.172**	**2.17**	**3.01**	**3.04**
NCC	[($\frac{H}{2}$, $\frac{W}{2}$,128), ($\frac{H}{4}$, $\frac{W}{4}$,256), ($\frac{H}{8}$, $\frac{W}{8}$,512)]	0.183	2.22	3.17	3.34

II. COMPARISONS OF D3STEREO WITH ADDITIONAL COST VOLUME AT FULL IMAGE RESOLUTION

The VGG backbone is capable of extracting feature maps at full image resolution. Therefore, we further conduct experiments by deploying D3Stereo on VGG with matching cost volumes at full image resolution, constructed using either normalized cross-correlation (NCC) or VGG feature similarity. The quantitative results presented in Table III indicate that performing D3Stereo with additional cost volumes at full image resolution based on both NCC and VGG feature similarity leads to improved EPE by 24.1-28.6% on the UDTIRI-Stereo dataset and 15.4-19.7% on the Middlebury dataset, respectively. Constructing full-resolution cost volumes with NCC can significantly increase D3Stereo's accuracy and building cost volumes with deep features provided by the pre-trained VGG yields even higher stereo matching accuracy and efficiency.

REFERENCES

[1] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

[2] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1959–1968.

[3] B. Xu *et al.*, "Bilateral grid learning for stereo matching networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 497–12 506.

[4] B. Liu *et al.*, "Local similarity pattern and cost self-reassembling for deep stereo matching networks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022, pp. 1647–1655.

[5] H. Xu *et al.*, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 941–13 958, 2023.

[6] J. Li *et al.*, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 263–16 272.

[7] B. Liu *et al.*, "GraftNet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 012–13 021.

[8] T. Chang *et al.*, "Domain generalized stereo matching via hierarchical visual transformation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 9559–9568.