Deep vanishing point detection: Geometric priors make dataset variations vanish Supplementary material

Yancong Lin, Ruben Wiersma, Silvia L. Pintea, Klaus Hildebrandt, Elmar Eisemann, and Jan C. van Gemert
Delft University of Technology, The Netherlands

1. Multi-scale sampling on the Gaussian sphere

Inspired by [8], we use a multi-scale sampling strategy to detect three orthogonal vanishing points in the Manhattan world. We start by uniformly sampling $N_{s=0}$ points at scale $s = 0$ on the entire hemisphere. We input these points into a spherical convolution network. Sequentially, we use the Manhattan assumption to choose 3 orthogonal vanishing points with the highest confidence as anchors. We uniformly sample $N_{s=1}$ points around each anchor in a local neighborhood defined by the radius $\delta_{s=1}$ at scale $s = 1$. Then, we feed these newly sampled points into a spherical convolution network. Finally, the point with highest confidence in each local neighborhood is considered as the anchor for sampling at the $(s+1)$th scale. Specifically, we set $\delta \approx \{90°, 13°, 4°\}$ and $N = \{512, 128, 128\}$. The spherical convolution networks share the same architecture while processing different number of samples. During training, we assign the nearest neighbors to the ground truth as positive samples while the others are considered as negative samples. We compute the cross-entropy losses averaged over positives and negatives respectively, at each scale.

2. Datasets

Tab. 1 shows a detailed comparison among all datasets, and Fig. 1 displays image examples from each dataset. The SU3 dataset is synthetic and all images are well-calibrated with sharp edges. The ScanNet dataset captures indoor scenes in the real-world environments, where image content varies significantly. The YUD dataset captures both indoor and outdoor scenes in urban cities and contains only 102 images. SU3, ScanNet and YUD datasets follow the Manhattan world assumption where there are 3 orthogonal vanishing points. In comparison, the NYU Depth dataset has a varying number of instances across images. Moreover, there are 1449 images in total, and therefore training deep networks is highly challenging on the NYU Depth dataset due to the lack of data.

3. Visualizations

We visualize predictions on the NYU Depth dataset in Fig. 2. We show the input images, labeled line segments and detected vanishing points on the hemisphere. Each color represents a group of lines and their corresponding vanishing point. In the third row, our model correctly detects all vanishing points, as the colored $\times$ and $\circ$ overlap. In comparison, CONSAC fails to localize the red one and J-Linkage is unable to detect the green one. In addition, CONSAC makes nearby predictions: e.g., the blue and pink $\times$ markers in second row. This is caused by the LSD [7] method producing a large number of outlier segments, resulting in incorrect predictions. Our method is suitable for real-world scenarios, where the image content varies substantially.

Fig. 3 and Fig. 4 show detected vanishing points from our model on the SU3 and YUD datasets, respectively. Since all methods make reasonably good prediction and the difference is hardly visible, we only visualize our results.

Fig. 5 compares detected vanishing points from all models on the ScanNet dataset. We compare all models in a column-wise manner, where the input image is on the top, while predictions from each method is displayed sequentially. We show the top 3 vanishing points for J-Linkage [3] and CONSAC [4]. In general, NeurVPS and ours are able to localize vanishing points more precisely than other non-learning approaches. As shown in the fourth example where the object is not orthogonally placed, Quasi-VP [5] fails due to the presence of strong outliers and the lack of inliers. This shows the disadvantage of non-learning method in dealing with complicated real-world scenarios. J-Linkage and CONSAC sometimes predict vanishing points far away from the ground truth (e.g., the fourth example), because they are originally designed for multiple vanishing point detection in non-Manhattan world, and do not enforce orthogonality explicitly. Ours show better performance in detecting orthogonal vanishing points from complex scenes thanks to the ability to learn semantic features from images directly in an end-to-end manner.

Datasets	Images	Manhattan	Resolution	number of VPs	Training	Validation	Testing
SU3 [9]	Synthetic	✓	512 × 512	3	18 400	2 300	2 300
ScanNet [1]	Real-world	✓	512 × 512	3	189 916	500	20 942
YUD [2]	Real-world	✓	480 × 640	3	25	-	77
NYU Depth [6]	Real-world	×	480 × 640	1-8	1000	224	225

Table 1. **Comparison of the four datasets.** The SU3, ScanNet and YUD datasets follow the Manhattan assumption with 3 orthogonal vanishing points, while the NYU Depth dataset is annotated with a varying number of instances. In addition, the size of the four datasets varies substantially. There are only 1000 and 25 training images in NYU and YUD datasets.

Figure 1. **Examples from the SU3, ScanNet, YUD and NYU Depth (labeled with ground truth lines) datasets.** Images from the SU3 dataset are well-calibrated with clear geometric cues, such as sharp edges and contours. In contrast, the other datasets capture real-world images where image content varies significantly. The NYU Depth dataset is labeled with multiple vanishing points (varying from 1-8).

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 2

[2] Patrick Denis, James H Elder, and Francisco J Estrada. Efficient edge-based methods for estimating manhattan frames in urban imagery. In *European conference on computer vision*, pages 197–210. Springer, 2008. 2

[3] Chen Feng, Fei Deng, and Vineet R Kamat. Semi-automatic 3d reconstruction of piecewise planar building models from single image. *CONVR (Sendai:)*, 2010. 1

[4] Florian Kluger, Eric Brachmann, Hanno Ackermann, Carsten Rother, Michael Ying Yang, and Bodo Rosenhahn. Consac: Robust multi-model fitting by conditional sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4643, 2020. 1

[5] Haoang Li, Ji Zhao, Jean-Charles Bazin, Wen Chen, Zhe Liu, and Yun-Hui Liu. Quasi-globally optimal and efficient vanishing point estimation in manhattan world. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1646–1654, 2019. 1

[6] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[7] Rafael Grompone Von Gioi, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):722–732, 2008. 1

[8] Yichao Zhou, Haozhi Qi, Jingwei Huang, and Yi Ma. Neurvps: Neural vanishing point scanning via conic convolution. In *Advances in Neural Information Processing Systems*, pages 866–875, 2019. 1

[9] Yichao Zhou, Haozhi Qi, Yuexiang Zhai, Qi Sun, Zhili Chen, Li-Yi Wei, and Yi Ma. Learning to reconstruct 3d manhattan wireframes from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7698–7707, 2019. 2

Figure 2. **Visualizations on the NYU Depth dataset.** The black ○ represents the ground truth, while the colored × indicates predictions. Each color corresponds to a set of lines and their related vanishing point. Our model is better at localizing multiple vanishing points in the non-Manhattan world, having predictions (colored cross ×) closer to the ground truth (black ○), while the predictions of the other methods scatter away from the ground truth, as shown in the first example.

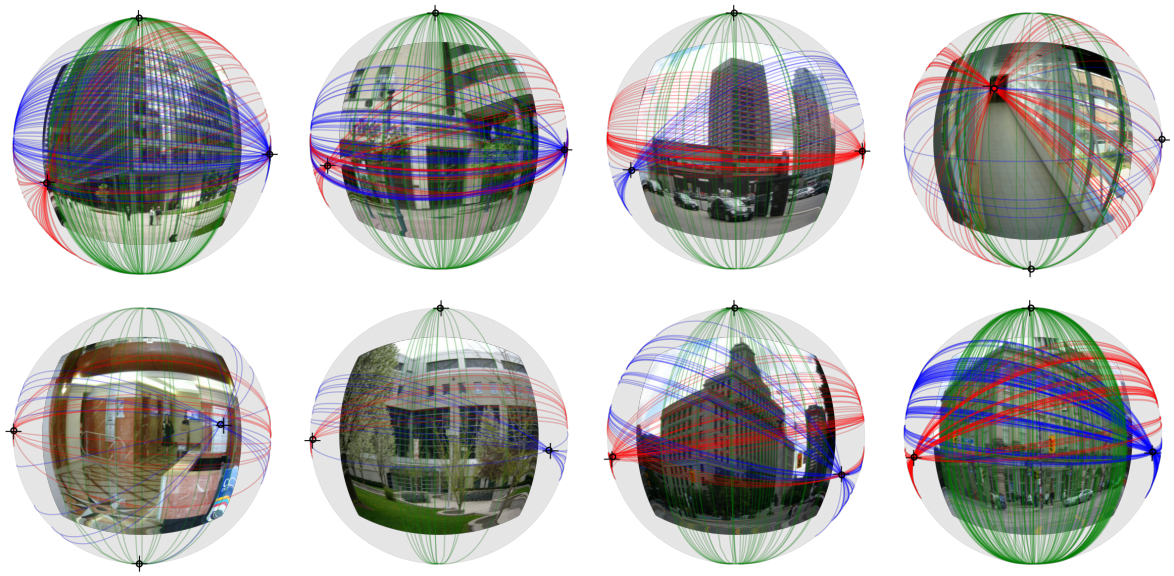

Figure 3. **Visualizations on YUD dataset.** We show ground-truth vanishing points ($\circ$) and our predictions ($\times$) on the Gaussian hemisphere, as well as ground truth lines. Our model accurately predicts vanishing points in man-made environments.

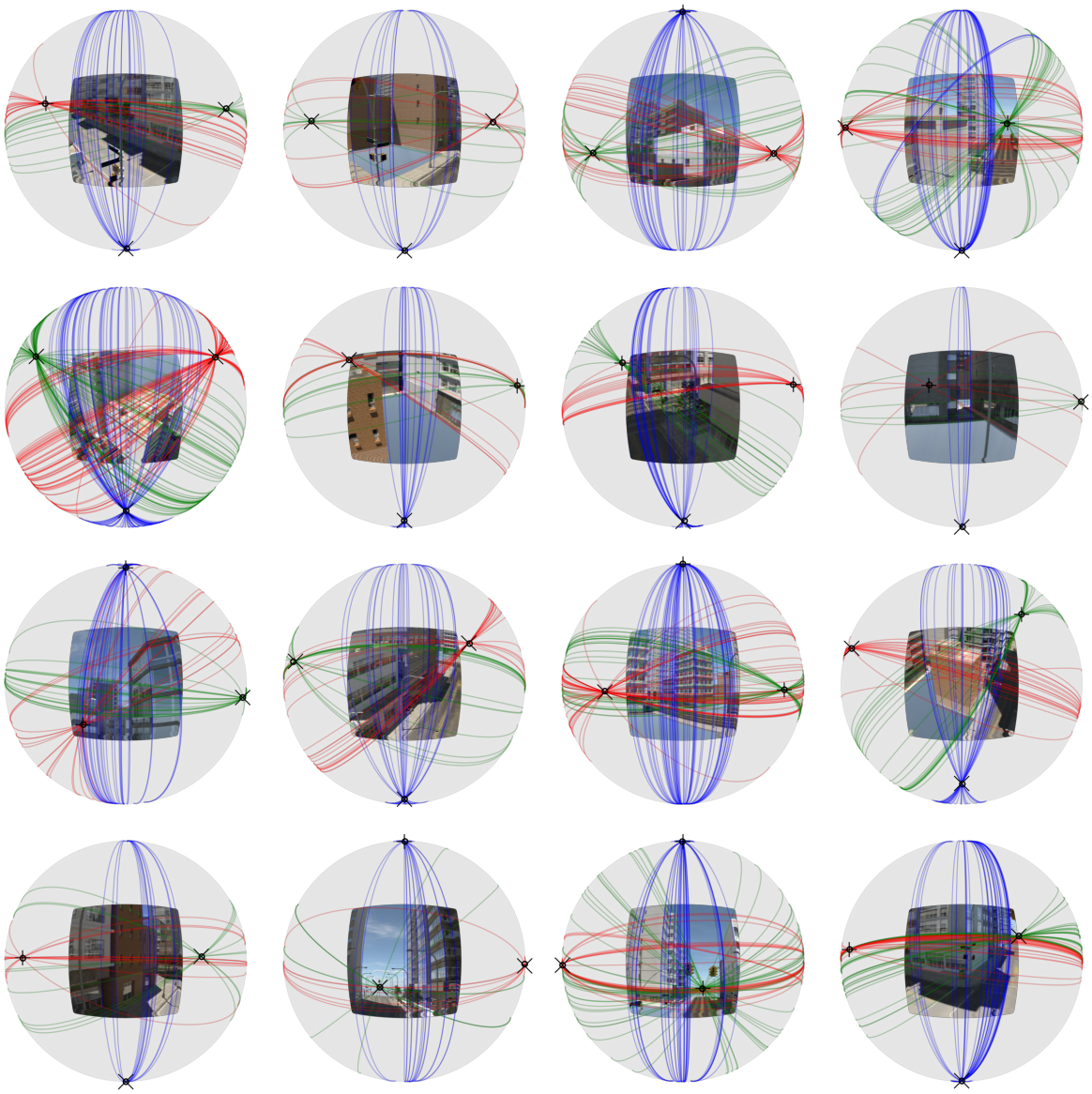

Figure 4. **Visualizations on SU3 dataset.** We show ground-truth vanishing points ($\circ$) and our predictions ($\times$) on the Gaussian hemisphere, as well as ground truth lines. Each color represents a cluster of lines that is related to a vanishing point. Our model accurately predicts vanishing points in man-made environments.

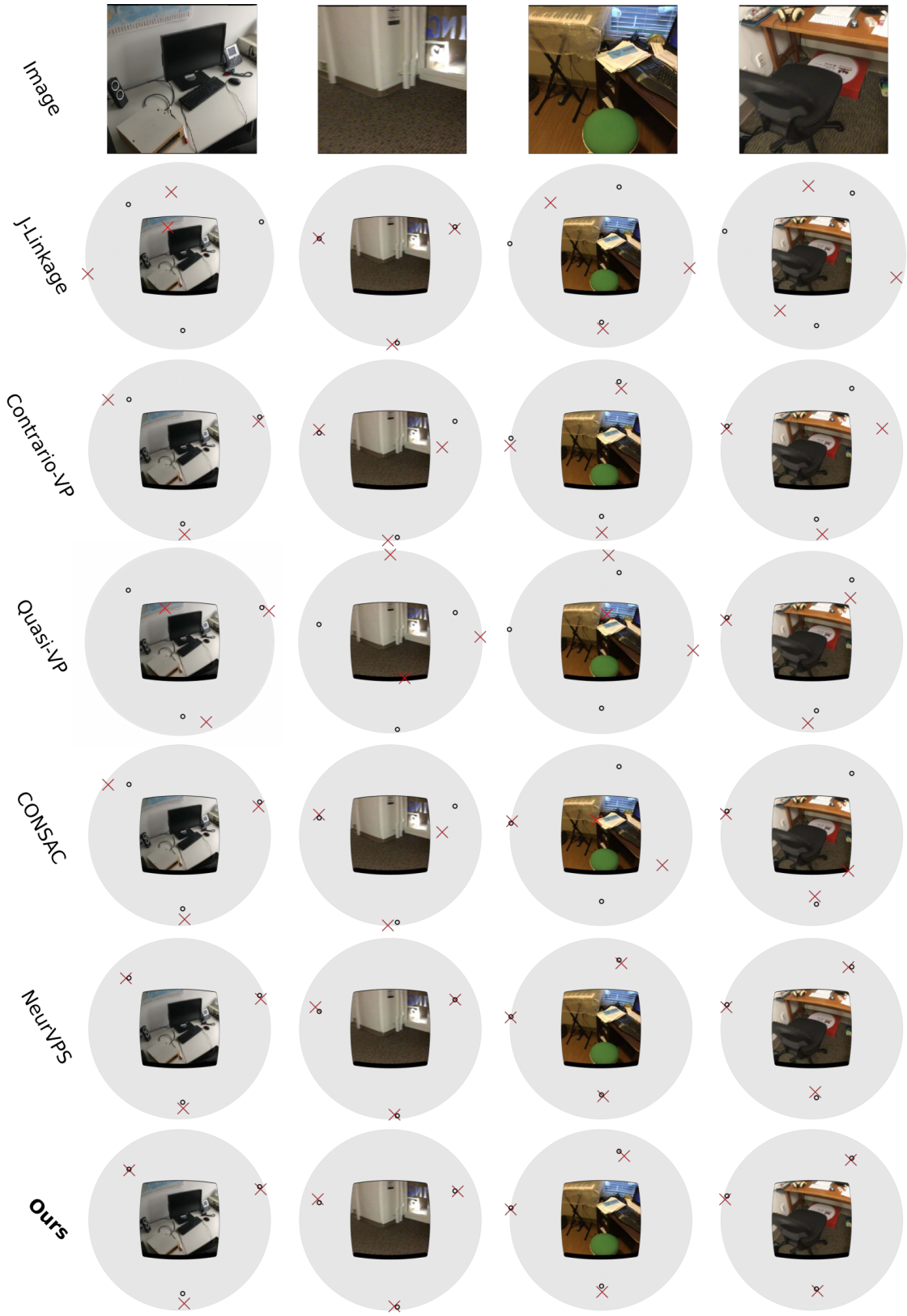

Figure 5. **Visualizations on ScanNet dataset.** We show ground-truth vanishing points (○) and predictions from all baseline methods (×) on the Gaussian hemisphere. Learning-based models shows superior performance to classic line segment-based approaches in complex real-world environments.