SP ATTACK: SINGLE-PERSPECTIVE ATTACK FOR GENERATING ADVERSARIAL OMNIDIRECTIONAL IMAGES

Yunjian Zhang[1,2] *Yanwei Liu*[1,*] *Jinxia Liu*[3] *Pengwei Zhan*[1,2] *Liming Wang*[1] *Zhen Xu*[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3] Zhejiang Wanli University

ABSTRACT

The safety of Deep Neural Networks (DNNs) processing omnidirectional images (ODIs) is an under-researched topic. In this paper, we propose a novel sparse attack, named Single-Perspective (SP) Attack, towards fooling these models by perturbing only one perspective image (PI) rendered from the target ODI. The attack is launched from the perspective domain, and finally the perturbation is transferred to the original ODI. To this end, we propose an effective PI position searching algorithm based on Bayesian Optimization, and then corrupt the PI centered on the desirable position with unconstrained/constrained perturbations. Extensive experiments on synthetic and real-world omnidirectional datasets demonstrate that SP Attack can overcome the projection deformation of ODIs, and mislead the neural networks by limiting the perturbations in a single patch on the target ODI.

Index Terms— Omnidirectional images, adversarial attack, bayesian optimization, sparse attack

1. INTRODUCTION

Present studies on adversarial attacks mainly focus on the tasks of 2D images, and only a few concern the 3D data such as point cloud [1] and 3D mesh [2]. However, another type of images, omnidirectional images (ODIs), which play a vital role in the advanced driver assistance systems (ADAS) [3] and autonomous navigation [4, 5], have been long neglected in the research of adversarial attacks.With an increasing number of learning models [6, 7, 8, 9] towards ODIs emerging for practical applications, it is urgent to evaluate the robustness of these models against adversarial attacks.

For attacking DNN models, sparse adversarial attack is a practically relevant topic of interest. Different from l_p-norm-based algorithms that modify all the pixels of the input image, sparse attacks aim at misleading DNNs by disturbing only a small portion of the image [10, 11]. There are two levels of constraints for sparse attacks. The first is the pixel level [12, 13], limiting the number of perturbed pixels, while the second is the patch level, constraining the disturbed elements in several patches on the image [14, 15]. In order to generate adversarial ODIs, an intuitive approach is to directly apply the 2D sparse attacks on the panoramas. However, it will introduce nonnegligible deformation when we project the raw ODIs to panoramas, reducing the aggressiveness of the attack. Therefore, a new attack suffering from less deformation is expected for generating adversarial ODIs.

ODIs can be represented by a set of perspective images (PIs), which suffer from less deformation due to their narrow views. Inspired by that, we propose to implement the adversarial attack to the DNN models for ODIs by perturbing a single perspective view, and we name our attack as **Single-Perspective Attack (SP Attack)**.

We focus on the score-based black-box scenario, in which the attackers can only access to the predicted scores of the target model. As different PIs corresponding to different view angles, they have divergent impacts on the prediction score of the neural network. Therefore, the core of our work is to select an desirable PI available for launching the attack.

In this paper, we propose a novel PI position searching approach based on Bayesian Optimization (BO), which helps us obtain a proper perspective plane center position in few iterations. After that, we design two attacks to perturb the PI rendered from the selected position. The first does not limit the scale of the added perturbation, called Perturbation-Unconstrained (PU)-SP attack, while the other limits the magnitude of the perturbation in an L_p-ball, named Perturbation-Constrained (PC)-SP attack. Then we remap the perturbed PI onto the spherical surface and generate the corrupted ODI with other regions unchanged. Consequently, the resulting image is the expected adversarial ODI. Overall, the key contributions of this paper are as the follows:

- A novel sparse attack is proposed towards DNNs processing ODIs. In our attack pipeline, the attacker only needs to perturb one PI rendered from the target ODI, which is concise, efficient, and practical.
- We propose a PI position selection method based on Bayesian Optimization, efficiently finding a desirable PI position for launching an attack.

*Yanwei Liu is the corresponding author.

- Extensive experiments are performed on synthetic and real-world ODIs, and the results show the effectiveness of the proposed SP Attack.

Algorithm 1 BO-based PI Position Selection Algorithm

Input: An ODI x^s with ground-truth label y_c, a classifier f_m, the PI position searching space V, and maximum iterations T
Output: A PI position $v_* = (\theta_*, \phi_*)$ for conducting attack

```
Set i = 0, and randomly initilize v, l with n PI positions
for i < T do
    Calculate the GP model l ~ GP(m(v), k(v, v'))
    Estimate μ(v) and δ(v)
    Use GP-UCB to sample the next PI position v_{n+1}
    Calculate l_{n+1}
    if l_{n+1} > 0
        v_* = v_{n+1}
        break
    else
        Add v_{n+1}, l_{n+1} into v, l
end for
if i = T
    No available PI position found
else
    return v_*
```

2. SINGLE-PERSPECTIVE ATTACK

As shown in Fig. 1, given an ODI x^s labeled as y_c and a target model f_m, our goal is to find a PI x^p rendered from x^s that can reserve adversarial perturbations against perspective projection, formulated by

$$\max L(F[x^s, R_e(P(x^p))], y_c), \tag{1}$$

where $L(\cdot)$ is the loss function, $P(\cdot)$ is the perturbed function on x^p, $R_e(\cdot)$ is the function to remap the perturbed PI onto the spherical surface, and $F[\cdot]$ denotes the operation to reconstruct the adversarial ODI. Eq. (1) reveals the pipeline of SP Attack: We firstly select a PI position to lanuch the attack, then perturbations are added to the PI rendered at the selected position, and finally the perturbed PI is remapped onto the spherical surface to generate the adversarial ODI.

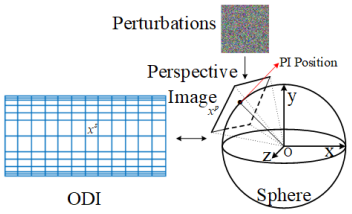


Fig. 1. SP Attack is launched on a PI from the ODI.

2.1. PI Position Selection based on BO

As the perturbation is limited in a single PI, the core of our attack is to select an optimal PI position to launch the attack. A challenge of our attack is that it is conducted in a black-box scenario, which means the gradients of f_m are unavailable, thus a gradient-free optimization algorithm is required for searching desirable PI position. In our work, we propose a PI position selecting approach based on BO, which is an effective and efficient searching algorithm that does not rely on calculating gradents.

Assuming x^p is rendered at spherical coordinates (θ^p, ϕ^p) on ODI, thus $x^p = Pr(x^s, \theta^p, \phi^p)$, where $Pr(\cdot)$ denotes the perspective projection. Then Eq. (1) can be rewritten as

$$\max_{\theta^p, \phi^p \in [0, 2\pi]} L(F[x^s, R_e(P(Pr(x^s, \theta^p, \phi^p)))], y_c). \tag{2}$$

For classification task, $L(\cdot)$ is formulated by

$$L(x_{in}, y_c) = \max(\max\{f_m(x_{in})_i : i \neq y_c\} - f_m(x_{in})_{y_c}, -k), \tag{3}$$

where x_{in} is the input image, and k is the confidence.

We starts from a set of randomly sampled PI positions and their losses, denoted as $\mathbf{v} = [v_1, v_2, ..., v_n]$ and $\mathbf{l} = [l_1, l_2, ..., l_n]$. Then we calculate a model to estimate the distribution of the loss l_* for an unknown PI position v_*, and we use the Gaussian Process (GP) model [16] to fit it

$$\mathbf{l} \sim GP(m(\mathbf{v}), k(\mathbf{v}, \mathbf{v}')), \tag{4}$$

where $m(\mathbf{v})$ is the mean function, and $k(\mathbf{v}, \mathbf{v}')$ is the covariance function.

Next, we utilize an acquisition function to sample another PI position v_{n+1} whose attack loss is more likely larger than the elements in $\mathbf{l}$. Following [17], we use the GP-Upper Confidence Bound (UCB) function in our work, and we have

$$v_{n+1} = \arg\max(\mu(\mathbf{v}) + \sqrt{\frac{2\log(N \cdot \pi^2 \cdot (n+1)^2)}{6\delta_0}}\delta(\mathbf{v})), \tag{5}$$

where $\mu(\mathbf{v})$ and $\delta(\mathbf{v})$ are the mean and standard deviation of $\mathbf{v}$, N is the size of searching space, and δ_0 is a sampling parameter. We add v_{n+1}, l_{n+1} into $\mathbf{v}$ and $\mathbf{l}$, then re-compute the GP model and sample the next PI position until successfully attack f_m. The searching algorithm is summarized in Alg. 1.

2.2. Perturbation Generation

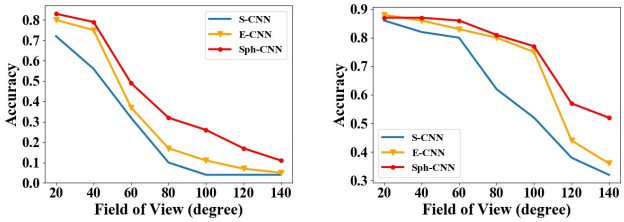
2.2.1. Perturbation-Unconstrained (PU)-SP Attack

The PU Attack does not require the perturbed image to be visually similar to the original image, which is also a general practice in previous work towards sparse attack.

For a PI x^p, we firstly replace it with a Gaussian noisy image which has the same size with x^p. Considering the high dimension of the image, we utilize random search to optimize the noisy image to attack the target model. Assuming the searching space of all pixels is V^p, in every iteration, we randomly select a combination of the pixels from V^p, and if the new perturbed image leads to a larger L, we update x^p with it, otherwise, we keep x^p unchanged.

Perturbations	Dataset	Attack	S-CNN	E-CNN	Sph-CNN
Unconstrained	Spherical ModelNet-40	SP Attack (0°)	0.84	0.86	0.87
		SP Attack (60°)	0.78	0.81	0.82
		LOAP (60°)	0.8	0.83	0.85
		LOAP	0.72	0.75	0.79
		PU-SP Attack	0.32	0.37	0.49
	Indoor Scene Panorama	SP Attack (0°)	0.84	0.86	0.86
		SP Attack (60°)	0.79	0.82	0.83
		LOAP (60°)	0.85	0.86	0.86
		LOAP	0.75	0.77	0.80
		PU-SP Attack	0.38	0.47	0.55
Constrained	Spherical ModelNet-40	SP Attack (0°)	0.88/0.86/0.78/0.76	0.9/0.89/0.79/0.77	0.87/0.86/0.86/0.81
		SP Attack (60°)	0.88/0.82/0.7/0.68	0.9/0.86/0.75/0.68	0.87/0.82/0.79/0.71
		LOAP (60°)	0.88/0.88/0.88/0.84	0.9/0.89/0.88/0.85	0.87/0.87/0.85/0.83
		LOAP	0.88/0.84/0.8/0.78	0.9/0.88/0.85/0.77	0.87/0.85/0.82/0.77
		PC-SP Attack	0.88/0.58/0.46/0.42	0.9/0.79/0.61/0.49	0.87/0.85/0.73/0.59
	Indoor Scene Panorama	SP Attack (0°)	0.90/0.87/0.81/0.77	0.91/0.89/0.88/0.79	0.91/0.90/0.88/0.82
		SP Attack (60°)	0.90/0.84/0.73/0.70	0.91/0.85/0.77/0.71	0.91/0.81/0.75/0.71
		LOAP (60°)	0.90/0.90/0.89/0.87	0.91/0.90/0.89/0.87	0.91/0.90/0.89/0.85
		LOAP	0.90/0.85/0.82/0.78	0.91/0.89/0.87/0.79	0.91/0.90/0.89/0.81
		PC-SP Attack	0.90/0.62/0.51/0.44	0.91/0.81/0.65/0.54	0.91/0.86/0.79/0.66

Table 1. Model accuracy against the PU-SP attacks and PC-SP attacks (with $\epsilon = 0/0.03/0.06/0.09$)



(a) (b)

Fig. 2. The attacking performance of (a) PU-SP Attack and (b) PC-SP Attack ($\epsilon = 0.09$) with the increasing FoV size.

2.2.2. Perturbation-Constrained (PC)-SP Attack

Compared to PU Attack, PC Attack aims for generating adversarial ODIs whose perturbations are imperceptible to human, thus the magnitude of the perturbation is limited by a parameter ϵ. To this end, we introduce a submodel f_s trained with PIs, and iteratively update the PI with its gradient. For the iteration t, the perturbed version of x^p is calculated by

$$x_t^p = x_{t-1}^p + \alpha \frac{\nabla_{x_{t-1}^p} L_c(f_m(x_{t-1}^p), y_c)}{||\nabla_{x_{t-1}^p} L_c(f_s(x_{t-1}^p), y_c)||}, \quad (6)$$

where ∇ is the differential operator, L_c is the loss function of f_s, and α controls the magnitude per iteration, after several iterations, the total perturbations are lower than ϵ finally.

2.3. Inverse Perspective Projection

For a given PI position P in spherical coordinates (θ, ϕ) on the sphere. If the Field of View (FoV) $f_h \times f_w$ and desired perspective resolution $h \times w$ are set, the mapping relation between the 2D coordinate (u, v) on the PI and the spherical coordinate can be obtained by the rectilinear projection.

Therefore, obtaining the perturbed PI x_a^p, we can re-project it onto the original ODI. As for the areas in ODI irrelevant to the selected PI, we keep them unchanged. Consequently, the final adversarial ODI $\widetilde{x_a^s}$ is calculated by

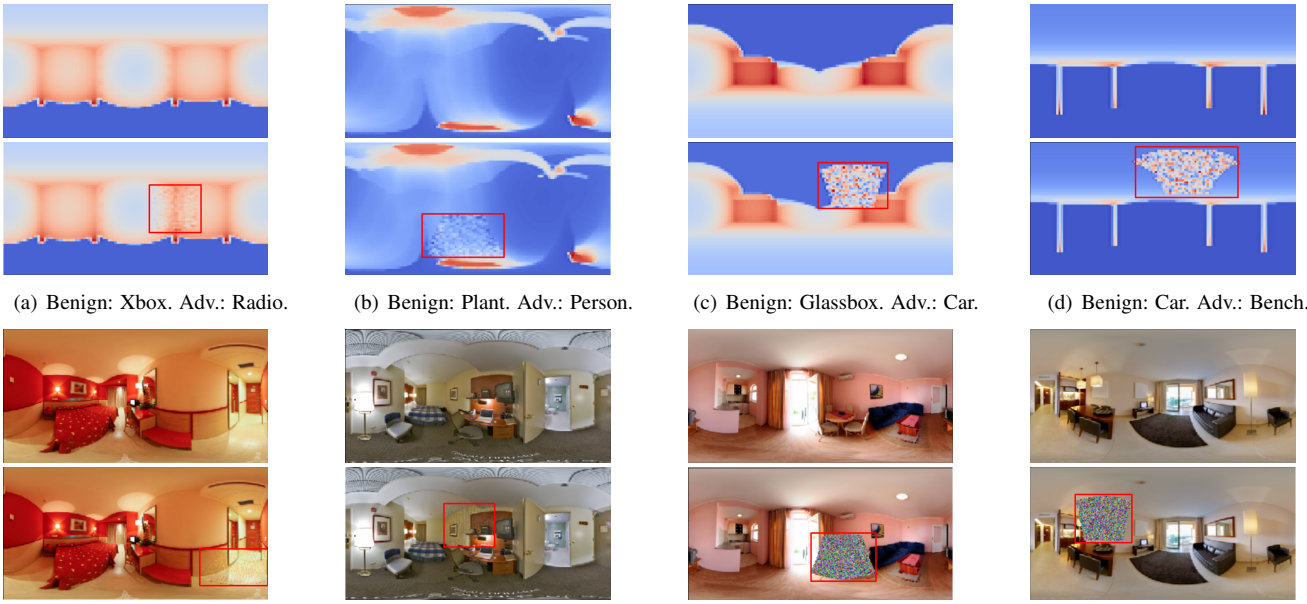
$$\widetilde{x_a^s} = M_s \odot x^s + x_a^s, \quad (7)$$

where $\odot$ is the element-wise multiplication operation, x_a^s is the region covered by the re-projection of x_a^p, and M_s is a mask formulated as

$$M_s(i,j) = \begin{cases} 0 & \text{if } (i,j) \text{ is covered by } x_a^s \\ 1 & \text{if } (i,j) \text{ is not covered by } x_a^s \end{cases} \quad (8)$$

3. EXPERIMENTS

We evaluate the SP Attack on two datasets. The first is the spherical ModelNet-40, used for shape classification, generated by projecting the ModelNet-40 dataset [18] onto the sphere. The second is an indoor scene dataset composed of PanoContext [19] and Standford 2D3D [20]. As there are no previous studies towards the attack on models for ODIs, we modify SP Attack by performing it on fixed latitudes (0° and 60°), and take them as two baselines (Fixed SP Attack (0°/60°)) to evaluate the effectiveness of the PI position searching strategy. In addition, for evaluating the necessity of conducting attack on the perspective domain rather than panorama, we consider an effective sparse attack LOAP [15] designed for planar images, and construct two baselines by directly performing it on the panorama (LOAP) or fixing its patches on the latitude of 60° (Fixed LOAP (60°)). In the attack, we consider three target models, including Spherical CNN (Sph-CNN) [6], EquiConv CNN (E-CNN) [9], and a Standard CNN (S-CNN) taking panoramas as inputs.



(a) Benign: Xbox. Adv.: Radio. (b) Benign: Plant. Adv.: Person. (c) Benign: Glassbox. Adv.: Car. (d) Benign: Car. Adv.: Bench.

(e) Benign: Bedroom. Adv.: Auditorium. (f) Benign: Bedroom. Adv.: Office. (g) Benign: Livingroom. Adv.: Bedroom. (h) Benign: Livingroom. Adv.: Conference room.

Fig. 3. Visualization of the adversarial ODIs generated by PU-SP attack (the left 4 pairs images) and PC-SP attack (the right 4 pairs of images, $\epsilon = 0.09$). In each subfigure, the upper is the benign image, and the lower is the adversarial image.

3.1. Attacking Performance

Firstly, we evaluate the effectiveness of PU-SP Attack, and the results are shown in Tab. 1. Our attack outperforms the baselines with $30\% \sim 50\%$ decline of accuracy. Specifically, the accuracies of the three victim models are around $0.3 \sim 0.5$ under the PU-SP Attack, while they are all above 0.7 under other attacks. Besides, the results of the PC-SP Attack are also shown in Tab. 1. It can be seen that all the attacks more or less affect the prediction of the victim models, although the models perform well on benign images. Among all attacks, the attack capability of the SP Attack is obviously superior to that of the baselines. Compared to the two LOAP-based attacks, the superiority of our attack indicates that attacking from the perspective domain is more aggressive than directly attacking from panorama, which is because that the PIs suffer from less deformation than panoramas. In addition, the accuracy under SP Attack is lower than that under SP Attack with fixed latitudes, showing that our proposed BO-based searching algorithm selects efficiently the desirable PI position on the whole ODI for attacking. Some examples generated by SP Attack are shown in Fig. 3. It can be observed that the perturbations are limited in a small region, but eventually mislead the prediction of the models. Besides, the patches disturbed by PC-SP Attack are visually similar to their preimages, which makes our attack more covert and practical.

3.2. Impact of FoV Size on Attacking Performance

We further measure the impact of the FoV on the performance of SP Attack with the Spherical ModelNet-40 dataset. It can be seen from Fig. 2 that the target models still perform well against the SP Attacks generated from the narrow FoVs, such as 20° and 40°. However, when the size of FoV increases, the accuracy of the models rapidly drops. When FoV size is equal to 120°, the accuracy on all the models of the attacks nearly reaches 0. We also notice that the model accuracy against PU-SP Attack declines faster than PC-SP Attack, which indicates that the PU-SP Attack is more powerful than the PC-SP Attack, because it is not expected to reserve the visual quality of the perturbed area on the ODI.

4. CONCLUSION

We investigate the vulnerability of DNNs processing ODIs against the adversarial attack with a novel sparse attack that only perturbs a single PI rendered from the target ODI. In order to select a proper PI to launch the attack, we propose a PI position selecting approach based on Bayesian Optimization, and also design two methods to perturb the selected PI. A systematic experiment demonstrates the effectiveness of the proposed attack.

5. ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China under Grant 61771469 and the Cooperation project between Chongqing Municipal undergraduate universities and institutes affiliated to CAS (HZ2021015).

6. REFERENCES

[1] Chong Xiang, Charles R Qi, and Bo Li, "Generating 3d adversarial point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Jong-Chyi Su, Matheus Gadelha, Rui Wang, and Subhransu Maji, "A deeper look at 3d shape classifiers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[3] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon, "Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan, "Vision-based navigation with language-based assistance via imitation learning with indirect intervention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[5] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al., "Spherical cnns on unstructured grids," in *International Conference on Learning Representations (ICLR)*, 2019.

[6] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis, "Learning so (3) equivariant representations with spherical cnns," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[7] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling, "Spherical cnns," in *International Conference on Learning Representations (ICLR)*, 2018.

[8] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Clara Fernandez-Labrador, José M Fácil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and José J Guerrero, "Corners for layout: End-to-end layout recovery from 360 images," *arXiv:1903.08094*, 2019.

[10] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP)*, 2017.

[11] Francesco Croce and Matthias Hein, "Sparse and imperceivable adversarial attacks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[12] Nina Narodytska and Shiva Prasad Kasiviswanathan, "Simple black-box adversarial attacks on deep neural networks.," in *CVPR Workshops*, 2017.

[13] Aram-Alexandre Pooladian, Chris Finlay, Tim Hoheisel, and Adam Oberman, "A principled approach for generating adversarial images under non-smooth dissimilarity metrics," in *International Conference on Artificial Intelligence and Statistics*, 2020.

[14] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[15] Sukrut Rao, David Stutz, and Bernt Schiele, "Adversarial training against location-optimized adversarial patches," in *European Conference on Computer Vision (ECCV)*, 2020.

[16] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2015.

[17] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," *arXiv preprint arXiv:0912.3995*, 2009.

[18] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[19] Y. Zhang, S. Song, T. Ping, and J. Xiao, "Panocontext: A whole-room 3d context model for panoramic scene understanding," in *European Conference on Computer Vision (ECCV)*, 2014.

[20] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv: 1702.01105*, 2017.