COLLAB-VP: Structure-enhanced VP Detector

Abdelkarim Elassam, Gilles Simon, Marie-Odile Berger

COLLABVP: STRUCTURE-ENHANCED VP DETECTOR

Abdelkarim Elassam *Gilles Simon* *Marie-Odile Berger*

INRIA, Université de Lorraine, LORIA, FR-54600

ABSTRACT

We introduce in this paper a novel structure-enhanced VP detector, CollabVP, which uses a multi-task learning framework to estimate multiple horizontal vanishing points (VPs) from a single RGB image. CollabVP exploits contextual information and scene structures through masks of vertical structures to accurately estimate the horizon line and VP positions. The proposed approach is not limited to Manhattan worlds and can detect any number of VPs.

***Index Terms*—** Vanishing points detection, Horizon line estimation, Scene structure segmentation, Multi-task learning

1. INTRODUCTION AND RELATED WORKS

With a pinhole camera, a vanishing point (VP) is a point in the image where parallel lines in 3D space appear to converge. Traditional VP detection methods start by extracting line segments from the image [1], then try to group them by using an algorithm such as RANSAC [2], J-linkage [3], the Hough transform [4] or more recently CONSAC, a learning-based robust estimator [5]. Recent papers [6, 7] dealing with man-made environments have shown that detecting first the horizon line, with a CNN approach in Zhai's paper [6], allows constraining the search of horizontal VPs on this line and greatly improve the accuracy of the VPs detected with line segments. In order that any set of lines that meet accidentally on the HL can generate a false positive, a robust method based on the *a-contrario* methodology [8] was proposed by Simon in [7], which leads to far fewer false positive and duplicate VPs than the optimization method proposed in [6].

Following the trend of learning-based methods, several CNN-based approaches for VP detection have been proposed, often with constraints on the number of detected VPs or on the environment. In [9, 10], only a dominant VP is detected inside a grid map of the image. Other works [3, 11, 12, 13] operate under the assumption of a Manhattan World, assuming three mutually orthogonal VPs, among which is the zenith. For example, [13] introduced TLC, a transformer-based line segment classifier designed to classify a line segment into one of the three Manhattan directions. A less strict hypothesis is the assumption of an Atlanta world where the zenith and an unknown number of horizontal VPs are assumed. Several works [6, 14, 15], including ours, rely on this assumption.

Other methods operate on the bounded Gaussian sphere instead of the unbounded image plane as in Kluger et al. [16] where VP detection is formulated as a multi-label classification task on the sphere. Zhou et al. [14] presented NeurVPS, a CNN with geometry-inspired convolutional operators for detecting VPs. Using an image and a candidate point on the unit sphere as input, their network predicts the probability of the point being near a ground-truth (GT) VP. To exploit the geometric properties of VPs as the intersection of parallel lines, the authors developed an operator named conic convolution, which explicitly enforces feature extractions and aggregations along the structural lines. However, this approach relies heavily on the initial random sampling on the sphere. Recently, Liu et al. introduced VaPiD [15], a more efficient version of NeurVPS that uses learned optimizers and a computation-sharing scheme to process VP anchors efficiently. VaPiD [15] performs better than all previous state-of-the-art VP detection approaches. However, this network and NeurVPS have been only evaluated against precision but not against recall on the Holicity dataset [17], which prevents evaluating the number of VPs found among those expected.

This paper proposes Collab-VP, a horizon-first VP detection method that exploits contextual information and structural VPs through a multi-task CNN. As in [14, 15], we consider a CNN supervised by couples $< image, VP >$. We additionally exploit the fact that man-made environments contain structural VPs which correspond to the intersection of vertical planes –whose orientations can be roughly inferred with learning-based techniques– with the horizontal plane. We thus propose a flexible two-branch framework (see Fig. 1) that outputs the horizon line, the distribution of VPs along the HL and the masks of discretized orientations of vertical structures. We finally propose a robust method for VP extraction inspired from [7] that jointly processes these outputs.

2. METHOD

2.1. Network architecture

The two-branch CollabVP architecture is shown in Fig. 1. The DirectVP branch uses the ResNet50 architecture [18]

This work is part of the MoveOn project between INRIA and the German Research Center for Artificial Intelligence (DFKI).

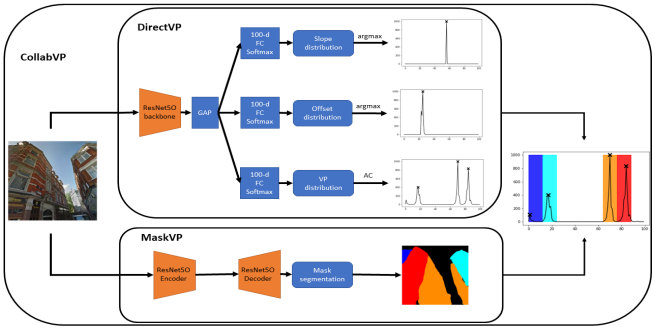


Fig. 1. Overview of CollabVP (better seen when zooming).

as a backbone, with the convolutional layers left unchanged. Three separate softmax classifiers replace the original classifier with 100 outputs each. The first two branches of DirectVP estimate the distribution of the slope and offset of the HL, respectively, while the third branch predicts the distribution of the VPs along the HL. The HL is parameterized by its slope angle $\theta \in$ [-π/2, π/2] and offset r $\in$ [-∞, +∞], where θ is the angle between the HL and the x-axis of the image and r is the y-intercept of the HL with the y-axis of the image. The x-coordinates of the VPs are sufficient to locate them on the HL. The infinite space is squashed to the interval [-π/2, π/2], using the inverse of the tangent function, to deal with infinite VPs and span the entire parameters spaces of the offset and the x-coordinates of VPs. HL and VP detection are considered a classification problem: the slope, squashed offset, and squashed x-coordinates of the VPs are converted into independent categorical classes by dividing their respective domains uniformly into $N = 100$ bins. A classification approach enables us to estimate a probability distribution over possible HL and VPs. Though training is done with only one VP for each image, several peaks can be obtained at the test stage in the output scores, corresponding to multiple VPs.

Though the DirectVP branch may be used alone, it may have difficulty detecting VPs associated with planar structures, such as facades, especially when they occupy a small portion of the image (see section 3.4). Therefore, we introduced a second branch, MaskVP, which is more specifically dedicated to detecting VPs from vertical planar structures. This problem is closely related to extracting the normal map from an image. Still, the clustering of normals turns out to be a particularly unstable task. For increased robustness, we have therefore chosen to learn a U-net classifier that only detects vertical structures and classifies their orientations into $N = 9$ classes, with $N - 1$ orientation bins spread between -π/2 and π/2, and one class for the rest. The masks with less than 500 px are discarded. Finally, the two-branch CNN has four outputs: the mask segmentation, the HL parameter distribution and the VP distribution.

2.2. Loss functions

To train the DirectVP branch, we use the cross-entropy loss of the distributions of the HL and VP parameters: $L_{parameter} = -\frac{1}{N}\sum_{i=1}^{N} y_i.log(\hat{y_i})$, where y_i and $\hat{y_i}$ are discretized in $N = 100$ bins for each parameter. Mask segmentation is also treated as a classification problem. For each pixel i, the cross-entropy loss is calculated between the predicted and GT masks and the final loss can be written as $L_{Mask} = -\frac{1}{M}\sum_{i=1}^{M} y_i.log(\hat{y_i})$, where M is the number of pixels in the input image. The overall losses are defined as:

$$L_{DirectVP} = L_{Slope} + L_{Offset} + L_{VP} \quad (1)$$

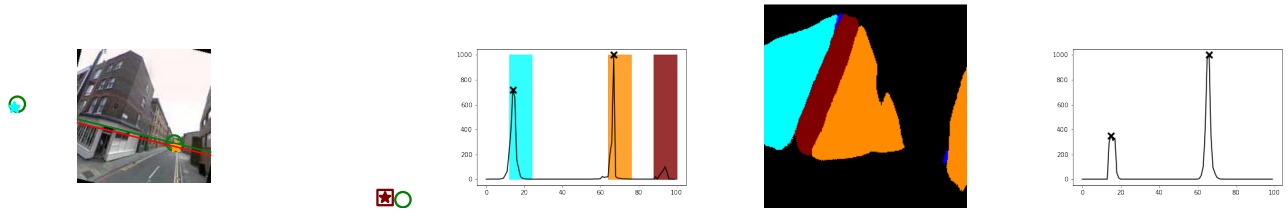
$$L_{CollabVP} = L_{DirectVP} + L_{Mask} \quad (2)$$

Most multitask CNNs are designed with branches sharing a common encoder. However, our experiments with a common encoder have proved disappointing. The network appeared to rely heavily on vertical structures while ignoring line cues useful for VP detection. For this reason, we opted for an alternative strategy described below that explicitly leverages both tasks during inference and improves the performance of each branch when compared to training them separately (see fig.2).

2.3. VP detection and collaboration strategy

The HL is computed from the slope and offset corresponding to the best scores of the related output distributions. VPs are detected as local maxima of the third distribution provided by DirectVP (Fig. 1). The probabilistic *a-contrario* (AC) framework [8] is used for this purpose, which has proven relevant to reliably detect modes of a histogram based on a parameter-free multi-scale approach. It is robust to irrelevant peaks and can detect peaks made of several consecutive bins. However, it can happen that VPs obtain low scores, judged as not meaningful by the AC method, mainly when they are associated with small or very low-textured structures. These scores are, in fact, judged not meaningful because other VPs obtain much higher ones. But isolated from the others, i.e. considered within an orientation interval provided by MaskVP, such VPs can be recovered.

Therefore, our collaboration strategy adds VPs in intervals found by MaskVP only if no VP was detected by the AC method in these intervals. In that case, the value with the highest score inside the mask is retained as VP (Fig. 1). This procedure is illustrated in Fig. 2. The input image is shown in Fig. 2a, with the three VPs finally detected in bold stars. The output scores of DirectVP are shown in Fig. 2b, with crosses showing the peaks detected by the AC method: two VPs were found. By contrast, MaskVP predicts three VP intervals (2c): two of them (blue and orange) contain a previously detected VP, but the third one (maroon) allows the detection of an additional VP whose score was too low to be detected by the AC method. This interval corresponds to the



(a) Input image with detected VPs. (b) Output of CollabVP. (c) Output of MaskVP. (d) Output of DirectVP (TA).

Fig. 2. Collaboration between DirectVP and MaskVP. GT HL and VPs (with circles) are in green. The detected VPs are marked with a star whose colour corresponds to that of the mask. A square indicates that the VP was only detected by MaskVP.

small facade in the middle of the left building (2c). In comparison, Fig. 2d shows the VP distribution obtained with DirectVP trained alone (TA). It can be seen that the third VP is not detected, whereas a small peak is obtained when DirectVP is trained together with MaskVP (fig.2.b).

3. EXPERIMENTAL RESULTS

3.1. Datasets and evaluation criteria

To conduct our experiments, we used the recently published Holicity dataset [17] both for training and for testing, and we also evaluated our model on the YU dataset [19], which provides GT VPs but lacks normal maps. The Holicity Dataset consists of 54354 real-world images covering downtown London. It provides labels for different 3D structures in urban environments, such as surface normal maps and VPs. We adopted the split the authors proposed, containing 45032 training samples and 2504 validation samples. We applied several data augmentation techniques to increase the training dataset's variety and quality. Firstly, we rotated the original images to different angles that follow a normal distribution with $\sigma = 10^{\circ}$ to simulate variations in the camera's roll. We also utilized homography transformations to introduce changes in the camera's pitch. Using these techniques, we generated more diverse training samples, resulting in better performance and improved generalization to new, unseen images. The York Urban dataset (YU) [19] consists of 102 indoor and outdoor scene images, each with provided camera parameters and GT VPs that satisfy the Manhattan world assumption.

Our method was evaluated using the mean and median angle errors. Also, the angle accuracy (**AA**) defined in [14] is used to quantify the quality of VP detections. We compute the angle difference between each predicted VP and the closest GT VP. The AA^{θ} value is defined as the area under the angle accuracy curve between [0, θ] divided by θ. We calculate AA^{θ} using various precision levels θ as in [14, 15]. A second metric called **AA-R** (in reference to the *recall* term) is introduced to quantify the ratio of GT VPs detected with the different methods. We compute the angle difference between each GT VP and the closest predicted VP. The AA-R$^{\theta}$ indicates the percentage of GT VPs detected at a certain precision level θ and shows the method's ability to detect multiple VPs.

3.2. Results on Holicity

Our method was compared on Holicity to state-of-the-art line-based methods such as [7] and to recent learning-based methods such as NeurVps [14] and VaPid [15], which are the closest to our work since working on non-Manhattan scenes. Since the source code of VaPid [15] is unavailable, we conducted a comparison on Holicity by referring to the tables provided in [15] and utilizing the same angular parameters (1, 2, and 10 degrees). The results presented in Tab. 1 show that CollabVP outperforms other baselines regarding the precision of detected VPs, as indicated by the AA and AA-R values. CollabVP achieves an AA^1 of 30.9% (AA^2 of 42.8% and AA^{10} of 79.5%), surpassing the previous best performance of VaPiD by a relative improvement $(AA_{new} - AA_{old})/(1 - AA_{old})$ of 11.3% (14.5%, 16.6%). Although the precision of NeurVPs [14] and VaPiD [15] methods on Holicity is reported, they do not provide the number of detected GT VPs (their method detects a fixed number of VPs). In contrast, we report the AA-R metrics: in the test set, there are 4973 GT VPs, and CollabVP detects AA-R$^{10} = 86.1\%$ of them with a precision of less than 10° (AA-R$^1 = 48.3\%$, AA-R$^2 = 65.1\%$). In comparison, Simon et al.'s method [7] only detects 80.5% of them with such precision (AA-R$^1 = 41.3\%$, AA-R$^2 = 56.6\%$). This demonstrates the advantage of extracting structures with similar normal vectors, as it enables the detection of a more comprehensive set of structural VPs. It is worth noting that although our method's mean error is slightly higher than VaPiD's, it detects additional points that are correct but not present in the ground truth. These additional points can significantly impact the mean error metric, as the angle difference between the predicted VP and the closest GT VP will be large, inflating this average.

3.3. Results on York Urban

We also compared CollabVP to several state-of-the-art methods, namely J-Linkage [3], Simon et al. [7], Li et al. [20], Wu et al. [21], Lu et al. [22], CONSAC [5], and NeurVPS

Method	AUC - HL	AA^1	AA^2	AA^{10}	Mean	Median
Simon et al. [7]	72.9	22.0	35.0	62.1	19.17°	1.61°
NeurVPS [14]	-	18.2	31.7	62.1	8.32°	1.78
VaPiD [15]	-	22.1	39.6	75.4	**3.00°**	1.19°
CollabVP	**93.1**	**30.9**	**48.4**	**79.5**	4.23°	**0.84°**

Table 1. Comparisons with baseline methods on HoliCity.

Method	AA^3	AA^5	AA^{10}
J-linkage [3]	40.2	50.5	64.1
Simon et al. [7]	40.1	58.2	77.5
Wu el al. [21]	44.3	61.4	77.4
Li et al. [20]	51.1	66.1	80.5
Lu et al. [22]	58.0	73.2	86.2
CONSAC [5]	62.1	73.7	84.1
NeurVPS [14]	39.9	50.3	65.0
Tong et al. [13]	65.5	77.1	**87.4**
CollabVP - trained on Holicity [17]	57.6	73.2	83.3
CollabVP - trained on NYU [23] + Holicity [17]	**66.8**	**77.8**	86.6

Table 2. Comparisons with baseline methods on YU [19].

[14], using the YU dataset [19]. The results are shown in Tab. 2 with the same accuracy parameters (3, 5, and 10 degrees) as in [13]. When trained solely on outdoor scenes, our method performs less than the current best performance [13], but still better than previous state-of-the-art methods. However, training our method on a combination of indoor [23] and outdoor scenes[1] resulted in better generalization capabilities (Fig. 3) and a relative improvement of 3.7% for AA^3 (3.1% for AA^5) compared to the previous best performance [13]. Additionally, CollabVP detects 98.2% of the GT VPs with a precision of less than 10° (81.5% with a precision of less than 2°). Notably, our method does not impose any constraints on the number VPs or their orthogonality, yet it outperforms methods that explicitly consider such constraints. It is important to note that the Holicity dataset has a lower recall metric (86.1% for 10°) compared to the YU (98.2%) due to the fact that the GT VPs in Holicity are generated from 3D CAO models, which results in some VPs that cannot be perceived in the images. In contrast, all images in YU were hand-labelled, resulting in a higher recall metric.

3.4. Origin of the detected vanishing points

Our experiments on the Holicity dataset reveal that out of the total detected VPs, 2573 were detected by both AC and masks, achieving a precision of 82.0% for 5 degrees. When relying on AC alone, 1039 VPs were detected, with an accuracy of 79.2% for 5 degrees. Similarly, when using masks alone, 1089 VPs were detected with an accuracy of 72.2% for 5 degrees. These findings emphasize that points detected using only masks are meaningful, even if their score in the VP distribution is low and were not directly extracted using the AC framework. Points exclusively detected by AC typically correspond to VPs from buildings located far away or behind occlusions, such as trees, while points solely detected by masks generally correspond to small or elongated surfaces without texture.

[1]The NYU dataset [23] contains 1449 images of indoor scenes with GT surface normal maps and hand-labelled VPs. Following [5], we partitioned the dataset into 1000 training samples, 224 validation samples, and 225 testing samples. To increase the size of the dataset and to enhance the diversity of the training dataset, we used data augmentation techniques similar to those used with Holicity. We froze the pre-trained model's weights (from Holicity) and only retrained the output layers on NYU using a very low learning rate.

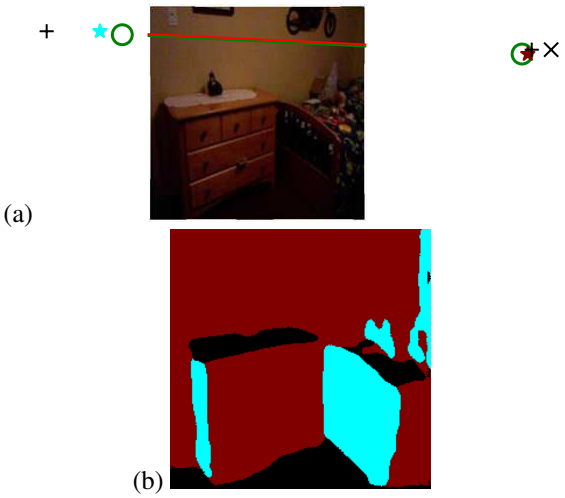


Fig. 3. Comparison of CollabVP (bold stars) with CONSAC [5] (+) and Simon [7] (x). Green circles indicate the GT VPs.

3.5. A sample result

Finally, Fig. 3a shows sample VPs obtained in an indoor scene, compared with CONSAC [5] (black crosses +) and Simon et al.'s [7] (black cross x) results. CONSAC [5] and CollabVP outperform the line-based method of Simon et al. [7] regarding prediction accuracy. Simon et al.'s performance is affected by brightness and shadows, which are not handled well by their method despite the scenes containing long lines. Notably, two GT VPs were identified, but Simon et al. [7] only detected one, while CONSAC [5] detected both. In contrast, CollabVP detected all VPs with better precision and accurately segmented the associated vertical structures (Fig. 3b).

4. CONCLUSION

CollabVP accurately estimates the HL and VP positions by leveraging contextual information and scene structures, outperforming traditional line-based and learning-based methods. Our approach does not impose constraints on the number or orthogonality of VPs, yet outperforms methods that do. This provides a more robust and flexible solution for VP detection in various environments. Although the generated masks are by-products, they could benefit robotics tasks such as semantic SLAM. We plan to explore this direction in our future work.

5. REFERENCES

[1] R. Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, pp. 722–32, 04 2010.

[2] R. Bolles and M. Fischler, "A ransac-based approach to model fitting and its application to finding cylinders in range data," in *IJCAI*, 1981.

[3] J.-C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I.-S. Kweon, and M. Pollefeys, "Globally optimal line clustering and vanishing point estimation in manhattan world," in *IEEE Conference on Computer Vision and Pattern Recognition*, 06 2012, pp. 638–645.

[4] P. Hough, "Machine analysis of bubble chamber pictures," 1959.

[5] F. Kluger, E. Brachmann, H. Ackermann, C. Rother, M. Y. Yang, and B. Rosenhahn, "Consac: Robust multi-model fitting by conditional sample consensus," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] M. Zhai, S. Workman, and N. Jacobs, "Detecting vanishing points using global image context in a non-manhattan world," *CoRR*, vol. abs/1608.05684, 2016.

[7] G. Simon, A. Fond, and M.-O. Berger, "A-Contrario Horizon-First Vanishing Point Detection Using Second-Order Grouping Laws," in *ECCV 2018 - European Conference on Computer Vision*, Munich, Germany, Sept. 2018, pp. 323–338. [Online]. Available: https://hal.inria.fr/hal-01865251

[8] A. Desolneux, L. Moisan, and J.-M. Morel, *From Gestalt Theory to Image Analysis: A Probabilistic Approach*, 1st ed. Springer Publishing Company, Incorporated, 2007.

[9] A. Borji, "Vanishing point detection with convolutional neural networks," *CoRR*, vol. abs/1609.00967, 2016. [Online]. Available: http://arxiv.org/abs/1609.00967

[10] Y. Liu, M. Zeng, and Q. Meng, "D-vpnet: A network for real-time dominant vanishing point detection in natural scenes," *CoRR*, vol. abs/2006.05407, 2020. [Online]. Available: https://arxiv.org/abs/2006.05407

[11] J. Košecká and W. Zhang, "Video compass," in *Computer Vision — ECCV 2002*, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 476–490.

[12] H. Wildenauer and A. Hanbury, "Robust camera self-calibration from monocular images of manhattan worlds," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2831–2838.

[13] X. Tong, X. Ying, Y. Shi, R. Wang, and J. Yang, "Transformer based line segment classifier with image context for real-time vanishing point detection in manhattan world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6093–6102.

[14] Y. Zhou, H. Qi, J. Huang, and Y. Ma, "NeurVPS: Neural Vanishing Point Scanning via Conic Convolution," *CoRR*, vol. abs/1910.06316, 2019. [Online]. Available: http://arxiv.org/abs/1910.06316

[15] S. Liu, Y. Zhou, and Y. Zhao, "Vapid: A rapid vanishing point detector via learned optimizers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 839–12 848.

[16] F. Kluger, H. Ackermann, M. Y. Yang, and B. Rosenhahn, "Deep learning for vanishing point detection using an inverse gnomonic projection," *CoRR*, vol. abs/1707.02427, 2017. [Online]. Available: http://arxiv.org/abs/1707.02427

[17] Y. Zhou, J. Huang, X. Dai, L. Luo, Z. Chen, and Y. Ma, "HoliCity: A city-scale data platform for learning holistic 3D structures," *CoRR*, 2020, arXiv:2008.03286 [cs.CV].

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[19] P. Denis, J. Elder, and F. Estrada, "Efficient edge-based methods for estimating manhattan frames in urban imagery," 10 2008, pp. 197–210.

[20] X. Lu, J. Yao, H. Li, and Y. Liu, "2-line exhaustive searching for real-time vanishing point estimation in manhattan world," 03 2017, pp. 345–353.

[21] J. Wu, L. Zhang, Y. Liu, and K. Chen, "Real-time vanishing point detector integrating under-parameterized ransac and hough transform," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 3732–3741.

[22] H. Li, J. Zhao, J.-C. Bazin, W. Chen, Z. Liu, and Y.-H. Liu, "Quasi-globally optimal and efficient vanishing point estimation in manhattan world," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[23] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.