DQ-HorizonNet: Enhancing Door Detection Accuracy in Panoramic Images via Dynamic Quantization

Cing-Jia Lin[1] Jheng-Wei Su [1] Kai-Wen Hsiao[1] Ting-Yu Yen[1] Chih-Yuan Yao[2] Hung-Kuo Chu[1]

[1]National Tsing Hua University, Taiwan
[2]National Taiwan University of Science and Technology, Taiwan

Abstract

This paper introduces DQ-HorizonNet, a novel learning-based methodology that incorporates vertical features to enhance doors detection in indoor panoramic images. Building upon HorizonNet, which excels in estimating 3D indoor layouts from panoramic images using 1D vectors to identify boundaries, we identify a key limitation: HorizonNet's dense, column-wise prediction output is ill-suited for object detection tasks due to the need for complex post-processing to separate true positives from numerous false-positive predictions. DQ-HorizonNet innovatively addresses this issue through dynamic quantization, which clusters column-wise outputs and assigns learning targets dynamically, improving accuracy via a U-axis distance cost matrix that evaluates the discrepancy between predictions and actual data. Our model, tested on the extensive Zillow indoor dataset (ZInD), significantly outperforms existing methods, including the original HorizonNet and the transformer-based DETR network, showcasing its superior ability to accurately detect doors in panoramic indoor imagery.

The code can be found on https://github.com/Lontoone/DQ-HorizonNet/.

1. Introduction

Door detection holds significance across diverse domains, including robot navigation, human visual assistance, and layout estimation, among others. However, detecting doors within a single panoramic image poses numerous challenges. These challenges encompass the diverse states of doors (closed, open, or semi-open) and the inherent distortion within panoramic images, rendering traditional bounding box annotations inadequate.

Door detection has been a subject of research for numerous years. Traditional approaches have relied on edge detectors such as the Sobel filter or Canny edge detection to extract features. However, these methods have been constrained by their representational capacity. Deep learning-based methodologies have tackled this limitation by harnessing deep neural networks. These networks provide a more robust and comprehensive representation of features, markedly improving the effectiveness of door detection tasks.

Classic object detection networks include Faster-RCNN [7], YOLO [14], and DETR [1]. These models excel in general-purpose object detection. However, they are restricted by the bounding box representation, which only allows for rectangular shapes and cannot align distortions in panoramic images. Moreover, the anchor-point strategy can result in unequal sampling issues, given the higher pixel density at the poles and sparser distribution at the equator in equirectangular images. Specifically, training the DETR model demands considerable effort. Lastly, these models do not fully take advantage of the vertical features of doors, which constitute a crucial aspect of door detection.

As shown in Figure 1, the characteristics of a perpendicular frame provide rich information about a door. Therefore, capturing the vertical features of a door is a crucial aspect of door recognition [10, 11, 20]. HorizonNet, a deep learning model with a height compression module that flattens features into 1D vectors, has demonstrated significant performance in estimating the layout of indoor panoramic images. However, directly applying HorizonNet to the door detection task will produce a dense column-wise prediction followed by a post-processing step such as non-maximum suppression to reduce false positives (see Figure 2).

Adapting a model originally designed for dense predictions to a task that requires sparse predictions presents a notable challenge, particularly regarding the assignment of learning targets. For instance, when one side of a door frame spans 5 columns in width, deciding which column's output should be used and back-propagated becomes a non-trivial sampling problem.

The techniques employed to sample and assign learning targets for each prediction can significantly influence the

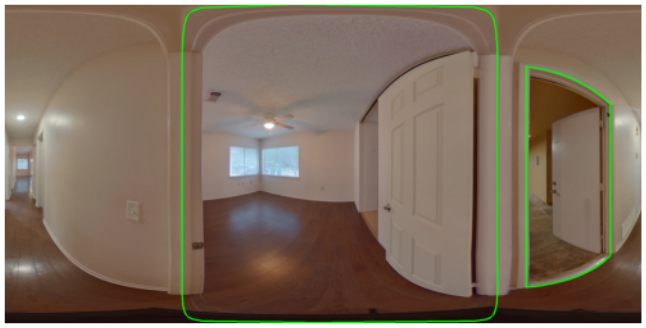

Figure 1. **Doors typically feature prominent vertical lines.** The vertical side of a door frame exhibits strong characteristics, unlike the horizontal side, which is significantly affected by distortion.

Figure 2. **The issue of false positives on small doors.** The lines depict the output of directly applying HorizonNet to door detection. However, several overlapping boxes were not successfully trimmed by non-maximum suppression, indicating a need for further tuning in the non-maximum suppression trimming threshold.

training process and the model's performance. Therefore, it is imperative to implement a customized grouping method to tackle this challenge effectively. In this work, we introduce DQ-HorizonNet, a novel deep neural network that incorporates vertical features and a dynamic quantization strategy to enhance the accuracy of door detection. This approach involves grouping the column output through a maximum filter and assigning the learning target according to a cost matrix. This matrix computes the U-axis distance between the predictions and the ground truths.

We carried out a thorough analysis to evaluate the performance of DQ-HorizonNet using the extensive Zillow Indoor Dataset (ZInD). The findings reveal that our approach surpasses other competing models, such as the original HorizonNet and the transformer-based DETR network, by a significant margin in terms of door detection capabilities.

In summary, we make the following contributions.

- We highlight the unique challenges in door detection within panoramic images, including diverse door states and image distortions.
- We introduce a new deep learning model DQ-HorizonNet which incorporates vertical features and a dynamic quantization strategy to accurately detect doors in panoramic images.
- We propose a customized grouping method to assign learning targets in scenarios that require sparse predictions, addressing a notable challenge in adapting models designed for dense predictions.
- DQ-HorizonNet presents superior performance compared to competing models in door detection tasks.

2. Related Work

2.1. Optimization-based approach.

The door frame offers valuable structural cues, as the presence of two adjacent vertical edges signifies a higher likelihood of a door [2, 6, 10, 11, 15]. Some further employ corner detection along with edge detection to maximize information derived from the relationships between edges and corners of a door [12, 20].

From traditional approaches, the vertical characteristics of doors have been found to provide essential information for door detection. However, reducing false positives remains a challenge. Therefore, a classifier is employed to address this issue. In this work, we employ a column-wise classifier to identify the location of a door on the u-axis of the image.

2.2. Learning-based approach.

In recent years, deep learning models have been extensively utilized in object detection tasks. Anchor-based approaches such as Faster-RCNN [7] and YOLO [14], as well as transformer-based models such as DETR [1], have demonstrated impressive performance in object detection of general purpose. However, bounding box annotations may not be sufficiently precise for panoramic images, as panoramic distortion can significantly distort the objects.

Therefore, SphereNet [4] proposes to project the CNN kernel onto a sphere. This technique is designed to manage distortions and maintain a uniform sampling area. However, Chou et al. [3] compared the performance of a standard convolution kernel with a spherical convolution, and it was found that the standard convolution kernel outperforms the spherical convolution.

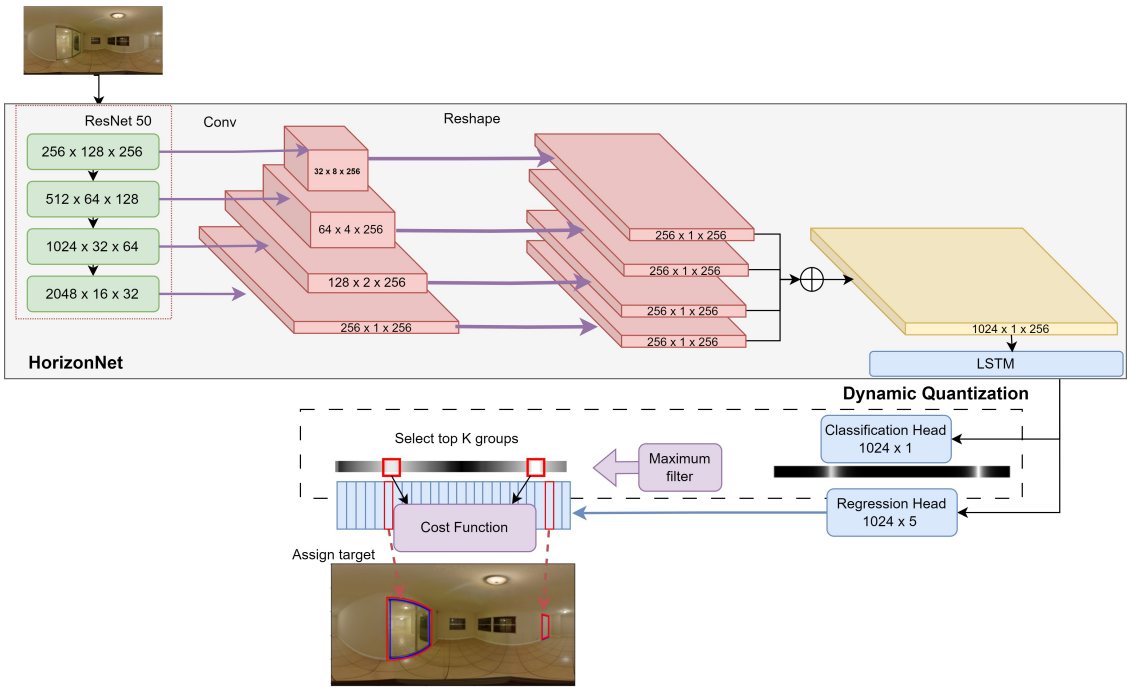

Figure 3. **Network architecture.** The upper part adheres to the HorizonNet architecture. We incorporate a single linear layer as the classification head and another linear layer for corner point regression. During training, the output logits from the classification are subjected to a maximum filter. Subsequently, k-groups with the highest logits are selected. A cost function, which calculates the U-axis distance, is employed to determine the appropriate column of corner point regression to match with the ground truth.

Feature compression Effectively addressing panoramic distortion presents a multifaceted challenge. Unlike the multiview approach [16] or the multi-projection strategy [19], HorizonNet is an approach to estimating the 3D room layout from a single panoramic image. It flattens the extracted features into 1D vectors and feeds them into a Recurrent Neural Network (RNN) to maintain output consistency across the image. HorizonNet represents room layout as three 1D vectors that encode, at each image column, the boundary positions of floor-wall and ceiling-wall, and the existence of wall-wall boundary.

HoHoNet [18] further refines this approach by using an efficient height compression module to extract flattened features and applying multi-head self-attention for refinement. SliceNet [13] proposes a decoder mechanism to produce a depth map with the same resolution as the input image. Recent studies, such as Zheng et al. [21], have utilized a bi-directional compression module that extracts compressed horizontal and vertical representations from the input image, thereby improving performance.

3. Methodology

3.1. Network architecture.

Figure 3 illustrates the network architecture of DQ-HorizonNet. Our model leverages HorizonNet and ResNet50 as the backbone, incorporating LSTM for sequence processing. Rather than using three 1D vectors, we employ two linear layers for outputting classification logits and performing corner points regression. Subsequently, a maximum filter is applied to the classification logits to establish a quantified logits distribution. During the training phase, we utilize a cost matrix to match the ground truth with predicted corner points from the column set that has the minimum total u-cost (U-axis distance cost matrix between predictions and ground truths).

Door representation. The column-wise classification enables us to identify the u-axis coordinates for the starting side of a door, which corresponds to the left side of the door-frame. Once the door has been classified, we can annotate it as shown in Figure 4. We denote 4 corner points of a door as follows:

$$(u_l, v_{lt}, v_{lb}, u_r, v_{rt}, v_{rb}), \tag{1}$$

where lt represents the left top, lb represents the left bottom, rt represents the right top, and rb represents the right bottom. We interpret the image in terms of UV coordinates, where the U axis corresponds to the X axis, and the V axis corresponds to the Y axis.

Dynamic quantization. To convert the original column-wise output to a fixed length and sample the output. Instead of quantizing the output to a fixed number, we propose a dynamic quantization method that combines a maximum filter and a group-based assignment. A maximum filter is a type of morphological filter that replaces each pixel value of a matrix with the maximum value of its neighborhood. It can be used for enhancing significant values, smoothing, or removing noise. Inspired by DETR, we can view the probability distribution as a set of groups after applying the maximum filter. Then we can compute the cost matrix composed of u_l distance. To find the optimal assignment of predicted boxes to ground truth boxes, we use the Hungarian algorithm, which is a method for solving the linear assignment problem. During training, the u_l distance is defined as follow.

$$U = L_1(\frac{topK(F(u_l, r), k)}{N}, \hat{u_l}), \tag{2}$$

where F denotes maximum filter. r denotes filter kernel size. $U \in (k, k)$ is the matrix of $L1$ distance in u-axis between selected u_l and ground truth $\hat{u_l}$. k is equal to the amount of ground truths. N denotes number of columns which is 1024 in HorizonNet.We are finding a permutation of k pairs of σ with the least u cost.

$$\sigma = argMin \sum_{i}^{k} U \tag{3}$$

4. Results

4.1. Experimental settings.

Dataset. We conducted all the experiments in the Zillow Indoor Dataset (ZInD) [5], which contains 67,448 indoor panoramas. We exclude images with empty annotations and follow the ZInD train-test split with respect to the annotations for 49,880 images and 6,248 images.

Baselines. We compare our method with HorizonNet [17] and DETR [1]. We use pretrained ResNet50 [8] as the backbone of HorizonNet and retain the LSTM [9] module as part of the model. We modify the HorizonNet output head to match ours, replacing the three 1D vectors with a classification output $\in (1024, 1)$ and a box regression output $\in (1024, 5)$. The Adam optimizer is used with a learning rate of 0.00035, β = (0.9, 0.999), and weight decay = 0. We train HorizonNet for 100 epochs with a batch size of 12 on four NVIDIA GTX 1080 Ti GPUs. Binary Cross-Entropy Loss is used for classification output. L1 loss is used for door regression. The output of HorizonNet is then post-processed with equirectangular projection and non-maximum suppression with a IoU threshold of 25%.

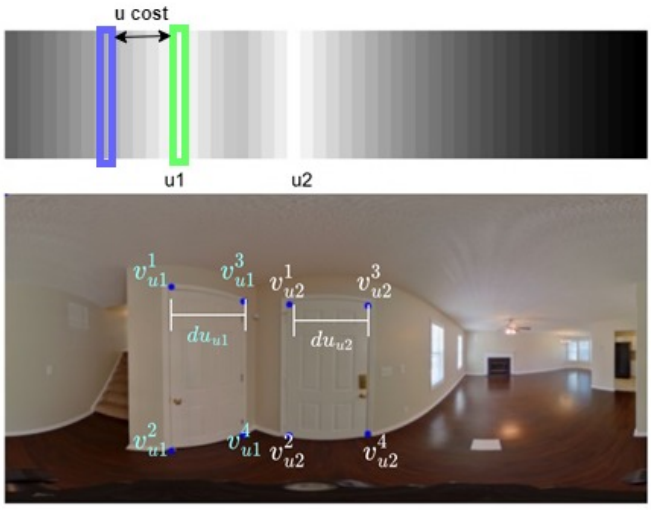

Figure 4. **An example of how classifier and regressor locate a door.** Top: This is an example of sparse columns of classification probability. The brighter the column, the higher the probability of the left side of a door being present. The value is an 1D vector $\in (1, N)$, where N is the column length. Bottom: As discussed in Section 3.1, we annotate each door using six floating-point values. The box regression output is represented as $\in (5, N)$ because we utilize the column index to locate the left side of a door.

We fine-tune the pretrained DETR model provided by HuggingFace for 80 epochs. The output of the DETR model is in the format of a list of bounding boxes, represented as $(c_x, c_y, width, height)$. Subsequently, we employ the same equirectangular projection technique used in HorizonNet to obtain the distorted door annotation and non-maximum suppression with an IoU threshold of 25%. All baseline comparisons are made against the same equirectangular-projected ground truth annotations. We conducted an evaluation of DETR without employing an equirectangular projection. We noticed a minor decrease in the mIoU by approximately 1%. This can be attributed to the higher prevalence of smaller doors in the ZInD dataset, which are less influenced by equirectangular distortion.

Evaluation metrics. We evaluate the performance of our models using average precision (AP), precision, recall, and mIoU. Precision (also known as positive predictive value) is the ratio of correctly predicted positive observations to the total predicted positives. Recall (also known as sensitivity or true positive rate) is the ratio of correctly predicted positive observations to all the observations. They are defined

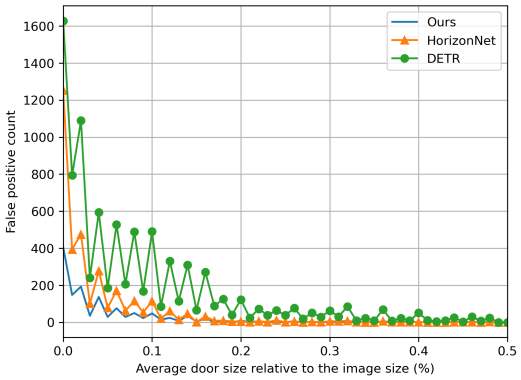

Figure 5. **Our method effectively reduces false positives, particularly in small doors.**The x-axis represents the average ratio of door size to image size at the pixel level for each image. The y-axis represents the number of false positives. Our method yields the fewest false positives when the door size ratio is close to zero.

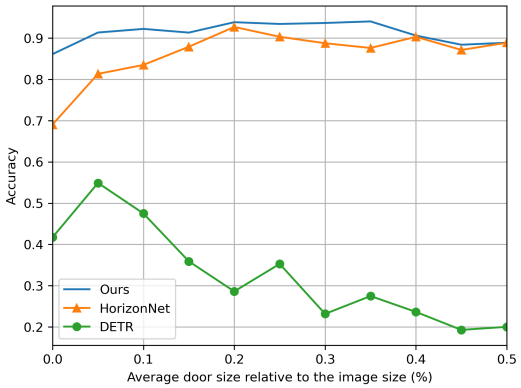

Figure 6. **Our method effectively improve accuracy particularly in small doors.**The x-axis represents the average ratio of door size to image size at the pixel level for each image. The y-axis represents the accuracy score, computed by dividing the number of correct predictions by the total predictions. Our method excels when the door size ratio approaches zero. As the size of the door increases, DETR tends to produce either no predictions or numerous false positives.

as :

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

where (TP) is again the number of true positives,(FP) is the number of false positives and (FN) is the number of false negatives. From the precision and recall, we observe how non-maximum suppression affects our baseline models. Tuning the threshold of non-maximum suppression is a trivial task, but it has a strong impact on the performance, as it determines how many false positives can be filtered out.

Unlike the baseline models, our method does not require non-maximum suppression, as we already group the columns by maximum filter.

Implementation details. We implemented our model in PyTorch with the PyTorch Lightning framework on four NVIDIA GTX 1080 Ti GPUs for 50 epochs. The resolution of the panoramas is resized to 1024×512. The Adam optimizer is used with a learning rate of 0.00035, β = (0.9, 0.999), and weight decay = 0. The batch size is set to 12.

4.2. Equirectangular correction.

To better align the annotation with the actual door shape in the panoramic image, we apply the equirectangular projection to the door annotation on both the ground truth and the prediction. Then we use the projected annotation to compute the IoU.

Equirectangular projection is a simple map projection that maps the longitude and latitude of a spherical surface to the horizontal and vertical coordinates of a flat plane. We first map the UV coordinate to a continuous 3D spherical space using the formula :

$$(x, y, z) = (cos(u)cos(v), cos(v)sin(u), sin(v)) \tag{6}$$

And connect the corner points with linear interpolation. Then we project each interpolated point back to UV coordinate using formula:

$$(u, v) = (atan2(y, x), arcsin(\frac{z}{\sqrt{x^2 + y^2 + z^2}})) \tag{7}$$

4.3. Comparison with baselines.

In Table 1, we present the evaluation results using the ZInD testing dataset for our method and the baselines. We compute mIoU only for true positives. Despite minor variations in mIoU, we noticed a substantial disparity in precision between our method and the original HorizonNet, leading to a superior AP score for our approach. To further elaborate, we illustrate the calculation of false positives in Figure 5, demonstrating that our method significantly improves quality and reduces the number of false positives, especially in the context of small doors. In Table 2, we observe that both HorizonNet and DETR tend to produce multiple overlapping doors, particularly in small door cases, and the non-maximum suppression is not efficient in removing them. Conversely, our method does not rely on non-maximum suppression and results in fewer false positives due to the quantization process and the implementation of a group-based training strategy.

4.4. Ablation study.

We compare our method with the fixed quantified HorizonNet, which modifies the classification head output to

Method	AP@5	AP@50	AP@75	Precision@50	Recall@50	mIoU(%)
Ours	83.0	75.5	50.4	87.7	85.8	74.3
HorizonNet	74.3	66.9	39.4	76.8	87.8	74.8
DETR	61.7	23.1	5.0	42.8	52.7	50.1

Table 1. **Quantitative result.** AP@5 denotes the Average Precision at an Intersection over Union (IoU) threshold of 5%, relative to the True Positive (TP) threshold, and so forth. Our method significantly outperforms others, primarily due to its effective reduction of false positives. Additionally, it is evident that bounding box annotations lack sufficient accuracy for object detection in panoramic images, as reflected by the AP75 metric of the DETR model.

Ours	HorizonNet	DETR

Table 2. **Qualitative results.**The figure depicts the predicted outcomes, denoted by a dotted red line, and the ground truth (GT), represented by a green line. Both HorizonNet and DETR tend to exhibit a higher incidence of false positives. In contrast, our method effectively reduces overlapping predictions, particularly in the context of small doors, thus decreasing the number of false positives.

$\in (1, n)$ and the door regression head to $\in (5, n)$, where n is a hyperparameter that controls the number of columns of quantization as shown in figure Figure 7. At Table 3, we empirically set n to 90 and observed a significant performance gap at AP@75.This gap may attributed to the precision restriction caused by a reduced number of columns.

We also experiment with different maximum filter kernel sizes to evaluate their effect on the result. We have observed that setting kernel size to 100 results in faster convergence and improved performance. It appears that the optimal value of kernel size may be correlated with the maximum density of target objects within the image.

Method	AP@5	AP@50	AP@75	mIoU(%)
Fixed quantization	75.9	65.8	37.0	67.8
Dynamic quantization	83	75.5	50.4	74.3

Table 3. Evaluation of fixed and dynamic strategies of quantization.

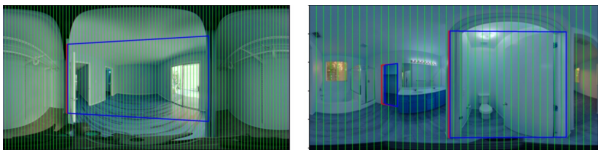

Figure 7. **Visualization of quantizing n columns.** Left: quantize to n = 90 columns; Right: quantize to n = 40 columns. Red line shows the quantized box annotation. Blue line shows the ground truth box annotation. From the right image, we observe that lower n leads to higher u_l error, especially for smaller doors.

Kernel size(r)	AP@5	AP@50	AP@75	mIoU(%)
0	82.6	49.4	18.3	60.8
29	83	75.5	50.4	74.3
100	84.8	76.5	52.1	75.5

Table 4. **Different quantization kernel sizes.** We experimented with different maximum filter kernel sizes of 0, 29, and 100. When kernel size set to 0 meaning not performing maximum filter.

5. Conclusions

In this paper, we conducted a concise review on door detection. To better align the distorted ground truth, we propose an annotation method that employs column-wise classification of the U-axis. To mitigate the issue of false positives, we introduce a dynamic quantization strategy as an alternative to non-maximum suppression. This strategy effectively reduces false positives without the need for threshold tuning. One limitation of our work is the assumption that the pitch and roll angles of the camera are fixed at 0. This method is primarily designed to detect vertically aligned objects. If this is not the case, calibrating the camera might be necessary. In our future research, we may investigate the impact of an uncalibrated camera on detection performance. Concurrently, we aim to broaden our methodology to encompass general-purpose object detection, moving beyond the scope of door detection, by further exploring the potential of feature compression.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2, 4

[2] Zhichao Chen and Stan Birchfield. Visual detection of lintel-occluded doors from a single image. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008. 2

[3] Shih-Han Chou, Cheng Sun, Wen-Yen Chang, Wan-Ting Hsu, Min Sun, and Jianlong Fu. 360-indoor: Towards learning real-world objects in 360deg indoor equirectangular images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 845–853, 2020. 2

[4] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 518–533, 2018. 2

[5] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360º panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2133–2143, June 2021. 4

[6] C Fernández-Caramés, Vidal Moreno, Belén Curto, Jesús Fernando Rodríguez-Aragón, and FJ Serrano. A real-time door detection system for domestic robotic navigation. *Journal of Intelligent & Robotic Systems*, 76(1):119–136, 2014. 2

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1, 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[10] Nosan Kwak, Hitoshi Arisumi, and Kazuhito Yokoi. Visual recognition of a door and its knob for a humanoid robot. In *2011 IEEE International Conference on Robotics and Automation*, pages 2079–2084. IEEE, 2011. 1, 2

[11] Iñaki Monasterio, Elena Lazkano, Iñaki Rañó, and Basilo Sierra. Learning to traverse doors using visual information. *Mathematics and computers in simulation*, 60(3-5):347–356, 2002. 1, 2

[12] Ana Cris Murillo, Jana Košecká, José Jesús Guerrero, and Carlos Sagüés. Visual door detection integrating appearance and shape cues. *Robotics and Autonomous Systems*, 56(6):512–521, 2008. 2

[13] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 11536–11545, 2021. 3

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 2

[15] Marwa M Shalaby, A Mohammed, Megeed Salem, Alaa Khamis, and Farid Melgani. Geometric model for vision-based door detection. In *2014 9th International Conference on Computer Engineering & Systems (ICCES)*, pages 41–46. IEEE, 2014. 2

[16] Jheng-Wei Su, Chi-Han Peng, Peter Wonka, and Hung-Kuo Chu. Gpr-net: Multi-view layout estimation via a geometry-aware panorama registration network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6468–6477, 2023. 3

[17] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 4

[18] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021. 3

[19] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3363–3372, 2019. 3

[20] Xiaodong Yang and Yingli Tian. Robust door detection in unfamiliar environments by combining edge and corner features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 57–64. IEEE, 2010. 1, 2

[21] Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, and Yao Zhao. Complementary bi-directional feature compression for indoor 360deg semantic segmentation with self-distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4501–4510, 2023. 3