VARIATIONAL DEPTH ESTIMATION ON HYPERSPHERE FOR PANORAMA

Jingbo Miao[⋆†] *Yanwei Liu*[⋆⋆] *Kan Wang*[⋆†] *Jinxia Liu*[‡] *Zhen Xu*[⋆]

[⋆] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[†]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[‡]Zhejiang Wanli University, Ningbo, China

ABSTRACT

Depth estimation for panorama is a key part of 3D scene understanding, and adopting discriminative models is the most common solution. However, due to the rectangular convolution kernel, these existing learning methods cannot efficiently extract the distorted features in panoramas. To this end, we propose OmniVAE, a generative model based on Conditional Variational Auto-Encoder (CVAE) and von Mises-Fisher (vMF) distribution, to strengthen the exclusive generative ability for spherical signals by mapping panoramas to hypersphere space. Further, to alleviate the side effects of manifold-mismatching caused by non-planar distribution, we put forward the Atypical Receptive Field (ARF) module to slightly shift the receptive field of the network and even take the distribution difference into account in the reconstruction loss. The quantitative and qualitative evaluations are performed on real-world and synthetic datasets, and the results show that OmniVAE outperforms the state-of-the-art methods.

Index Terms— Panorama, depth estimation, hypersphere space, conditional variational auto-encoder

1. INTRODUCTION

In recent years, predicting depth from omnidirectional images has been widely demanded for 3D reconstruction applications However, early studies, including manually defining monocular cues [1, 2] and next learning-based methods [3, 4], are designed only for perspective images.

The rich semantic information in omnidirectional images conduce to improving the prediction accuracy. But due to the impact of projection distortion, depth estimation for panorama becomes more challenging. Omnidepth [5] follows an autoencoder structure, using two networks to extract planar and spherical features, respectively. Jin et al. [6] adopt layout information, and Wang et al. [7] use the inputs of multiple projection formats to construct a depth extraction framework. The encoder network designed by Sun et al. [8] compresses the equirectangular projection (ERP) image longitudinally as the disentangled representation for various image tasks. Most of these existing methods build discriminative models based on rectangular convolution kernel. Nevertheless, the receptive field of the traditional square convolution is subject to Gaussian distribution [9] after several iterations. As an ERP image naturally represents the spherical signal, the potential receptive mode on a plane cannot accurately extract spherical-shaped features or be customized in discriminative structures. Though the classic conditional variational auto-encoder (CVAE) [10] is one way to summarize signals onto a specific space explicitly [11, 12], most of these methods, being improved under the premise of the conventional Gaussian distribution, are limited to perspective images and inefficient for spherical panoramas.

In order to solve the above issues, we propose a generative model OmniVAE based on CVAE with von Mises-Fisher (vMF) distribution for panorama depth estimation. For one thing, the specific encoder-decoder scheme of VAE contributes to getting a clearer patch boundary and explicitly defining representation space; for another, the adopted CVAE models well the mapping from one input to many possible outs by multiple sampling, which is helpful to solve the uncertainty and ambiguity in depth inference. The main contributions are summarized as follows:

- We propose a generative model for panorama depth estimation, generating latent variables on a hypersphere space with the more appropriate vMF distribution.
- We put forward an atypical encoder to slightly disrupt the receptive field of the standard convolution to adapt to the discomfort caused by the distribution changes.
- We limit the depth value decoded from hyperspherical hidden variables to linear space by extending the distribution difference to the reconstruction loss.

2. PROPOSED METHODS

2.1. OmniVAE for Panorama Depth Estimation

The classic structure of VAE compresses the observed sample $\mathbf{x}$ into specific latent variables $\mathbf{z}$ (usually subject to Gaussian

*Corresponding author. This work was supported by the Cooperation project between Chongqing Municipal undergraduate universities and institutes affiliated to CAS (HZ2021015).

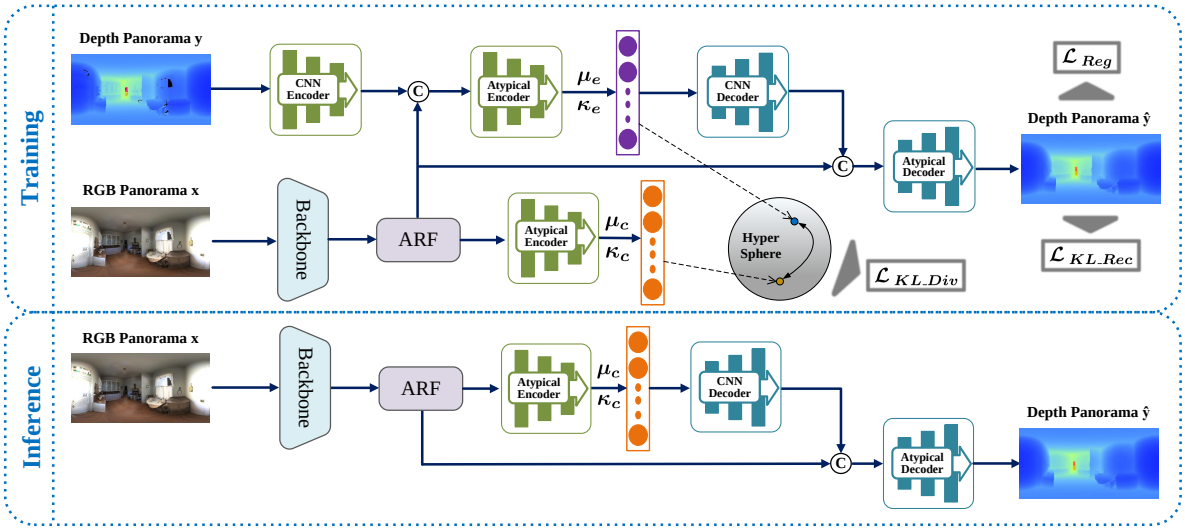

Fig. 1: An overview of the proposed OmniVAE framework for panorama depth estimation. '©' denotes concatenation. The inputs include the conditional RGB panorama $\mathbf{x}$ and the corresponding ground truth $\mathbf{y}$ during the training stage. Features of the two inputs are extracted separately and then merged to be encoded onto the hypersphere space. After fusing with the ARF features, the latent variables are processed by the decoder to infer depth map $\hat{\mathbf{y}}$. In the inference stage, the latent vector from the prior network is re-parameterized to participate in the decoding process. Note that we use the Monte Carlo (MC) sampling to draw multi-samples and take the average value of the outputs as the final result.

distribution $p(\mathbf{z})$ on a plane) through the encoder $q_\psi(\mathbf{z}|\mathbf{x})$ (a parameterized network). It then reconstructs $\mathbf{x}$ by decoder $p_\phi(\mathbf{x}|\mathbf{z})$.

Actually, we need to generate many possible candidate depth values for one pixel in a panorama for the depth estimation task. Accordingly, we adopt the CVAE model as the basic structure, as shown in Fig. 1. The feature from the depth map $\mathbf{y}$ is extracted by the CNN-Encoder $q_\psi(\mathbf{z}|\mathbf{x},\mathbf{y})$ conditioned by the original RGB image $\mathbf{x}$ and then be squeezed onto a hypersphere as the posterior distribution of latent variables $\mathbf{z}$. Here we choose the *von Mises-Fisher* (vMF) distribution,

$$F_d(\boldsymbol{\mu},\boldsymbol{\kappa}) = \frac{\boldsymbol{\kappa}^{d/2-1}e^{\boldsymbol{\kappa}\boldsymbol{\mu}'\mathbf{z}}}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\boldsymbol{\kappa})}, \mathbf{z}\in\mathcal{R}^{d-1}, \qquad (1)$$

as the priori to the hidden variables since panorama is derived from spherical imaging. The parameter $\mathcal{R}^{d-1}$ indicates d-sphere, $\boldsymbol{\mu}$ the mean direction, and $\boldsymbol{\kappa}$ the concentration. where $\mathcal{I}_v$ indicates the modified Bessel function of the first kind traversed in the direction v. We parameterize the distorted features by the vMF manifold since panoramas essentially represent the spherical signal. Note that vMF, as mentioned in $\mathcal{S}$-VAE [13], is often taken as the normal Gaussian distribution on a hypersphere and therefore more conducive to expressing spherical information and eliminating the soap bubble effect.

The inputs of OmniVAE during the inference stage only contain the conditional part (i.e., RGB panorama). We design a prior network to generate the prior distribution $F_d(\boldsymbol{\mu_c},\boldsymbol{\kappa_c})$ from the conditional RGB panorama $\mathbf{x}$, which yields better reconstruction results by approximating the posterior distribution $F_d(\boldsymbol{\mu_e},\boldsymbol{\kappa_e})$ produced by the encoder during the training stage.

As the conditional part, feature extraction of RGB panoramas is carried out in both the training and inference stages and even the only information source during the inference stage. In order to improve the utilization of these features, we use the relatively more complex module ARF for feature extraction with the shifted receptive field. The receptive field of the following convolution, called Atypical Encoder (Decoder), will be slightly irregular in contrast to the standard CNN Encoder (Decoder). It should be emphasized that the term atypical refers to the variation of the feature map receptive field at each stage of the network rather than the variation of the network structure. The feature from ARF is directly processed by the decoder, as shown in Fig. 1, which is also a critical factor for the encoder-decoder model with skip connection to restore a clear target boundary throughout the network path.

2.2. Atypical Receptive Field (ARF) Module

The receptive field of the feature map after multi-layer traditional convolution is subject to the Gaussian distribution as described in Section 1. Still, our goal is to represent the panorama signal in a hypersphere space, which brings the manifold-mismatch problem. As a result, we add more nonlinear characteristics to the feature extraction network through an attention-based weighted-RFs structure called the atypical

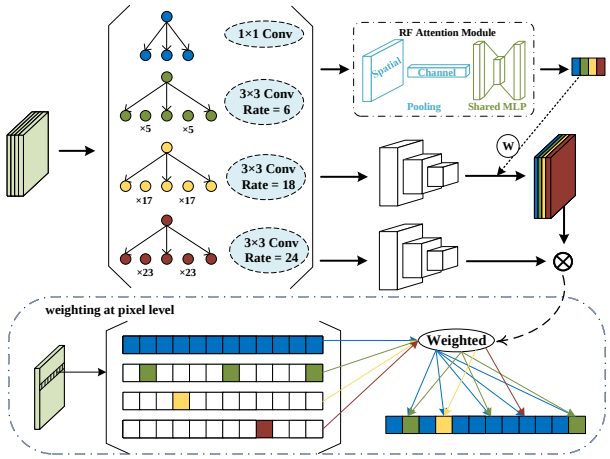

Fig. 2: An overview of the ARF Module. '$\textcircled{W}$' indicates weighting operation, and '$\otimes$' indicates correlation computation. The RF attention module weights the feature map with different sizes of RFs from multi-scale atrous convolution. Visualization of the weighting at the pixel level is shown in the dotted frame.

receptive field module to offset the receptive fields. Through the atypical nonlinearity of the receptive field, the network can express the panorama signal on the hypersphere.

After the backbone network, as shown in Fig. 2, multi-scale atrous convolution [14] is employed to produce feature maps with different sizes of receptive fields. After concatenation along the channel dimension, these feature maps are processed by the RF attention module to generate the corresponding weights for each atrous rate. There are two pooling operations in the RF attention module—along the channel and spatial dimension successively—to obtain RF-granularity weights before the MLP (Multilayer Perception). Then the similarities between the weighted feature maps and the original ones (we choose matrix multiplication here) are computed to generate the final outputs. In order to eliminate aliasing effects, we add additional convolutional layers before weighting and correlation operations. Fig. 2 also shows the pixel-level weighting results. The weights obtained from the attention module determine the contribution of different RFs to the same pixel.

2.3. Loss Function

Taking RGB panorama as the condition, the ELBO is formatted as:

$$\begin{aligned} ELBO = E_{\mathbf{z}\sim q_\psi(\mathbf{z}|\mathbf{x},\mathbf{y})} \log p_\phi(\mathbf{y}|\mathbf{x},\mathbf{z}) \\ - D_{\mathrm{KL}}(q_\psi(\mathbf{z}|\mathbf{x},\mathbf{y})\|p_\phi(\mathbf{z}|\mathbf{x})), \end{aligned} \tag{2}$$

which is derived from standard VAE. D_{KL} is the function of Kullback-Leibler (KL) divergence, and $\mathbf{z}$ is a differentiable reparameterization transformation of the output of the prior network. We minimize the KL divergence between the posterior distribution from the depth map $\mathbf{y}$ conditioned by RGB $\mathbf{x}$ and the prior distribution generated by $\mathbf{x}$ only as:

$$\mathcal{L}_{KL\_Div} = \int q_\psi(\mathbf{z}|\mathbf{x},\mathbf{y}) \log \frac{q_\psi(\mathbf{z}|\mathbf{x},\mathbf{y})}{p_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}. \tag{3}$$

Since we explicitly fix the latent variables of the panorama on the hypersphere space, we can reasonably assume that the output $\hat{\mathbf{y}}$ is subject to spherical distribution $q_s(\hat{\mathbf{y}})$, while the depth value $\mathbf{y}$ is subject to the Dirac distribution $p_\delta(\mathbf{y})$ on the linear space. More specifically, it should be taken into account that the depth information in nature follows a linear smooth variation but we transfer the image signal to the hypersphere, which introduces the problem of misfitting with the linear distribution. Therefore, we constrain the depth value in the loss function by splitting the reconstruction loss into KL divergence and the regression loss inspired by KL Loss [15]. The total loss function is formalized as:

$$\begin{aligned} \mathcal{L} &= \alpha \cdot \mathcal{L}_{KL\_Div} + \beta \cdot \mathcal{L}_{KL\_Rec} + \mathcal{L}_{Reg} \\ &= \alpha \cdot \mathcal{L}_{KL\_Div} + \beta \cdot \int_{\mathbf{y}} q_s(\hat{\mathbf{y}}) \log \frac{q_s(\hat{\mathbf{y}})}{p_\delta(\mathbf{y})} d\mathbf{y} + \|\mathbf{y} - \hat{\mathbf{y}}\|, \end{aligned} \tag{4}$$

where the balance parameters α and β are empirically set to 0.5 and 0.1 according to their respective importance.

3. EXPERIMENTS

3.1. Datasets and Settings

We evaluate the proposed OmniVAE with two datasets, Stanford2D3D [16] and 3D60 [5]. The real-world dataset Stanford2D3D contains 1413 panoramas, and the fifth area is used for testing. Provided by Omnidepth, 3D60 is generated from real-world datasets Stanford2D3D, Matterport3D [17], and synthetic dataset SUNCG [18]. According to the official guidance, we use the provided script to split the training and testing sets. The network is trained on NVIDIA TITAN XP (12G) GPU with the batch size of 10. We choose Xception [19] as the backbone and optimize the network by stochastic gradient descent (SGD) with the momentum of 0.9 and the learning rate of $1e^{-3}$ decaying by a rate of $5e^{-4}$.

3.2. Performance Comparison

We use the common depth estimation error metrics—MAE, MRE, RMSE, RMSElog, and three accuracy metrics δ with different thresholds, to compare the performance of the proposed network with that of the state-of-the-art discriminative models, OmniDepth [5] and BiFuse [7]. As shown in Table 1, OmniVAE takes the lead in estimation performances of all three methods. Specifically, OmniVAE reduces the MAE by more than 22% and improves the accuracy metric of $\delta < 1.25$

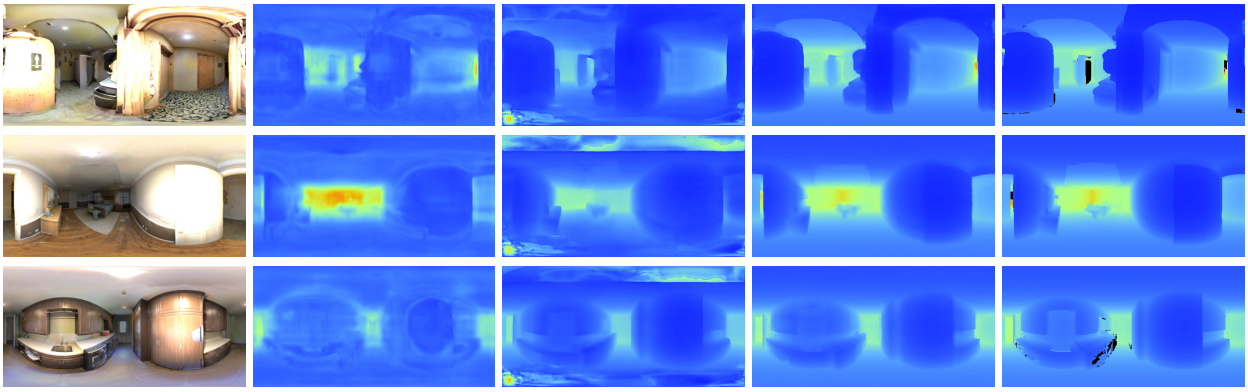

Fig. 3: Qualitative comparisons of the estimated dense depth on 3D60.

Table 1: Quantitative comparisons on Stanford2D3D and 3D60. The numbers in bold indicate the best results.

Dataset	Method	MAE	MRE	RMSE	RMSE(log)	$\delta_{1.25^1}$	$\delta_{1.25^2}$	$\delta_{1.25^3}$
Stanford2D3D	OmniDepth [5]	0.3743	0.1996	0.6152	0.1212	0.6877	0.8891	0.9578
	BiFuse [7]	0.2343	0.1209	0.4142	0.0787	0.8660	0.9580	0.9860
	Ours	**0.1814**	**0.0917**	**0.2876**	**0.0515**	**0.9354**	**0.9928**	**0.9980**
3D60	OmniDepth [5]	0.1706	0.0931	0.3171	0.0725	0.9092	0.9702	0.9851
	BiFuse [7]	0.1143	0.0615	0.2440	0.0428	0.9699	0.9927	0.9969
	Ours	**0.0994**	**0.0447**	**0.2156**	**0.0358**	**0.9728**	**0.9940**	**0.9979**

Table 2: Ablation study results on Stanford2D3D.

Method	CVAE (Gaussian)	OmniVAE (vMF)	OmniVAE (+ARF)	OmniVAE (+KL_Rec)
MAE	0.3637	0.3095	0.2460	0.2892
RMSE	0.4426	0.3924	0.3371	0.3748
$\delta_{1.25^1}$	0.8009	0.8553	0.9036	0.8733

by 6.94% on Stanford2D3D, compared to BiFuse. It should be noted that the performance of OmniVAE is achieved by inputs of ERP format-only, while BiFuse combines the ERP and Cube map formats of omnidirectional images. Besides that, since 3D60 comprises multi-modal stereo renders of scenes from realistic and synthetic datasets, the outperformance on it indicates our proposal has strong generalization capability.

Fig. 3 shows the qualitative comparisons of several samples, the dark region of the ground truth indicates unavailable depth. We use the officially provided model for testing. For better observation, we employ color mapping to pseudo-colorize grayscale depth images. Given the distinct distortion levels in different latitudes in ERP, we can intuitively see that OmniVAE makes the boundary clearer and better predicts fine-grained details than other methods in various positions thanks to the hypersphere manifold and the shiftable receptive field.

3.3. Ablation Studies

We perform ablation studies on Stanford2D3D. The baseline is the classic CVAE model with Gaussian distribution. As shown in Table 2, OmniVAE compresses latent variables on hypersphere to improve the accuracy of $\delta < 1.25$ and reduce the MAE by more than 0.05. Due to the better perception of spherical signal, the ARF module significantly improves the performance, reducing MAE and RMSE by more than 0.06 and 0.05, respectively. It also indicates that strengthening the feature extraction ability of the conditional branch can effectively improve the performance of CVAE. Moreover, introducing KL Loss to reconstruction loss enhances the accuracy metric by nearly 2%.

4. CONCLUSION

We propose a generative model OmniVAE for 360° panorama depth estimation. We encode the image information onto the hypersphere space to reduce the side effects of distortion and introduce the ARF module for the encoding part of the network. In addition, we consider the KL divergence between the vMF and Dirac distribution in the loss function to linearize the outputs. The experimental results show that OmniVAE can achieve better performance than the state-of-the-art methods in multiple metrics on real-world and synthetic datasets.

5. REFERENCES

[1] Beyang Liu, Stephen Gould, and Daphne Koller, "Single image depth estimation from predicted semantic labels," in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2010, pp. 1253–1260.

[2] Ashutosh Saxena, Min Sun, and Andrew Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 5, pp. 824–840, 2009.

[3] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab, "Deeper depth prediction with fully convolutional residual networks," in *Conference on 3D Vision (3DV)*, 2016, pp. 239–248.

[4] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao, "Structure-guided ranking loss for single image depth prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 611–620.

[5] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 453–471.

[6] Lei Jin, Yanyu Xu, Jia Zheng, Junfei Zhang, Rui Tang, Shugong Xu, Jingyi Yu, and Shenghua Gao, "Geometric structure based and regularized depth estimation from 360 indoor imagery," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 889–898.

[7] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai, "Bifuse: Monocular 360 depth estimation via bi-projection fusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 462–471.

[8] Cheng Sun, Min Sun, and Hwann-Tzong Chen, "Hohonet: 360 indoor holistic understanding with latent horizontal features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2573–2582.

[9] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Conference on Neural Information Processing Systems (NIPS)*, 2016, vol. 29, pp. 4898–4906.

[10] Kihyuk Sohn, Xinchen Yan, and Honglak Lee, "Learning structured output representation using deep conditional generative models," in *Conference on Neural Information Processing Systems (NIPS)*, 2015, vol. 28, pp. 3483–3491.

[11] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus H. Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," in *Conference on Neural Information Processing Systems (NIPS)*, 2018, vol. 31, pp. 6965–6975.

[12] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua, "Cvae-gan: Fine-grained image generation through asymmetric training," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2764–2773.

[13] Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak, "Hyperspherical variational auto-encoders," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018, pp. 856–865.

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 4, pp. 834–848, 2018.

[15] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang, "Bounding box regression with uncertainty for accurate object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2888–2897.

[16] Iro Armeni, Sasha Sax, Amir Roshan Zamir, and Silvio Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv:1702.01105*, 2017.

[17] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," in *International Conference on 3D Vision (3DV)*, 2017, pp. 667–676.

[18] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser, "Semantic scene completion from a single depth image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 190–198.

[19] Francois Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.