Supplementary Material : PSMNet-FusionX3 : LiDAR-Guided Deep Learning Stereo Dense Matching On Aerial Images

Teng Wu, Bruno Vallet, Marc Pierrot-Deseilligny
LASTIG, Univ Gustave Eiffel, IGN-ENSG
firstname.lastname@ign.fr

1. Detail for Equation (1)

In the main paper, the a_1, b_1, c_1, and other variables may not be clearly listed due to the length and readability of the paper. Therefore, we have provided a clear list of these variables below.

$$\begin{aligned}
L &= x_1y_2 + x_3y_1 + x_2y_3 - x_3y_2 - x_1y_3 - x_2y_1 \\
a_1 &= \frac{y_2 - y_3}{L} \\
a_2 &= \frac{y_3 - y_1}{L} \\
a_3 &= \frac{y_1 - y_2}{L} \\
b_1 &= \frac{x_3 - x_2}{L} \\
b_2 &= \frac{x_1 - x_3}{L} \\
b_3 &= \frac{x_2 - x_1}{L} \\
c_1 &= \frac{x_2y_3 - x_3y_2}{L} \\
c_2 &= \frac{x_3y - 1 - x_1y_3}{L} \\
c_3 &= \frac{x_1y - 2 - x_2y_1}{L}
\end{aligned} \tag{1}$$

2. Experiment data introduction

2.1. Toulouse2020 dataset

Since the DublinCity dataset has been detailedly introduced in [4], we will provide detailed information about the Toulouse2020 dataset. The Toulouse2020 dataset was collected by IGN (French Mapping Agency), in the urban area of Toulouse using a Vexcel camera with a focal length of 120mm and a flight height of 1676m on June 30th, 2020. The overlap along the flight is about 82%, and the cross-flight overlap is about 70%. The data is projected using the Lambert-93 system. The area used for the experiment is shown in Figure 1. Additionally, the LiDAR point cloud data was collected during the same flight.

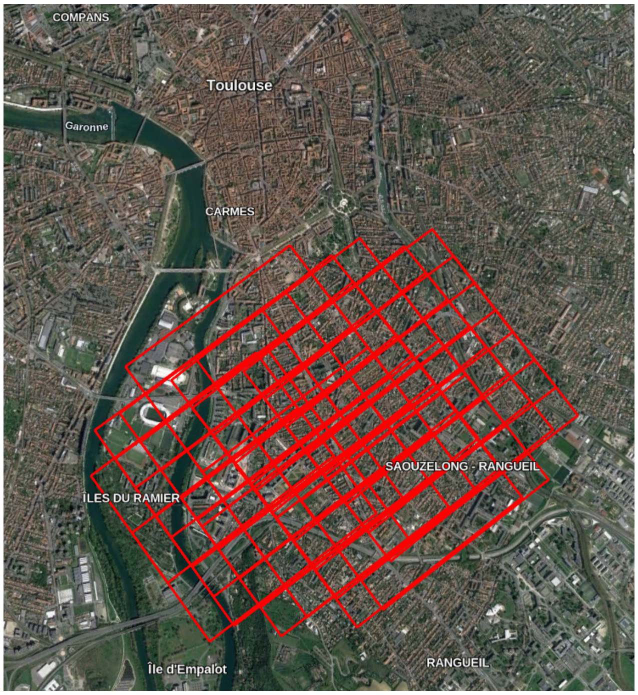
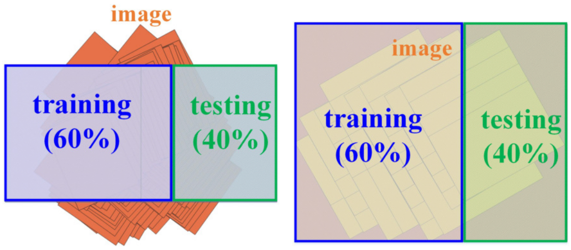


Figure 1. The coverage of Toulouse2020 dataset on Google Earth, the red rectangle are the image frames.

2.2. Dataset generation

The training data is generated automatically using the method described in [7]. To facilitate training and testing, the data is segregated based on geographical location, as depicted in Figure 2. The division is established by the LiDAR point cloud, with training data encompassing 60% of the entire region and testing data accounting for the remaining 40%. Precisely, the training and testing datasets correspond to distinct regions within the same city.



(a) DublinCity

(b) Toulous2020

Figure 2. Illustration of dataset partitioning based on LiDAR area splitting and image frames. The geographical-based LiDAR areas utilized for dividing the training and testing datasets are highlighted.

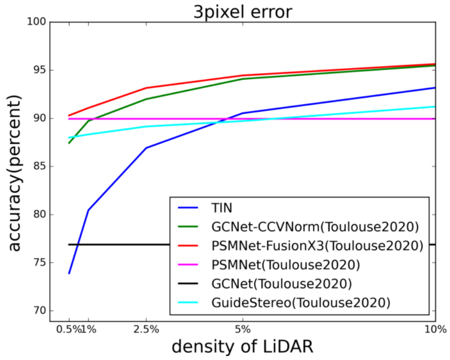


Figure 3. : This figure shows the influence of LiDAR density on the 3-pixel error for the Toulouse2020 dataset. When using image-only deep image matching (DIM), the PSMNet outperforms the GCNet. However, GuideStereo shows improvement only at a 10% density ratio. The TIN approach introduces more errors, but is better than PSMNet at a high-density ratio of 5%. For very low density (0.5%), the GCNet-CCVNorm outperforms GuideStereo and the image-only DIM. Our proposed PSMNet-FusionX3 is consistently better than GCNet-CCVNorm and is the only guided method that outperforms PSMNet at the lowest density of 0.5%.

3. Detail of the experiment

In the main paper, the figures only show the 1-pixel error for all ratios of LiDAR guidance, which makes it easier to draw conclusions on the influence of density at a glance. While the 1-pixel error reflects the precision of the methods, the 3-pixel error indicates their robustness. Further details will be presented later. However, the difference between the error values cannot be read from the figures. Therefore, we provide the quantity of errors in the following.

3.1. Density analysis

Figure 3 shows the 3-pixel error on the Toulouse2020 dataset. The deep learning (DL) based methods exhibit superior performance compared to the 1-pixel error, which suggests the smoothness of these methods.

Figure 4 illustrates the 3-pixel error on the DublinCity dataset, which demonstrates a similar phenomenon to the Toulouse2020 result.

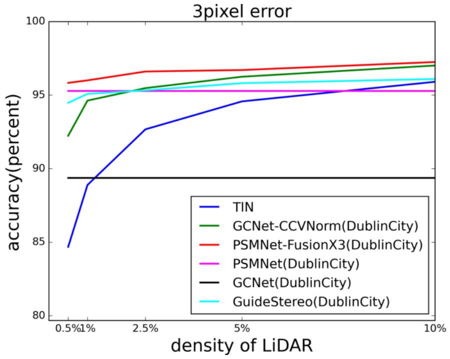


Figure 4. Influence of LiDAR density on 3-pixel error on DublinCity. The behavior is similar to Toulouse2020, GCNet performs better than TIN below 1%. GCNet-CCVNorm only have advantages at large ratio, our PSMNet-FusionX3 is always better than GCNet-CCVNorm.

3.2. Toulouse2020

The detailed analysis of pixel error on the experiment on Toulouse2020 is presented in the Table 1, where the 1,2,3,5, and 9-pixel errors are listed. From the table, we can observe that a 1-pixel error indicates accuracy, while a 3-pixel error indicates robustness. SGM [2] performs well on the 3-pixel error metric. Even when using triangulated irregular network (TIN) interpolation, the results are not bad at a ratio of 0.5%. However, the performance of DL-based methods heavily relies on the training data. PSMNet [1] outperforms GCNet [3], but its performance is more dependent on the training data. On the other hand, GCNet-CCVNorm [6] does not work well when the ratio is low. However, when the ratio increases, PSMNet-FusionX3 does not outperform GCNet-CCVNorm much. Using sparse disparity guidance to train PSMNet, the result is slightly worse than using the dense ground truth.

3.3. DublinCity

DublinCity is a high-resolution image dataset [4], and the dataset is located in the center of the city, making it less challenging than other areas. We can observe that SGM's performance with a 3-pixel error is already good, and PSMNet achieves excellent results. Even with low-density LiDAR, PSMNet-FusionX3 outperforms PSMNet. However, the results of GuideStereo and GCNet-CCVNorm deteriorate with low-density LiDAR. When the LiDAR density is high, PSMNet-FusionX3 does not outperform GCNet-CCVNorm significantly. Using sparse disparity guidance to train PSMNet, the result is nearly the same as using the dense ground truth.

method	**ratio**	**training**	**<1-pixel**	**<2-pixel**	**<3-pixel**	**<5-pixel**	**<9-pixel**
SGM	–	–	**48.66**	67.73	**75.85**	83.17	88.73
GCNet	–	Toulous2020	47.21	67.59	76.89	85.04	90.84
PSMNet	–	Toulous2020	**73.78**	85.48	**89.95**	94.22	97.52
GCNet	–	*DublinCity*	46.59	67.67	76.97	84.16	89.77
PSMNet	–	*DublinCity*	56.83	74.79	81.20	86.92	91.92
TIN	0.5%	–	58.33	67.61	73.91	82.16	90.22
GuideStereo	0.5%	Toulous2020	69.28	82.95	88.01	92.64	96.37
GCNet-CCVNorm	0.5%	Toulous2020	**70.57**	82.51	**87.45**	92.17	96.12
PSMNet-FusionX3	0.5%	Toulous2020	**76.16**	86.09	**90.30**	94.36	97.57
PSMNet	0.5%	Toulous2020[+]	69.48	83.04	88.00	92.55	96.25
GuideStereo	0.5%	*DublinCity*	57.80	74.90	81.43	87.38	92.48
GCNet-CCVNorm	0.5%	*DublinCity*	**65.97**	79.84	**85.06**	90.10	94.52
PSMNet-FusionX3	0.5%	*DublinCity*	**71.15**	82.33	**87.21**	92.06	96.18
TIN	1%	–	65.95	74.77	80.46	87.46	93.72
GuideStereo	1%	Toulous2020	70.70	83.36	88.33	92.92	96.60
GCNet-CCVNorm	1%	Toulous2020	76.32	85.77	89.73	93.68	97.13
PSMNet-FusionX3	1%	Toulous2020	78.28	87.13	91.08	94.93	97.93
PSMNet	1%	Toulous2020[+]	70.35	83.44	88.26	92.70	96.31
GuideStereo	1%	*DublinCity*	61.74	77.53	83.39	88.84	93.61
GCNet-CCVNorm	1%	*DublinCity*	71.70	83.38	87.98	92.57	96.59
PSMNet-FusionX3	1%	*DublinCity*	74.96	84.49	88.89	93.41	97.10
TIN	2.5%	–	74.47	82.23	86.92	92.31	96.60
GuideStereo	2.5%	Toulous2020	73.03	84.54	89.16	93.56	97.06
GCNet-CCVNorm	2.5%	Toulous2020	80.56	88.58	92.00	95.35	98.11
PSMNet-FusionX3	2.5%	Toulous2020	82.64	89.89	93.16	96.34	98.76
PSMNet	2.5%	Toulous2020[+]	70.95	83.94	88.67	93.01	96.51
GuideStereo	2.5%	*DublinCity*	67.17	79.81	85.04	90.28	94.83
GCNet-CCVNorm	2.5%	*DublinCity*	80.05	88.34	91.78	95.11	97.95
PSMNet-FusionX3	2.5%	*DublinCity*	80.16	87.84	91.49	95.25	98.20
TIN	5%	–	79.90	86.66	90.54	94.80	97.96
GuideStereo	5%	Toulous2020	73.51	85.16	89.72	93.96	97.30
GCNet-CCVNorm	5%	Toulous2020	**85.58**	91.57	**94.10**	96.65	98.90
PSMNet-FusionX3	5%	Toulous2020	**85.95**	91.83	**94.46**	97.08	99.16
PSMNet	5%	Toulous2020[+]	70.83	84.07	88.78	93.09	96.58
GuideStereo	5%	*DublinCity*	71.35	82.07	86.72	91.53	95.69
GCNet-CCVNorm	5%	*DublinCity*	84.04	90.83	93.64	96.39	98.70
PSMNet-FusionX3	5%	*DublinCity*	84.93	91.09	93.91	96.77	98.99
TIN	10%	–	84.49	90.09	93.18	96.49	98.81
GuideStereo	10%	Toulous2020	78.92	87.39	91.21	94.96	97.91
GCNet-CCVNorm	10%	Toulous2020	88.13	93.38	95.48	97.58	99.37
PSMNet-FusionX3	10%	Toulous2020	88.72	93.59	95.64	97.73	99.45
PSMNet	10%	Toulous2020[+]	**71.31**	84.04	88.70	93.01	96.49
GuideStereo	10%	*DublinCity*	74.27	83.99	88.32	92.79	96.59
GCNet-CCVNorm	10%	*DublinCity*	87.76	93.21	95.36	97.49	99.32
PSMNet-FusionX3	10%	*DublinCity*	88.01	93.24	95.49	97.75	99.46

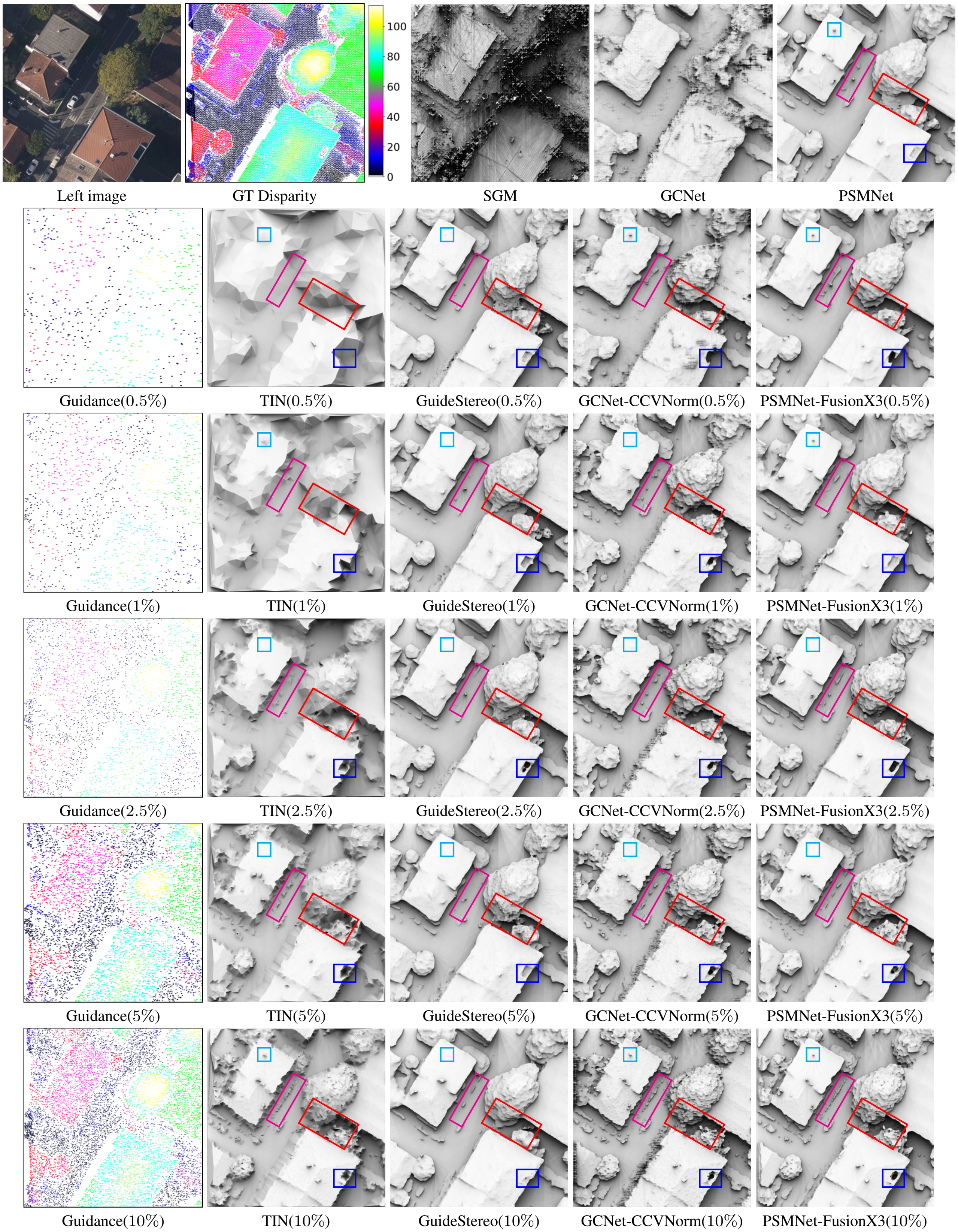
[+] Training with the sparsely sampled disparity as ground truth.

Table 1. The n-pixel error on the Toulouse2020. DL-based methods give a better result than SGM. Despite GCNet's poor performance, the impact of training data on its performance is not significant. However, PSMNet's performance is more dependent on the training data. By using LiDAR as guidance, the PSMNet-FusionX3 model trained on the DublinCity dataset achieves impressive results.

method	ratio	training	<1-pixel	<2-pixel	<3-pixel	<5-pixel	<9-pixel
SGM	–	–	**66.85**	81.54	**86.76**	91.19	94.40
GCNet	–	DublinCity	**66.17**	83.06	89.36	93.62	96.65
PSMNet	–	DublinCity	**84.43**	92.75	**95.28**	97.35	98.98
GCNet	–	*Toulous2020*	**59.66**	80.15	87.04	92.13	95.57
PSMNet	–	*Toulous2020*	75.60	88.23	92.12	95.23	97.65
TIN	0.5%	–	70.82	79.64	84.68	90.24	94.87
GuideStereo	0.5%	DublinCity	**81.45**	91.42	94.47	96.93	98.76
GCNet-CCVNorm	0.5%	DublinCity	**78.54**	88.73	92.24	95.23	97.64
PSMNet-FusionX3	0.5%	DublinCity	**87.11**	93.67	95.83	97.71	99.22
PSMNet	0.5%	DublinCity[+]	83.95	92.52	95.11	97.24	98.93
GuideStereo	0.5%	*Toulous2020*	74.95	87.80	91.79	95.03	97.53
GCNet-CCVNorm	0.5%	*Toulous2020*	80.23	89.88	93.26	96.13	98.36
PSMNet-FusionX3	0.5%	*Toulous2020*	83.13	91.38	94.23	96.69	98.67
TIN	1%	–	77.03	84.70	88.89	93.35	96.82
GuideStereo	1%	DublinCity	**83.94**	92.44	95.08	97.28	98.97
GCNet-CCVNorm	1%	DublinCity	70.87	84.13	89.49	93.92	96.98
PSMNet-FusionX3	1%	DublinCity	87.87	93.92	96.00	97.85	99.31
PSMNet	1%	DublinCity[+]	84.30	92.59	95.14	97.26	98.94
GuideStereo	1%	*Toulous2020*	74.64	87.59	91.66	94.94	97.52
GCNet-CCVNorm	1%	*Toulous2020*	84.43	91.78	94.43	96.83	98.81
PSMNet-FusionX3	1%	*Toulous2020*	85.17	92.25	94.80	97.07	98.89
TIN	2.5%	–	83.32	89.46	92.67	95.97	98.35
GuideStereo	2.5%	DublinCity	85.12	92.83	95.32	97.42	99.07
GCNet-CCVNorm	2.5%	DublinCity	86.99	93.20	95.48	97.49	99.15
PSMNet-FusionX3	2.5%	DublinCity	90.13	94.87	96.60	98.23	99.53
PSMNet	2.5%	DublinCity[+]	84.39	92.71	95.24	97.32	98.97
GuideStereo	2.5%	*Toulous2020*	79.50	89.90	93.26	96.03	98.23
GCNet-CCVNorm	2.5%	*Toulous2020*	85.98	92.65	95.15	97.38	99.13
PSMNet-FusionX3	2.5%	*Toulous2020*	88.67	93.92	95.95	97.88	99.40
TIN	5%	–	86.88	91.96	94.57	97.20	99.01
GuideStereo	5%	DublinCity	87.18	93.62	95.81	97.72	99.23
GCNet-CCVNorm	5%	DublinCity	**89.13**	94.33	96.25	98.01	99.43
PSMNet-FusionX3	5%	DublinCity	**90.87**	95.05	96.70	98.32	99.58
PSMNet	5%	DublinCity[+]	84.46	92.74	95.27	97.34	98.98
GuideStereo	5%	*Toulous2020*	79.36	90.18	93.58	96.32	98.42
GCNet-CCVNorm	5%	*Toulous2020*	89.91	94.49	96.26	97.98	99.47
PSMNet-FusionX3	5%	*Toulous2020*	90.52	94.80	96.54	98.25	99.60
TIN	10%	–	89.66	93.80	95.90	98.01	99.42
GuideStereo	10%	DublinCity	88.22	94.05	96.09	97.91	99.35
GCNet-CCVNorm	10%	DublinCity	91.56	95.52	97.01	98.44	99.66
PSMNet-FusionX3	10%	DublinCity	92.30	95.85	**97.25**	98.64	99.74
PSMNet	10%	DublinCity[+]	**84.53**	92.79	95.30	97.37	99.00
GuideStereo	10%	*Toulous2020*	84.55	92.01	94.71	97.09	98.93
GCNet-CCVNorm	10%	*Toulous2020*	90.64	95.06	96.74	98.32	99.64
PSMNet-FusionX3	10%	*Toulous2020*	91.86	95.81	**97.25**	98.62	99.74

[+] Training with the sparsely sampled disparity as ground truth.

Table 2. The n-pixel error on the DublinCity. The performance of GCNet trained on the same dataset is much better than when trained on the Toulouse2020 dataset. PSMNet achieves good results, while GuideStereo performs worse than PSMNet, especially when the ratio is low. With a low ratio, PSMNet-FusionX3 significantly outperforms GCNet-CCVNorm. However, when the ratio exceeds 5%, the difference in performance between PSMNet-FusionX3 and GCNet-CCVNorm diminishes. When the ratio is high, the performance of PSMNet-FusionX3 trained on the Toulouse2020 dataset is nearly equivalent to that of the model trained on the DublinCity dataset.



Left image | GT Disparity | SGM | GCNet | PSMNet

Guidance(0.5%) | TIN(0.5%) | GuideStereo(0.5%) | GCNet-CCVNorm(0.5%) | PSMNet-FusionX3(0.5%)

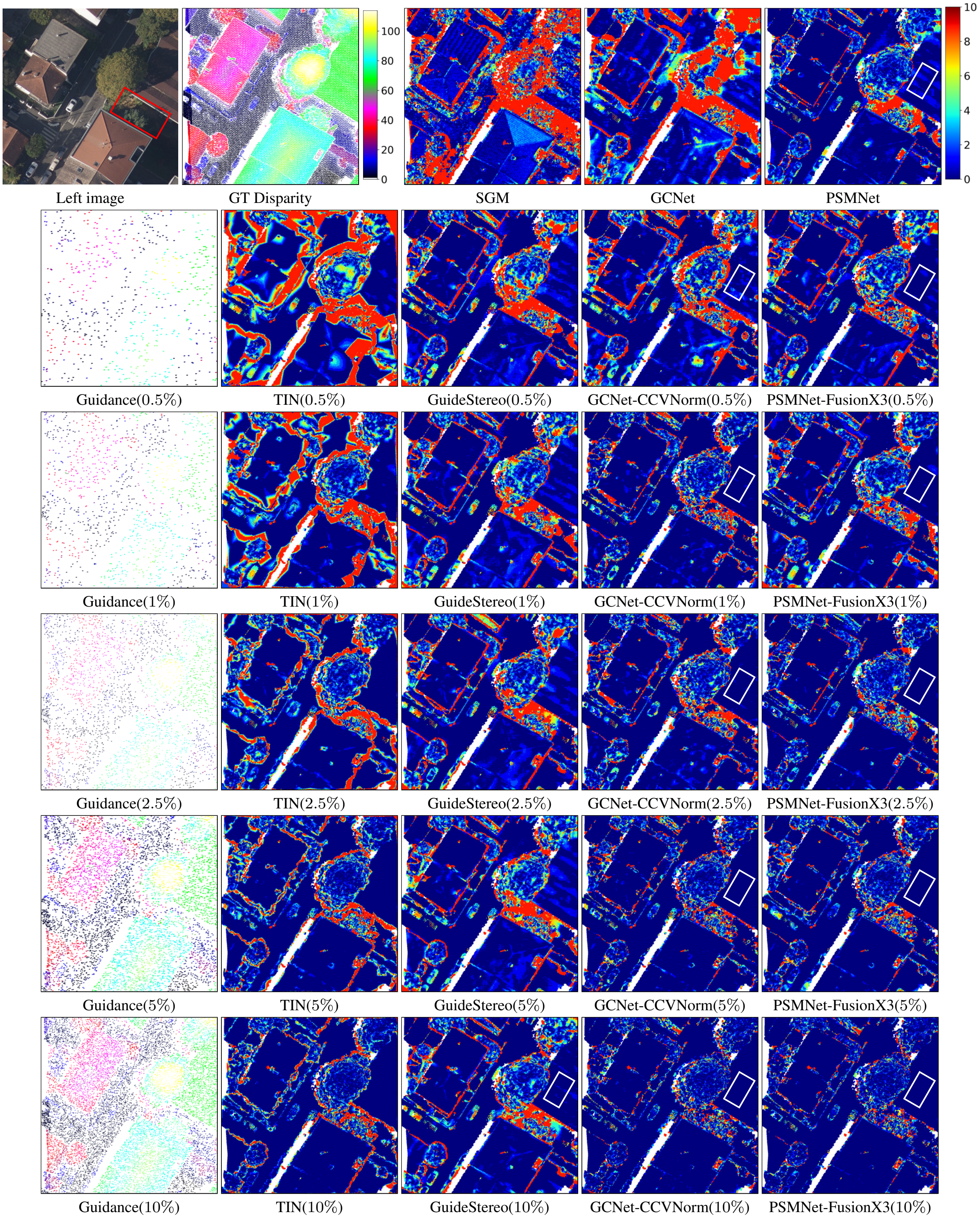
Guidance(1%) | TIN(1%) | GuideStereo(1%) | GCNet-CCVNorm(1%) | PSMNet-FusionX3(1%)

Guidance(2.5%) | TIN(2.5%) | GuideStereo(2.5%) | GCNet-CCVNorm(2.5%) | PSMNet-FusionX3(2.5%)

Guidance(5%) | TIN(5%) | GuideStereo(5%) | GCNet-CCVNorm(5%) | PSMNet-FusionX3(5%)

Guidance(10%) | TIN(10%) | GuideStereo(10%) | GCNet-CCVNorm(10%) | PSMNet-FusionX3(10%)

Figure 5. An example of the disparity result of Toulouse2020 shading using MicMac. The cyan rectangle highlights the detail area, where the performance of PSMNet-FusionX3 depends on the input guidance. The red rectangle represents the shadow area, where PSMNet produces unsatisfactory results, but utilizing LiDAR as guidance can enhance the outcome. The blue rectangle corresponds to the disparity discontinuity area, where both GCNet-CCVNorm and PSMNet-FusionX3 can preserve the discontinuity. Finally, the magenta rectangle represents small details, and with an increase in the ratio, more and more detail can be reconstructed.



Left image GT Disparity SGM GCNet PSMNet

Guidance(0.5%) TIN(0.5%) GuideStereo(0.5%) GCNet-CCVNorm(0.5%) PSMNet-FusionX3(0.5%)

Guidance(1%) TIN(1%) GuideStereo(1%) GCNet-CCVNorm(1%) PSMNet-FusionX3(1%)

Guidance(2.5%) TIN(2.5%) GuideStereo(2.5%) GCNet-CCVNorm(2.5%) PSMNet-FusionX3(2.5%)

Guidance(5%) TIN(5%) GuideStereo(5%) GCNet-CCVNorm(5%) PSMNet-FusionX3(5%)

Guidance(10%) TIN(10%) GuideStereo(10%) GCNet-CCVNorm(10%) PSMNet-FusionX3(10%)

Figure 6. The correspondence error map of the disparity result. PSMNet outperforms SGM and GCNet which produce larger errors. The shadow area in the left image is particularly challenging, with the most errors occurring there, PSMNet and GuideStereo perform poorly in this area. Similarly, the roof in the shadow also presents a challenge, and both PSMNet and GuideStereo struggle to produce accurate results. GuideStereo achieves good results only when the error ratio is 10%, whereas GCNet-CCVNorm and PSMNet-FusionX3 produce much fewer errors when the ratio is 1%.

4. Visual assessment

In the main paper, an example of DublinCity is presented. Here, we will provide an example from the Toulouse2020 dataset, and the disparity result is displayed in Figure 5, while the correspondence error map is shown in Figure 6.

For visualizing the disparity maps, an ambient occlusion shading technique implemented in MicMac [5] was used. This method is particularly effective in highlighting stereo matching in remote sensing. Due to the presence of shadows, SGM performs poorly and results in reconstruction of the building roof. GCNet performs slightly better than SGM but is still not sufficient, while PSMNet produces an outstanding outcome. The cyan rectangle represents a detail area where GuideStereo fails, and PSMNet-FusionX3 performance depends on the input guidance. At ratios of 0.5%, 0.1% and 10%, the points are present, and PSMNet-FusionX3 can reconstruct them. However, in ratios of 2.5% and 5%, it fails. The red rectangle shows the shadow area, where PSMNet performs poorly, but LiDAR guidance can enhance the results. The blue rectangle represents the disparity discontinuity area, where GCNet-CCVNorm and PSMNet-FusionX3 can preserve the discontinuity. The magenta rectangle represents small details, where with an increase in the ratio, more LiDAR guidance is selected, and more detail can be reconstructed.

The error map displayed in Figure 6 reveals the performance of different stereo matching methods. SGM and GC-Net exhibit a large error, although the latter has fewer errors than the former. On the other hand, PSMNet demonstrates much better performance than GCNet. In the left image, the red rectangle corresponds to the shadow area, which is where most errors occur. In this region, both PSMNet and GuideStereo perform poorly. Even when the ratio is 5% and the error is small, GuideStereo still cannot fuse the LiDAR thoroughly. The white rectangle represents the roof in the shadow, and PSMNet and GuideStereo exhibit poor performance. Only when the ratio is 10%, the result of GuideStereo is good. However, a ratio of 2.5% of TIN is sufficient. Finally, when the ratio is 1%, the error of both GCNet-CCVNorm and PSMNet-FusionX3 is significantly lower.

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[2] Daniel Hernandez-Juarez, Alejandro Chacón, Antonio Espinosa, David Vázquez, Juan Carlos Moure, and Antonio M López. Embedded real-time stereo estimation via semi-global matching on the gpu. *Procedia Computer Science*, 80:143–153, 2016.

[3] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[4] Debra F Laefer, Saleh Abuwarda, Anh-Vu Vo, Linh Truong-Hong, and Hamid Gharibi. 2015 aerial laser and photogrammetry survey of dublin city collection record. 2017.

[5] Marc Pierrot-Deseilligny, D Jouin, J Belvaux, G Maillet, L Girod, E Rupnik, J Muller, M Daakir, G Choqueux, and M Deveau. Micmac, apero, pastis and other beverages in a nutshell. *Institut Géographique National*, 2014.

[6] Tsun-Hsuan Wang, Hou-Ning Hu, Chieh Hubert Lin, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5895–5902. IEEE, 2019.

[7] T Wu, B Vallet, M Pierrot-Deseilligny, and E Rupnik. A new stereo dense matching benchmark dataset for deep learning. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:405–412, 2021.