Active Stereo Without Pattern Projector – supplementary material

Luca Bartolomei*,† Matteo Poggi† Fabio Tosi† Andrea Conti† Stefano Mattoccia*,†
*Advanced Research Center on Electronic System (ARCES)
†Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{luca.bartolomei5, m.poggi, fabio.tosi5, andrea.conti35, stefano.mattoccia}@unibo.it
https://vppstereo.github.io/

This document provides additional details regarding ICCV paper "Active Stereo Without Pattern Projector".

We used 5% of sparse depth points for all the experiments reported in this document.

1. Occlusion handling – Heuristic grid search

In this section, we report further details about our occlusion handling methodology used to estimate the occluded points in VPP. We performed a grid search to find the optimal hyper-parameters for the heuristic function used for this purpose. Specifically, we fixed the range of r_x and r_y to $5, 7, 9, 11, 13$, and $\lambda \in 1, 2, 3$, $\gamma \in \{0, 0.0714, 0.1429, 0.2143, 0.2857, 0.3571, 0.4286, 0.5, 0.5714, 0.6429, 0.7143, 0.7857, 0.8571, 0.9286, 1\}$, and $t \in 1, 2, 3$. To evaluate the effect of each configuration, we used the Middlebury [10] *Additional* split (Midd-A), over which we measure how good our heuristic is at classifying pixels as occluded by reporting accuracy, precision, recall and F1-score metrics. The grid search results are shown in Table I, showing only the top-50 entries sorted by F1-Score. We found that r_x and r_y were the most important hyper-parameters, as they remained the same for all top-50 entries. Furthermore, λ preferred values greater than one (i.e., none of the top-50 configuration uses $\lambda = 1$), while γ and t were less influential hyper-parameters. These findings suggest that the choice of r_x and r_y is crucial in estimating occluded points accurately and that increasing the value of λ can lead to improved performance.

All fifty configurations selected during the grid search show comparable performance based on F1-score. However, we chose $r_x = 9$, $r_y = 7$, $\lambda = 2$, $\gamma = 0.4375$, and $t = 1$, as they provided a slight improvement in recall. We argue that reducing false negative classification errors could be beneficial, as points classified as occluded are still projected in the "FGD"-projection configuration and thus could lead to ambiguities if not properly detected.

Midd-A

r_x	r_y	λ	γ	t	Accuracy	Precision	Recall	F1-Score
9	7	3	0.2857	2	98.6	92.8	93.7	93.2
9	7	2	0.5714	3	98.6	92.7	93.8	93.2
9	7	3	0.5000	1	98.6	92.7	93.8	93.2
9	7	3	0.7143	2	98.6	92.9	93.6	93.2
9	7	3	0.5714	2	98.6	92.8	93.7	93.2
9	7	3	0.7857	3	98.6	93.0	93.4	93.2
9	7	3	1.0000	3	98.6	92.8	93.6	93.2
9	7	3	0.2143	3	98.6	93.0	93.5	93.2
9	7	2	0.5000	3	98.6	92.6	93.8	93.2
9	7	3	0.7143	1	98.6	92.6	93.9	93.2
9	7	3	0.7857	2	98.6	92.8	93.6	93.2
9	7	3	0.3571	3	98.6	93.0	93.5	93.2
9	7	3	0.5000	2	98.6	92.8	93.6	93.2
9	7	3	0.8571	2	98.6	92.8	93.6	93.2
9	7	2	0.3571	3	98.6	92.7	93.7	93.2
9	7	3	0.2857	1	98.6	92.6	93.8	93.2
9	7	3	0.7857	1	98.6	92.6	93.8	93.1
9	7	2	0.7857	2	98.6	92.5	93.9	93.1
9	7	2	0.6429	2	98.6	92.5	93.9	93.1
9	7	2	0.6429	3	98.6	92.7	93.7	93.1
9	7	3	0.2143	2	98.6	92.7	93.6	93.1
9	7	3	0.3571	1	98.6	92.6	93.8	93.1
9	7	2	0.8571	3	98.6	92.6	93.8	93.1
9	7	2	0.2857	3	98.6	92.6	93.8	93.1
9	7	3	0.3571	2	98.6	92.7	93.7	93.1
9	7	3	0.2143	1	98.5	92.5	93.9	93.1
9	7	2	0.7143	3	98.6	92.6	93.8	93.1
9	7	3	0.5714	1	98.6	92.6	93.7	93.1
9	7	3	0.2857	3	98.6	92.9	93.5	93.1
9	7	3	0.1429	2	98.6	92.6	93.7	93.1
9	7	2	0.3571	2	98.6	92.5	93.9	93.1
9	7	2	0.4375	3	98.5	92.6	93.8	93.1
9	7	2	0.1429	3	98.6	92.6	93.8	93.1
9	7	2	0.4375	2	98.5	92.4	94.0	93.1
9	7	3	0.6429	1	98.6	92.6	93.8	93.1
9	7	3	0.5000	3	98.6	93.0	93.3	93.1
9	7	2	0.2857	2	98.5	92.5	93.9	93.1
9	7	3	0.4375	2	98.6	92.7	93.6	93.1
9	7	2	0.5714	2	98.5	92.5	93.9	93.1
9	7	3	0.8571	3	98.6	92.9	93.5	93.1
9	7	3	0.7143	3	98.6	93.0	93.3	93.1
9	7	3	0.4375	1	98.6	92.6	93.7	93.1
9	7	2	0.5000	2	98.5	92.5	93.9	93.1
9	7	2	0.7143	2	98.5	92.4	93.9	93.1
9	7	3	0.5714	3	98.6	92.9	93.4	93.1
9	7	3	0.6429	3	98.6	93.0	93.3	93.1
9	7	3	0.6429	2	98.6	92.8	93.5	93.1
9	7	2	0.2143	3	98.5	92.6	93.7	93.1
9	7	2	1.0000	3	98.5	92.6	93.7	93.1
9	7	2	0.4375	1	98.5	92.3	94.1	93.1

Table I: **Occlusion handling – grid search.** We evaluate the effect of the hyper-parameters $r_x, r_y, \lambda, \gamma, t$.

2. Complete Ablation Experiment

We extend the ablation study reported in the main paper (Tab. 1) with an exhaustive study with a larger number of configurations. Tab. II collects the results achieved by playing with the three occlusion handling strategies – i.e., "BKGD", "NO", and "FGD", respectively, in sub-tables (a), (b) and (c) –, the different virtual patterns (ii)-(vii) presented in the main paper, and also measuring the impact of Alpha-blending (α) with different intensities

Middd-A

VPP	hyper-parameters				Error Rate (%) > 2		
	Pattern	α	Patch	Occ.	RAFT-St. [4]	PSMNet [1]	rSGM
✗	✗	✗	✗	✗	11.5	29.3	34.3
✓	(ii)	✗	✗	BKGD	5.2	15.3	20.6
✓	(ii)	0.8	✗	BKGD	5.3	15.5	20.7
✓	(ii)	0.4	✗	BKGD	5.8	16.7	21.2
✓	(iii)	✗	✗	BKGD	5.1	15.2	20.2
✓	(iii)	0.8	✗	BKGD	5.2	15.3	20.3
✓	(iii)	0.4	✗	BKGD	5.6	16.1	20.5
✓	(vi)	✗	3×3	BKGD	5.2	16.0	15.5
✓	(vi)	0.8	3×3	BKGD	5.1	15.7	15.5
✓	(vi)	0.4	3×3	BKGD	5.0	15.3	15.9
✓	(iv)	✗	3×3	BKGD	4.8	16.0	16.6
✓	(iv)	0.8	3×3	BKGD	4.8	15.6	16.5
✓	(iv)	0.4	3×3	BKGD	4.9	15.0	16.7
✓	(vii)	✗	3×3	BKGD	4.9	16.1	16.0
✓	(vii)	0.8	3×3	BKGD	4.9	15.8	15.9
✓	(vii)	0.4	3×3	BKGD	5.0	15.2	15.9
✓	(v)	✗	3×3	BKGD	4.8	16.4	16.6
✓	(v)	0.8	3×3	BKGD	4.8	15.9	16.3
✓	(v)	0.4	3×3	BKGD	5.0	15.1	16.2
✓	(vi)	✗	5×5	BKGD	5.8	17.9	14.8
✓	(vi)	0.8	5×5	BKGD	5.5	17.1	14.8
✓	(vi)	0.4	5×5	BKGD	5.0	15.7	14.9
✓	(iv)	✗	5×5	BKGD	5.3	18.0	18.8
✓	(iv)	0.8	5×5	BKGD	5.2	17.3	17.1
✓	(iv)	0.4	5×5	BKGD	5.0	15.8	17.0
✓	(vii)	✗	5×5	BKGD	5.3	17.8	15.9
✓	(vii)	0.8	5×5	BKGD	5.0	17.2	15.6
✓	(vii)	0.4	5×5	BKGD	4.9	16.0	15.5
✓	(v)	✗	5×5	BKGD	5.2	18.1	19.3
✓	(v)	0.8	5×5	BKGD	5.1	17.5	17.3
✓	(v)	0.4	5×5	BKGD	4.9	16.3	16.9

(a)

Middd-A

VPP	hyper-parameters				Error Rate (%) > 2		
	Pattern	α	Patch	Occ.	RAFT-St. [4]	PSMNet [1]	rSGM
✗	✗	✗	✗	✗	11.5	29.3	34.3
✓	(ii)	✗	✗	FGD	5.2	15.0	20.5
✓	(ii)	0.8	✗	FGD	5.3	15.3	20.6
✓	(ii)	0.4	✗	FGD	5.8	16.6	21.1
✓	(iii)	✗	✗	FGD	5.2	14.9	20.1
✓	(iii)	0.8	✗	FGD	5.3	15.0	20.2
✓	(iii)	0.4	✗	FGD	5.6	16.0	20.5
✓	(vi)	✗	3×3	FGD	4.8	14.9	14.9
✓	(vi)	0.8	3×3	FGD	4.8	14.7	15.1
✓	(vi)	0.4	3×3	FGD	5.0	14.6	15.6
✓	(iv)	✗	3×3	FGD	4.6	14.7	15.7
✓	(iv)	0.8	3×3	FGD	4.6	14.6	15.7
✓	(iv)	0.4	3×3	FGD	4.8	**14.4**	16.1
✓	(vii)	✗	3×3	FGD	4.6	14.8	15.2
✓	(vii)	0.8	3×3	FGD	4.6	14.7	15.1
✓	(vii)	0.4	3×3	FGD	4.8	**14.4**	15.3
✓	(v)	✗	3×3	FGD	4.5	14.8	15.5
✓	(v)	0.8	3×3	FGD	4.5	14.6	15.3
✓	(v)	0.4	3×3	FGD	4.8	**14.4**	15.6
✓	(vi)	✗	5×5	FGD	4.6	15.6	**13.6**
✓	(vi)	0.8	5×5	FGD	4.6	15.2	13.7
✓	(vi)	0.4	5×5	FGD	4.7	14.7	14.0
✓	(iv)	✗	5×5	FGD	4.4	15.4	17.0
✓	(iv)	0.8	5×5	FGD	**4.3**	15.1	15.4
✓	(iv)	0.4	5×5	FGD	4.6	14.7	15.8
✓	(vii)	✗	5×5	FGD	**4.3**	15.6	14.3
✓	(vii)	0.8	5×5	FGD	**4.3**	15.2	14.1
✓	(vii)	0.4	5×5	FGD	4.5	14.6	14.3
✓	(v)	✗	5×5	FGD	**4.3**	15.5	17.3
✓	(v)	0.8	5×5	FGD	**4.3**	15.2	15.4
✓	(v)	0.4	5×5	FGD	4.4	14.7	15.4

(b)

Middd-A

VPP	hyper-parameters				Error Rate (%) > 2		
	Pattern	α	Patch	Occ.	RAFT-St. [4]	PSMNet [1]	rSGM
✗	✗	✗	✗	✗	11.5	29.3	34.3
✓	(ii)	✗	✗	NO	5.3	15.2	20.5
✓	(ii)	0.8	✗	NO	5.4	15.4	20.7
✓	(ii)	0.4	✗	NO	5.8	16.6	21.2
✓	(iii)	✗	✗	NO	5.2	15.0	20.2
✓	(iii)	0.8	✗	NO	5.3	15.2	20.3
✓	(iii)	0.4	✗	NO	5.6	16.0	20.5
✓	(vi)	✗	3×3	NO	5.1	15.5	15.1
✓	(vi)	0.8	3×3	NO	5.0	15.2	15.2
✓	(vi)	0.4	3×3	NO	5.1	14.9	15.7
✓	(iv)	✗	3×3	NO	4.8	15.3	15.9
✓	(iv)	0.8	3×3	NO	4.8	15.1	15.8
✓	(iv)	0.4	3×3	NO	4.9	14.7	16.2
✓	(vii)	✗	3×3	NO	4.8	15.5	15.4
✓	(vii)	0.8	3×3	NO	4.8	15.4	15.2
✓	(vii)	0.4	3×3	NO	4.9	14.9	15.4
✓	(v)	✗	3×3	NO	4.7	15.5	15.7
✓	(v)	0.8	3×3	NO	4.8	15.3	15.5
✓	(v)	0.4	3×3	NO	4.9	14.8	15.7
✓	(vi)	✗	5×5	NO	5.2	16.3	13.8
✓	(vi)	0.8	5×5	NO	5.1	15.9	13.8
✓	(vi)	0.4	5×5	NO	5.0	15.2	14.3
✓	(iv)	✗	5×5	NO	4.9	16.4	17.0
✓	(iv)	0.8	5×5	NO	4.9	16.0	15.5
✓	(iv)	0.4	5×5	NO	4.9	15.2	15.8
✓	(vii)	✗	5×5	NO	4.8	16.4	14.4
✓	(vii)	0.8	5×5	NO	4.8	16.2	14.2
✓	(vii)	0.4	5×5	NO	4.8	15.5	14.5
✓	(v)	✗	5×5	NO	4.8	16.6	17.4
✓	(v)	0.8	5×5	NO	4.7	16.2	15.5
✓	(v)	0.4	5×5	NO	4.8	15.5	15.6

(c)

Table II: **Ablation on main projection hyper-parameters.** Results on Midd-A. Networks trained on synthetic data.

(none (i.e., 1.0, 0.4 and 0.8) and the patch size (none (i.e., 1×1), 3×3 or 5×5).

As in the main paper, we study the impact of the different VPP configurations on the disparity maps predicted by RAFT-Stereo [4], PSMNet [1] or estimated by the rSGM algorithm [7] on the Middlebury [10] *Additional* split (Midd-A). For RAFT-Stereo and PSMNet, we use weights obtained after training on synthetic images only. From a first look, we can notice how determining the best overall configuration is not trivial: indeed, the configuration we select for the experiments in the main paper (in yellow) is not the absolute winner in this study, yet allowing to outperform existing techniques already [6, 2]. Nonetheless, from this exhaustive study, we draw some general take-home messages to define the best set of configurations to deploy VPP.

Occlusion-handling. By comparing sub-tables (a), (b) and (c), we can notice that a proper handling of the pattern projection near occlusions is helpful, most times, to attain the best results. Indeed, applying the virtual pattern according to "FGD" strategy (b) – i.e., by encouraging the matching of an occluded region with its own occluder – yields better results with respect to not handling occlusions at all (a) or to the naïve projection of the pattern in the not occluded regions only (c), when the remaining hyper-parameters remain unchanged.

Patch size. In general, projecting a pattern on local patches (iv)-(vii) rather than single pixels (ii)-(iii) yields better results. Not surprisingly, acting on a local region allows for easing the visual correspondence task much more than intervening on single pixels alone. Increasing the patch size from 3×3 to 5×5 can further improve the results with rSGM, while it yields mixed results with RAFT-Stereo (often improving) and PSMNet (which seems to benefit more from 3×3 patches). We argue that this behavior is caused by the higher-level cues used by deep stereo networks to perform matching, not necessarily affected by the local visual distinctiveness in the same manner as traditional algorithms.

Alpha-blending (α). Since image content is crucial for modern stereo networks to compute correspondences, a proper Alpha-blending of the projected pattern with the original color images is often beneficial. In particular, it allows for compensating for the possibly erroneous correspondence being enforced by large patches near object boundaries. On the contrary, Alpha-blending always reduces the effectiveness of the virtual pattern when running rSGM, since it dampens the high distinctiveness of the pattern, which is crucial to ease correspondences when using hand-crafted matching functions.

Different networks/algorithms. Finally, our exhaustive ablation study highlights how different stereo networks and algorithms react differently to the configurations we experimented. Indeed, depending on the network or algorithm, we have different configurations yielding the best results. Specifically, RAFT-stereo [4] obtains the best results with 5×5 patterns (iv)-(v) or (vii), often by setting a high Alpha-blending factor ($\alpha = 0.8$) or no Alpha-blending factor at all ($\alpha = 1.0$); PSMNet [1], on the contrary, registers the best improvement with 3×3 patterns (iv)-(v) or (vii) and a moderate Alpha-blending factor ($\alpha = 0.4$); finally, rSGM [7] attains the best results with 5×5 pattern (vi), without any Alpha-blending being applied. For this reason, we selected a configuration for our main experiments that could balance the accuracy of the three methods without favoring any in particular.

				Midd-14					Midd-21					ETH3D					KITTI 142				
		Depth Points		Error Rate std. (%)				avg. std.	Error Rate std. (%)				avg. std.	Error Rate std. (%)				avg. std.	Error Rate std. (%)				avg. std.
Model	Model name	Train	Test	> 1	> 2	> 3	> 4	(px)	> 1	> 2	> 3	> 4	(px)	> 1	> 2	> 3	> 4	(px)	> 1	> 2	> 3	> 4	(px)
RAFT-Stereo-*vpp* [4]	Sceneflow	✗	✓	0.06	0.05	0.05	0.07	0.04	0.00	0.02	0.00	0.02	0.00	0.10	0.08	0.07	0.06	0.01	0.00	0.00	0.00	0.00	0.00
RAFT-Stereo-*vpp* [4]	Middlebury	✗	✓	0.06	0.07	0.09	0.08	0.02	0.06	0.07	0.05	0.06	0.02	0.07	0.05	0.04	0.03	0.00	0.03	0.00	0.02	0.01	0.00
RAFT-Stereo-*vpp* [4]	ETH3D	✗	✓	0.08	0.08	0.11	0.13	0.02	0.04	0.04	0.04	0.03	0.01	0.03	0.03	0.04	0.03	0.01	0.02	0.02	0.01	0.01	0.00
GMStereo-*vpp** [11]	Sceneflow	✗	✓	0.01	0.01	0.05	0.06	0.01	0.05	0.04	0.05	0.05	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GMStereo-*vpp** [11]	Mixdata	✗	✓	0.00	0.03	0.02	0.02	0.00	0.07	0.03	0.05	0.06	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CFNet-*vpp** [9]	Sceneflow	✗	✓	0.12	0.11	0.10	0.09	0.23	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CFNet-*vpp** [9]	Middlebury	✗	✓	0.04	0.03	0.02	0.03	0.02	0.01	0.01	0.01	0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
HSMNet-*vpp* [12]	Middlebury	✗	✓	0.02	0.00	0.01	0.02	0.00	0.06	0.01	0.01	0.00	0.01	0.14	0.02	0.03	0.02	0.01	0.00	0.00	0.00	0.00	0.00
CREStereo-*vpp** [3]	ETH3D	✗	✓	0.04	0.02	0.04	0.04	0.01	0.04	0.04	0.06	0.05	0.03	0.05	0.03	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
PSMNet-*vpp* [1]	Sceneflow	✗	✓	0.07	0.06	0.05	0.04	0.02	0.01	0.00	0.01	0.00	0.00	0.03	0.04	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00
rSGM-*vpp* [7]	-	✗	✓	0.01	0.02	0.00	0.01	0.09	0.00	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00

Table III: **Stability of random pattern of type (vi).** We present an analysis of the standard deviation of type (vi) virtual pattern based on a sample size of 5. Our findings indicate that, despite the non-deterministic nature of this pattern, the observed errors exhibit stability. * uses $\alpha = 0.2$ for blending.

3. Stability of random pattern

As stated in the main paper, the random patch pattern (vi) proposed and utilized is a non-deterministic technique. To assess the stability of this approach concerning artifacts, we conduct five inference runs to compute the standard deviation for various stereo methods, as many as we used in the main paper to calculate the mean errors. Our results, as shown in Table III, demonstrate that despite the non-deterministic nature of the approach, errors remain relatively constant. This evidence indicates that this pattern can be effectively employed in diverse environments and attain stable results. Specifically, the standard deviation remains below 0.14% for any error rate threshold and below 0.1 px for avg. error, except for CFNet [9] being slightly less stable (0.23 px on Midd-14).

4. Additional implementation details

When evaluating the accuracy of stereo algorithms and networks with and without VPP, different resizing strategies were applied to the input images, often following the suggestions of the original authors.

RAFT-Stereo [4] and HSMNet [12] were evaluated using the original image resolution on Middlebury 2014 [7], Middlebury 2021 [10], ETH3D [8], and KITTI [5] datasets. CFNet [9], rSGM [7], and PSMNet [1] were evaluated using half the original image resolution on Middlebury 2014 and Middlebury 2021 (because of memory limitations). GMStereo [11] and CREStereo [3] on Middlebury 2014 and Middlebury 2021 were evaluated with image resizing to 1024×1536. GMStereo on ETH3D was evaluated with image resizing to 512×768, while CREStereo on ETH3D was evaluated with image resizing to 768×1024. GMStereo on KITTI was evaluated with image resizing to 352×1216, while CREStereo on KITTI was evaluated with image resizing to 1024×1536.

5. Qualitative results

We conclude by showing some additional qualitative results to support the effectiveness of VPP further.

Middlebury 2021. Fig. I shows a scene from the Middlebury 2021 dataset, in which we sample 5% points from the ground-truth to project our virtual pattern. We test four stereo networks trained on synthetic data and the rSGM algorithm. We can notice how the original disparity maps suffer from several artifacts and, in most cases, very high error rates. Only RAFT-Stereo achieves reasonable results. In contrast, by enabling VPP, the results by any method fall below 13% error rate, with RAFT-Stereo achieving a bad2 error lower than 2%. To better highlight the effect of VPP on images, we show both the original and hallucinated frames in Fig. II.

KITTI 2015. Fig. III shows an example from the KITTI 2015 training set. Specifically, we select the *000104* stereo pair, being known by practitioners in the field for its very low exposure. Since raw LiDAR measurements are not available for this specific pair, we sample again 5% points from ground-truth. We can appreciate how, even in this prohibitive environment, VPP can largely improve the results of stereo networks and the rSGM algorithm. To better highlight the effect of VPP on images, we show both the original and hallucinated frames in Fig. IV.

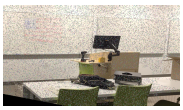
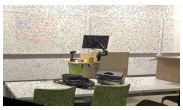
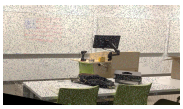
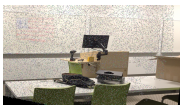
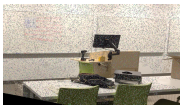
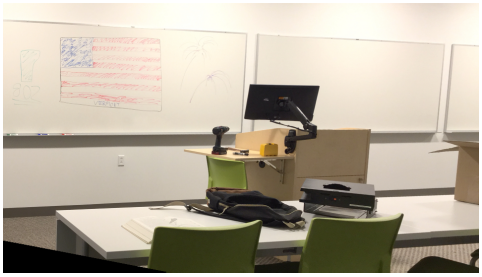
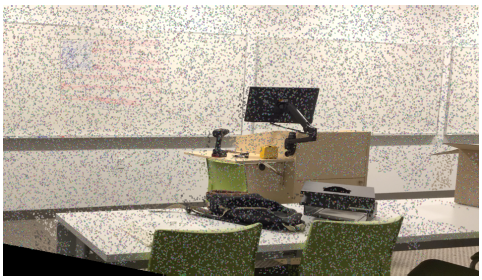
Model	RGB left	RGB right	Disparity map map Error Rate (> 2) %	VPP left	VPP right	VPP disparity map Error Rate (> 2) %
PSMNet [1]			 74.36			 7.16
RAFT-Stereo [4]			 15.91			 1.94
CFNet* [9]			 77.76			 11.59
GMStereo* [11]			 81.40			 8.65
rSGM [7]			 81.31			 12.70

Figure I: **Cross-domain generalization enhancement: qualitative results on indoor.** All networks are trained on synthetic data and tested on an unseen scenario (Midd-21, scene podium1): our VPP approach significantly improves cross-domain generalization compared to the vanilla baselines. Additionally, traditional stereo algorithms, such as rSGM, also benefit from our approach. * uses $\alpha = 0.2$ for blending.

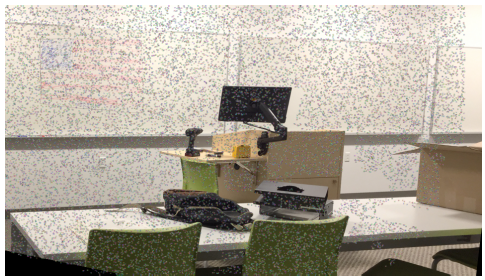


RGB Left

RGB Right



VPP Left



VPP Right

Figure II: **Qualitative view of virtual pattern in indoor scenario.** Midd-21, scene podium1

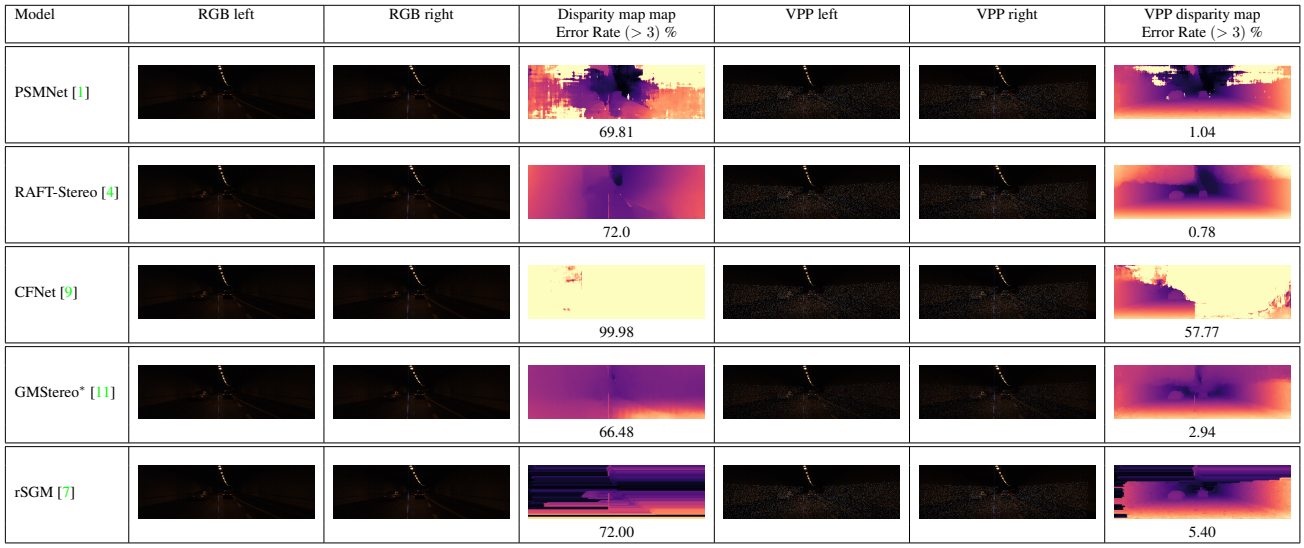


Figure III: **Cross-domain generalization enhancement: qualitative results on outdoor.** All networks are trained on synthetic data and tested on an unseen scenario (KITTI 2015, scene 104): our VPP approach significantly improves cross-domain generalization compared to the vanilla baselines, except for CFNet which achieves mild results. Additionally, traditional stereo algorithms, such as rSGM, also benefit from our approach. * uses $\alpha = 0.2$ for blending.

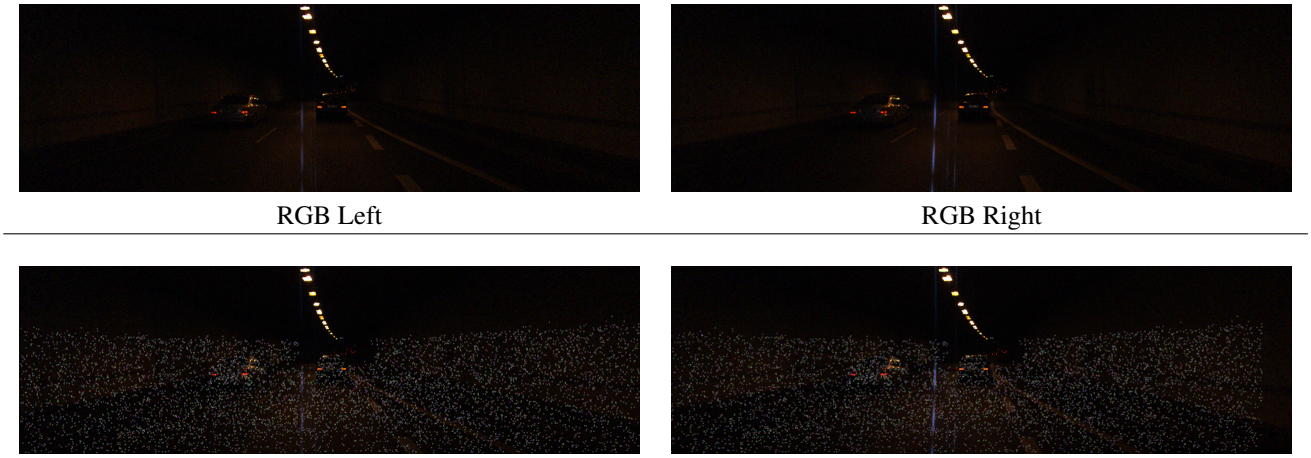


Figure IV: **Qualitative view of virtual pattern in outdoor scenario.** KITTI 2015, scene 104

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5418, 2018. 2, 3, 4, 5

[2] Xuelian Cheng, Yiran Zhong, Yuchao Dai, Pan Ji, and Hongdong Li. Noise-aware unsupervised deep lidar-stereo fusion. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. 2

[3] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 3

[4] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, 2021. 2, 3, 4, 5

[5] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[6] Matteo Poggi, Davide Pallotti, Fabio Tosi, and Stefano Mattoccia. Guided stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 979–988, 2019. 2

[7] 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. 2, 3, 4, 5

[8] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[9] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, June 2021. 3, 4, 5

[10] Robert Spangenberg, Tobias Langner, Sven Adfeldt, and Raúl Rojas. Large scale semi-global matching on the cpu. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 195–201. IEEE, 2014. 1, 2, 3

[11] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *arXiv preprint arXiv:2211.05783*, 2022. 3, 4, 5

[12] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 3