Cross-Domain Synthetic-to-Real In-the-Wild Depth and Normal Estimation for 3D Scene Understanding

Jay Bhanushali[1] Manivannan Muniyandi[1] Praneeth Chakravarthula[2]
[1]Indian Institute of Technology Madras
[2]UNC Chapel Hill

Abstract

We present a cross-domain inference technique that learns from synthetic data to estimate depth and normals for in-the-wild omnidirectional 3D scenes encountered in real-world uncontrolled settings. To this end, we introduce UBotNet, an architecture that combines UNet and Bottleneck Transformer elements to predict consistent scene normals and depth. We also introduce the OmniHorizon synthetic dataset containing 24,335 omnidirectional images that represent a wide variety of outdoor environments, including buildings, streets, and diverse vegetation. This dataset is generated from expansive, lifelike virtual spaces and encompasses dynamic scene elements, such as changing lighting conditions, different times of day, pedestrians, and vehicles. Our experiments show that UBotNet achieves significantly improved accuracy in depth estimation and normal estimation compared to existing models. Lastly, we validate cross-domain synthetic-to-real depth and normal estimation on real outdoor images using UBotNet trained solely on our synthetic OmniHorizon dataset, demonstrating the potential of both the synthetic dataset and the proposed network for real-world scene understanding applications. The dataset and accompanying code are available at omnihorizon.github.io.

1. Introduction

The task of estimating depth from omnidirectional images using a single camera has received significant attention in recent years [8, 19, 20, 27, 32, 41, 55]. It comes with specific challenges, such as handling distortions from equirectangular projections, and the quality and diversity of datasets are crucial for reliable depth estimation [63]. Similarly, accurately estimating surface normals is vital for understanding scenes [29], especially in diverse and real-world environments [10]. In fact, previous works like GeoNet [38] and Cross-Task Consistency [59] have shown the benefits of jointly learning depth and surface normals. Despite a growing interest in realistic representations of real-world scenes, obtaining accurate per-pixel data from real omnidirectional images is challenging and expensive [7, 34, 58]. Existing synthetic datasets often focus on indoor spaces with limited depth range [62], making them less suitable for generalizing to outdoor scenarios with diverse scene components and larger depth ranges [1, 24]. While simulators like CARLA [15] and datasets like SYNTHIA [45] and Virtual KITTI [6] cater to autonomous driving applications, there is a notable absence of comprehensive omnidirectional datasets and robust methods for understanding scenes in various outdoor environments. This gap in research, particularly for in-the-wild monocular scene depth and normal estimation, remains significant.

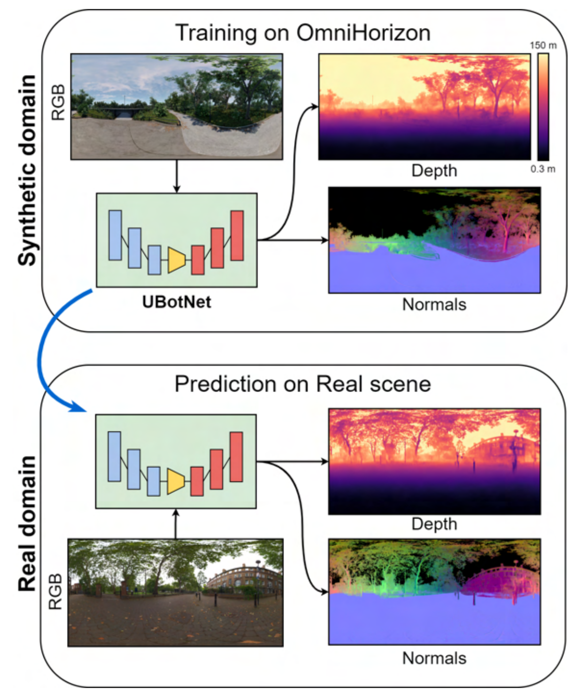

Figure 1. *Synthetic to Real cross-domain inference.* The proposed synthetic OmniHorizon dataset and the UBotNet performs cross-domain inference of scene-consistent depth and normals on real-world images captured outdoors in-the-wild.

In this work, we overcome these challenges and close the research gap by introducing a cross-domain synthetic-to-real

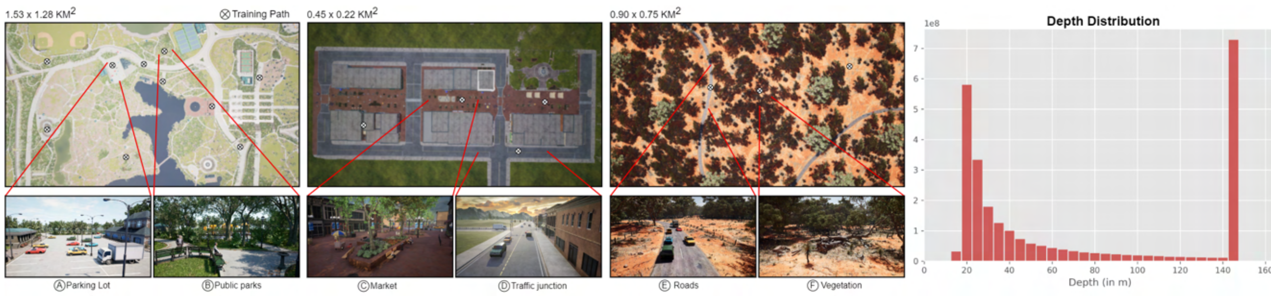

Figure 2. *Overview of the OmniHorizon dataset.* Our dataset models urban areas, vegetation and various outdoor components with pedestrians and vehicles with varied depth distribution across the scenes as visualized.

Table 1. *Comparison between the various proposed omnidirectional datasets.* While existing datasets are predominantly indoors, our proposed dataset models various outdoor environments and dynamic scene participants.

Dataset	Domain	Type	No. of panoramic views	Scene Lighting	Dynamic components
Matterport3D 360° [41]	Real	Indoors	9,684	Static	✗
Replica 360° 2k/4k RGBD [41]	Real	Indoors	130	Static	✗
Stanford 2D-3D [2]	Real	Indoors	1,413	Static	✗
PanoSUNCG [54]	Synthetic	Indoors	25,000	Static	✗
Zillow [13]	Real	Indoors	71,474	Static	✗
Fukuoka [34]	Real	Outdoors	650	Static	Vehicles & Pedestrians
OmniHorizon	Synthetic	Outdoors	24,335	Dynamic	Vehicles & Pedestrians

neural network-based approach for depth and normal estimation for the in-the-wild 3D scene understanding. To this end, we introduce a synthetic omnidirectional dataset rendered from life-sized and diverse virtual environments, featuring randomly placed scene agents (see Figure 2). The primary objective is to enable the joint estimation of scene depth and normal information across various outdoor scenarios. Subsequently, this can offer potential applications in immersive Virtual Reality [5, 42] and Visual SLAM [57]. Notably, OmniHorizon dataset includes urban environments, natural elements like vegetation and rocks, and introduces dynamic elements such as pedestrians and vehicles. Additionally, the dataset encompasses different times of day, allowing for robust depth and normal estimation under varied lighting conditions.

We aim to achieve depth and normal estimation in real-world, in-the-wild scenes using a network trained exclusively on just synthetic dataset. In doing so, we also examine and address limitations in existing neural network architectures designed for depth and normal estimation. We propose an enhanced network architecture named UBotNet, drawing inspiration from U-Net [44] and the Bottleneck transformer [49], which notably improves depth and normal estimation for both synthetic and real-world scenes. Furthermore, we conduct a thorough analysis of the cross-domain inference performance of UBotNet, trained on our OmniHorizon dataset, and the state-of-the-art Fukuoka dataset [34]. The introduced dataset and neural network demonstrate significant advancements in cross-domain inference, showcasing the capability to train the network on synthetic scenes and successfully apply it to comprehend real-world, in-the-wild scenes, see Figure 1.

In summary, we make two key contributions:

- **OmniHorizon**: We introduce a synthetic omnidirectional dataset comprising over 24,000 images, designed for comprehensive scene depth and normal estimation. This dataset is well-suited for cross-domain inference, featuring diverse landscapes, dynamic elements such as varying lighting, cloud formations, pedestrians, and vehicles.

- **UBotNet**: We propose a novel network architecture, UBotNet, inspired by U-Net and the Bottleneck Transformer. UBotNet is tailored for efficient depth and consistent scene normals estimation, demonstrating generalizability for cross-domain inference. Additionally, we introduce a streamlined variant, UBotNet Lite, with 71% fewer parameters, emphasizing compactness and efficiency in the network design.

2. Related Work

We categorize omnidirectional datasets in the literature concerning depth and normal estimation into two main groups. These are based on whether the data is collected from real-world scenarios (Real Datasets) or generated using a 3D rendering engine (Synthetic Datasets). Table 1 provides an overview of these datasets.

Real Datasets Matterport 3D [7] is a real-world dataset capturing indoor scenes, comprising of 10,800 panoramic views from 90 building-scale environments. It provides data including depth, normals, surface reconstruction, camera poses, and semantic segmentations derived from these scenes. Matterport3D 360° [41] is an extension, adding 9,684 high-resolution 360 samples specifically designed for monocular depth estimation. Gibson [58] offers a virtual environment based on real-world settings, delivering photo-realistic interiors with RGB images, depth information, surface normals, and semantic annotations for selected spaces. Stanford2D3D [2] presents a dataset gathered from six large-scale indoor areas, consisting of 1,413 equirectangular RGB images along with corresponding depths, surface normals, and additional data. HM3D [39] stands out as the most extensive dataset for 3D indoor spaces, providing 1.4 to 3.7 times the navigable space compared to other datasets. Replica [50] comprises 18 3D indoor scene reconstructions, while Replica 360° 2k/4k RGBD [41] extends this dataset, offering 130 RGB-D pairs rendered at resolutions of 2048×1024 and 4096×2048. Zillow [13] is one of the largest indoor datasets, featuring 71,474 panoramas, 21,596 room layouts, and 2,564 floor plans, all captured from 1,524 homes. Fukuoka [34], designed for place categorization challenges, is an outdoor dataset, providing 650 panoramic RGB views, 3D depth, and reflectance maps. The dataset encompasses various outdoor settings such as forests, urban areas, coastal regions, parking lots, and residential areas.

Synthetic Datasets Structured 3D [62] is a synthetic indoor dataset featuring 3,500 scenes, each offering various furniture configurations. The dataset also incorporates diverse lighting conditions, including warm and cold settings. PanoSUNCG [54] contributes 103 scenes, rendering 25,000 omnidirectional images using environments from SUNCG [48]. 360D, introduced by Zioulis et al. [63], includes 360 color images with corresponding depth, rendered from two synthetic datasets (SunCG, SceneNet [33]) and two realistic datasets (Matterport 3D, Stanford2D3D). As highlighted in Table 1, our OmniHorizon dataset stands out by encompassing outdoor virtual spaces, complete with dynamic scene lighting and diverse scene participants. This addresses a significant gap in existing datasets, which predominantly focus on indoor environments with static scene components and lack contextual information for outdoor spaces.

Monocular Omnidirectional Depth and Normals Early approaches to monocular omnidirectional depth estimation, pioneered by [52] and [63], involved adapting traditional CNNs for spherical images, either through distortion-aware training on perspective images or by introducing a rendered spherical dataset. Notably, Pano Popups [16] concurrently predicted depth and surface orientation, emphasizing the challenges in approximating planar regions. The significance of spatially imbalanced predictions in 360° depth estimation was addressed by Generalized Mapped Convolutions [17], showcasing the importance of accounting for distortion in equirectangular projections. Omnidirectional extension networks [11] introduced a near field-of-view (NFoV) perspective depth camera alongside a spherical one, enhancing detail preservation in inferred depth maps. Recent works have explored diverse paths, with approaches like BiFuse [56] and UniFuse [27] focusing on the fusion of cubemap and equirectangular features. HoHoNet [51] adapted classical CNNs for 360° images, flattening meridians to DCT coefficients for efficient dense feature reconstruction in monocular depth estimation from spherical panoramas. Other studies, including [28, 60], investigate the relationship between layout and depth estimation, while [21] explores joint optimization of depth and surface orientation using a UNet model. However, these methods encounter challenges in joint estimation of depth and normals, and generalizing to images in real-world scenarios due to the limited scene diversity inherent in existing datasets acquired by depth sensors. In addition, normals estimated using depth is considerably inaccurate in comparison to direct normal estimation [4]. In contrast, our proposed hybrid neural network architecture allows for cross-domain inference and joint estimation of depth and normals, generalizing well to in-the-wild scenes.

3. Dataset

The OmniHorizon dataset was generated using Unreal Engine 4 [18], featuring color images, stereo scene depth, and world normals in a top-bottom format, all rendered at 1024×512 resolution. Utilizing assets from the Unreal Marketplace, we designed an animated *training path* with 1521 frames captured using a moving camera for each scene (see Figure 2), resulting in 24,335 omnidirectional views for outdoor scene depth and normal estimation. Scenes were scaled appropriately, and the dataset includes underpass, stairs, uneven terrain, buildings, and pedestrians. Depth is capped at 150m (Unreal units), and world-space normals are employed for normal maps, see supplementary material for discussion. The training-validation split is 85:15, with 85% (20,685) for training and 15% (3,650) for validation. Notably, we generate an unseen scene sequence with diverse elements, including underpasses, stairs, uneven terrain, buildings, and pedestrians, isolated from training data, to serve as a test set (1520 images). Additional attributes of the dataset are discussed in subsequent subsections.

3.1. Scene Attributes

The performance of neural networks in real-world scene inferences is significantly influenced by scene attributes and context [1]. Our dataset is designed to cover a diverse range of scene attributes, including urban environments with buildings and roads, as well as naturally occurring uneven terrains,

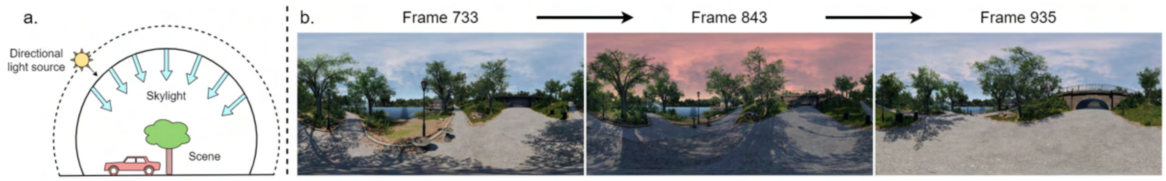

Figure 3. *Dynamic lighting and varying time of day settings.* a) The lighting of the scene is varied by modulating the directional light (sun) and secondary light source (skylight). b) Changes in the scene lighting condition achieved using the modulation of the light sources.

vegetation, and various outdoor elements. Figure 2 provides a snapshot overview of our dataset, featuring scenes like Downtown [37] and CityPark [47], which represent urban areas with buildings, houses, parks, and street props. CityPark scenes include wider roads, while Downtown scenes consist of markets, narrower streets, and alleyways. Additionally, scenes like Desert [25] feature rocks, roads, uneven terrain, and wild vegetation.

3.2. Dynamic Lighting

In outdoor environments, lighting conditions vary dynamically based on the time of day and intricate cloud patterns in the sky. Existing datasets often lack the modeling of such dynamic lighting changes, leading to compromised performance in trained neural networks, particularly for in-the-wild scene understanding tasks in outdoor settings. Note that scene depth and normals remain independent of scene brightness or color. To address this, we adopted a two-pass rendering approach, isolating scene depth and normal data from scene color. This separation allowed us to prototype changes in scene lighting, brightness, and color while ensuring consistent depth and normal data generation during rendering. We simulated dynamic lighting changes by modulating the position and intensity of both a directional light source (sunlight) and a secondary light source (diffuse light from the sky) throughout an animated sequence, mimicking a full day. Figure 3 illustrates example changes in lighting conditions achieved by modulating these light sources. To capture more complex lighting variations resulting from different cloud formations in the sky, we utilized a sky plugin [22] to render various sky-cloud settings, including Stratus, Cumulus, and Cirriform clouds. Cloud coverage was varied from very light to extremely heavy, and the dataset spans early morning to late evening time settings.

3.3. Dynamic Scene Participants

Dynamic scene elements such as vehicles and pedestrians are predominant and play crucial roles in outdoor spaces. To accurately represent these components, we modeled various classes of vehicles, including trucks, hatchbacks, SUVs, pickup trucks, and sports cars. These vehicles were randomly generated and placed in outdoor environments, as well as manually positioned in parking lots and on roads (see Figure 2). Additionally, we introduced visual diversity in

Figure 4. *Examples of pedestrians in OmniHorizon dataset.* a) virtual avatars sitting in a cafeteria, b) pedestrian walking on the street (spline path is highlighted in pink) and c) casual group hangout.

pedestrians by incorporating 3D scanned avatars [40] and high-fidelity realistic Metahumans [23]. Metahumans, with diverse skin tones and detailed grooming, were utilized at the highest Level of Detail (LOD 0) to enhance visual realism (see supplementary material). The pedestrians exhibit three distinct settings: idle poses, sitting, and walking. Walking behavior and trajectories are controlled using spline paths and Unreal Engine's blueprints. Figure 4 showcases examples of human avatars strategically placed throughout the dataset, engaging in realistic activities such as sitting outside a cafeteria, walking on the street, and engaging in group discussions.

4. Neural Cross-domain Inference

In this section, we present the UBotNet architecture, inspired from U-Net [44] and Bottleneck transformer [49], for cross-domain inference, along with our network training methodology. Our evaluation comprises four distinct experiments: a) Benchmarking on the OmniHorizon dataset, b) Ablation study of the dataset, c) Sim-to-Real domain transfer performance, and d) In-the-wild depth and normal estimation from real-world omnidirectional images.

4.1. UBotNet Architecture

In comparison to high-capacity encoders like ResNet and DenseNet architectures, UNet with skip-connections has exhibited superior performance in Pano3D benchmarks [1]. However, U-Net architecture is unsuitable for predicting consistent normals across both synthetic and real-world scenes (see Figure 6 and additional results in supplementary material). To address this limitation and enable the network to capture information in a broader context with long-range dependencies, we sought inspiration from Vision Transformers (ViT). ViT, known for achieving state-of-the-art results in image classification using a pure transformer architecture [14], has demonstrated a wider receptive field compared to CNNs,

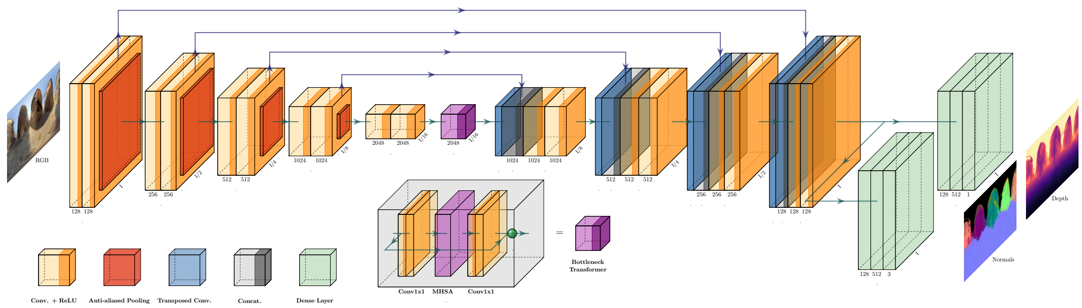

Figure 5. *Proposed UBotNet architecture.* UBotNet is a hybrid architecture based on UNet and Bottleneck Transformer (BoTNet). Anti-aliased max pooling is used for the pooling operation. The transformer block is placed in the middle of the encoder and decoder paths of the UNet. UBotNet Lite uses separable convolutions in place of standard convolution layers; otherwise, it is identical to UBotNet. A simplified illustration of BoTNet is also shown which contains Multi-Head Self-Attention (MHSA) for learning global context.

allowing global integration of information across an image. Recent studies replacing the final layers of ResNet with a bottleneck transformer have shown improved performance in instance segmentation and object detection tasks [49]. The fusion of U-Net with attention or transformer-based architectures has also been explored, particularly in medical image segmentation [9, 35].

Building upon these insights from prior research, we introduce an enhanced architecture named *UBotNet*, that can efficiently learn local features through convolutional layers and integrate self-attention for global context aggregation. UBotNet combines elements from U-Net and the Bottleneck transformer. Specifically, we position the self-attention transformer block at the lowest resolution feature maps in the U-Net bottleneck, as self-attention involves $O(n^2d)$ memory and compute [53]. Figure 5 provides an overview of our proposed architecture. To enhance its performance, we replace the traditional max-pooling layer with an anti-aliased max-pooling layer [61]. Additionally, we present a streamlined compact version of UBotNet, dubbed *UBotNet Lite*, employing separable convolution [12] to significantly reduce the number of parameters. UBotNet Lite (38.3M) has a 71.2% reduction in parameters compared to its larger counterpart, UBotNet (133M). Towards the end, we incorporate two branches of fully-connected layers with sigmoid activation to predict scene depth and consistent normals. The CNN blocks focus on capturing local image features, while the Multi-Head Self-Attention (MHSA) block from the bottleneck transformer learns global contextual features. See supplementary material for additional details. We demonstrate and validate the impact of learning local and global-scale features in the experiments detailed in Section 5.

4.2. Network Training and Experiments

We maintain the following setup and configuration for training and testing in all our benchmark, evaluation and ablation experiments.

Training configuration. We used an Nvidia RTX 3090 with 24GB onboard memory for training all network models. The batch size is set to 4 and Adam optimizer [30] is used with a learning rate of 1×10^{-4} and decay rate of 1×10^{-5}. Due to memory constraints, the images were rescaled to a resolution of 512×256 for training and evaluation. All the networks were trained for 40 epochs.

Loss Functions. The networks were trained to jointly learn both depth and normal information from the input monocular omnidirectional images. We used $\mathcal{L}_{berHu}$ (Reverse-Huber) function [31] as the loss objective for depth and $\mathcal{L}_1$ penalty as the objective function for estimating scene normals . The overall loss function for joint learning is, therefore, a sum of both depth and normal objectives:

$$\mathcal{L}_{Total} = \mathcal{L}_{Depth} + \mathcal{L}_{Normal} \tag{1}$$

Data Augmentation. We employ two techniques to augment the color data of the input images, namely Channel Shuffle [43] and Color Jitter [43]. Additionally, we use a third technique for rotation-based augmentation [26, 27].

Baseline Architectures and Evaluation Criteria. We evaluated our dataset using various architectures, including SliceNet [20], BiFuse [55], Panoformer [46], HoHoNet [51], UResNet [63], RectNet [63], UNet_{128} and the proposed UBotNet and UBotNet Lite architecture. Modifications were applied to the final layers of UResNet and RectNet to enable joint learning of depth and normals. UNet128, employing a base of 128 feature channels extending to 2048 channels, is left similar to the vanilla architecture. Networks like HoHoNet, BiFuse, SliceNet, and Panoformer were left unmodified due to their complexity, training solely for depth estimation. Our depth estimation criteria include standard metrics such as Root Mean Square Error (RMSE), Mean Relative Error (MRE), Root Mean Square Error in log space (RMSE log), and accuracy metrics ($\delta 1, \delta 2$ and $\delta 3$ with a threshold of 1.25) [1, 27, 55]. For normal estimation, evaluation criteria include metrics such as RMSE, mean, median, and accuracy metrics at angles of $5°$, $7.5°$ and $11.25°$ [1, 3, 10].

Table 2. *Quantitative Results for the benchmark evaluated on the OmniHorizon dataset.* Values in **bold** highlight best results. (* denotes networks that only perform depth estimation)

Method	# parameters	Depth Error ↓			Depth Accuracy ↑			Normal Error ↓			Normal Accuracy ↑		
		RMSE	MRE	RMSE log	$\delta1 < 1.25$	$\delta2 < 1.25^2$	$\delta3 < 1.25^3$	Mean	Median	RMSE	5.0°	7.5°	11.25°
RectNet [63]	8.9 M	0.646	23.786	1.213	0.247	0.265	0.283	9.84	5.49	14.53	48.84	56.06	65.85
UResNet [63]	50.8 M	0.097	0.487	0.260	0.424	0.614	0.768	11.50	7.18	16.32	44.50	49.01	55.73
HoHoNet* [51]	49.5 M	0.092	0.547	0.228	0.510	0.717	0.838	X	X	X	X	X	X
SliceNet* [20]	79.5 M	0.087	0.425	0.232	0.583	0.784	0.868	X	X	X	X	X	X
UBotNet Lite (Ours)	38.3 M	0.063	0.403	0.181	0.657	0.844	0.896	8.00	4.19	12.57	54.86	64.51	75.36
Bifuse* [55]	212 M	0.067	0.345	0.174	0.646	0.846	0.908	X	X	X	X	X	X
Panoformer* [46]	20.4 M	0.062	0.311	0.159	0.661	0.842	0.913	X	X	X	X	X	X
UNet_{128}	124 M	**0.052**	**0.259**	0.157	0.641	0.849	0.925	9.01	4.01	14.71	54.00	62.58	72.68
UBotNet (Ours)	133 M	0.054	0.271	**0.151**	**0.712**	**0.874**	**0.929**	**7.44**	**3.61**	**12.12**	**56.80**	**67.29**	**78.52**

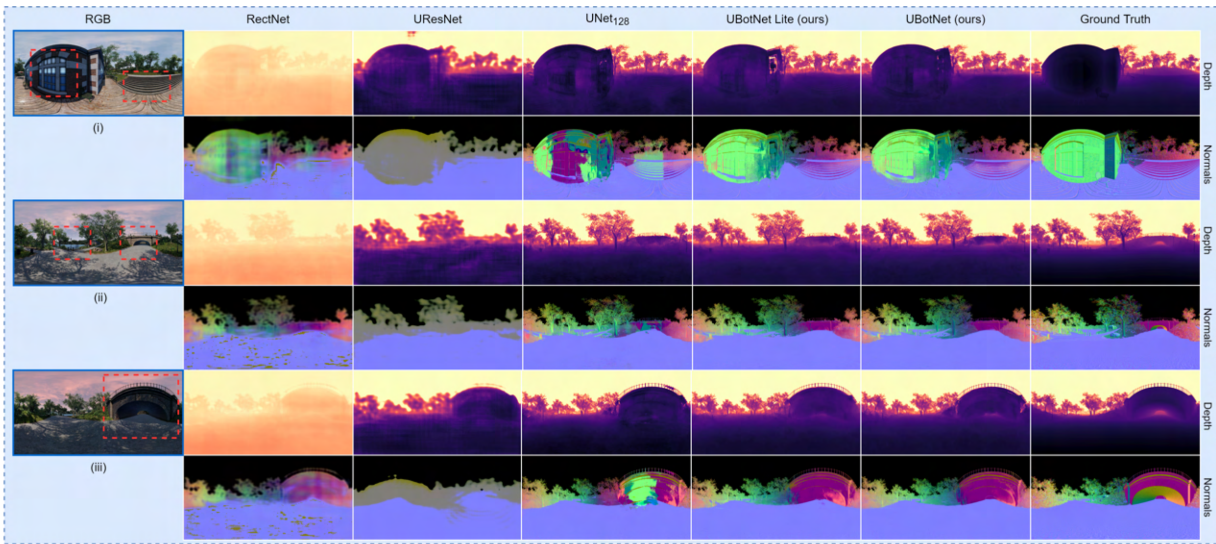

Figure 6. *Qualitative Results from the benchmark on OmniHorizon dataset.* Three different instances of varying depths and lighting conditions are compared between all networks. UBotNet performs consistently better than UNet_{128} and other architectures when estimating depth and normals. UBotNet Lite shows small artefact in depth estimates but still preserves the global context required to learn normals.

5. Discussion and Evaluation

5.1. Benchmark Results on OmniHorizon

Quantitative results. Quantitative results for depth and normal estimation across all networks are shown in Table 2. RectNet and UResNet architectures exhibit suboptimal performance, with RectNet failing to converge after early iterations. Conversely, other networks, including UBotNet, show superior outcomes in the benchmark. Notably, UBotNet consistently outperforms other architectures, including UNet_{128}, across all metrics except for Mean Relative Error (MRE). In terms of normal metrics, UBotNet demonstrates a performance improvement of 14.92% for normal error and 4.45% for normal accuracy compared to UNet_{128}. UBotNet Lite (38.3M params) performs slightly lower than UNet_{128} (124M params) for specific metrics but shows better results for normal metrics while having 70% fewer parameters.

Qualitative results. In Figure 6, visual comparisons for depth[1] and normals across various architectures are presented, showcasing their validation against Ground Truth (GT) data. The first image highlights structures in close proximity, where UBotNet exhibits superior depth estimation for building windows and stairs compared to UNet_{128}. Additionally, UBotNet provides more accurate normal estimates for the stairs and building structure. The second and third images focus on distant elements, testing the networks' ability to identify trees and underpass structures in shadows. Note that UNet_{128} struggles to identify the distant part of the tunnel in the third image, while UBotNet successfully detects it. Furthermore, UBotNet and UBotNet Lite shows estimates closer to GT and outperforms UNet_{128}, which completely fails, in estimating normals for the underpass structure in both images. This validates that the proposed architectures demonstrate strong performance in normal estimation, benefiting from the global context extracted by the Multi-Head Self-Attention (MHSA) from the encoder features. Qualitative results for other networks (marked with * in Table 2) are available in the supplementary material.

[1] Depth maps have been normalised for visualisation purposes.

Table 3. *Quantitative results for the ablation study on the OmniHorizon dataset.* Various versions of the dataset are compared by removing the dynamic elements from the scene. VP - Vehicles & Pedestrians and DL - Dynamic Lighting

	Depth Error ↓			**Depth Accuracy** ↑			**Normal Error** ↓			**Normal Accuracy** ↑		
Method	RMSE	MRE	RMSE log	$\delta1 < 1.25$	$\delta2 < 1.25^2$	$\delta3 < 1.25^3$	Mean	Median	RMSE	5.0°	7.5°	11.25°
Static	0.055	0.293	0.155	0.656	0.854	0.924	7.67	3.74	12.55	56.16	66.49	77.60
Static + VP	**0.053**	0.289	0.154	**0.713**	0.868	0.924	7.53	3.64	12.26	56.72	67.05	78.18
Static + VP + DL	0.054	**0.271**	**0.151**	0.712	**0.875**	**0.926**	**7.44**	**3.61**	**12.12**	**56.80**	**67.28**	**78.52**

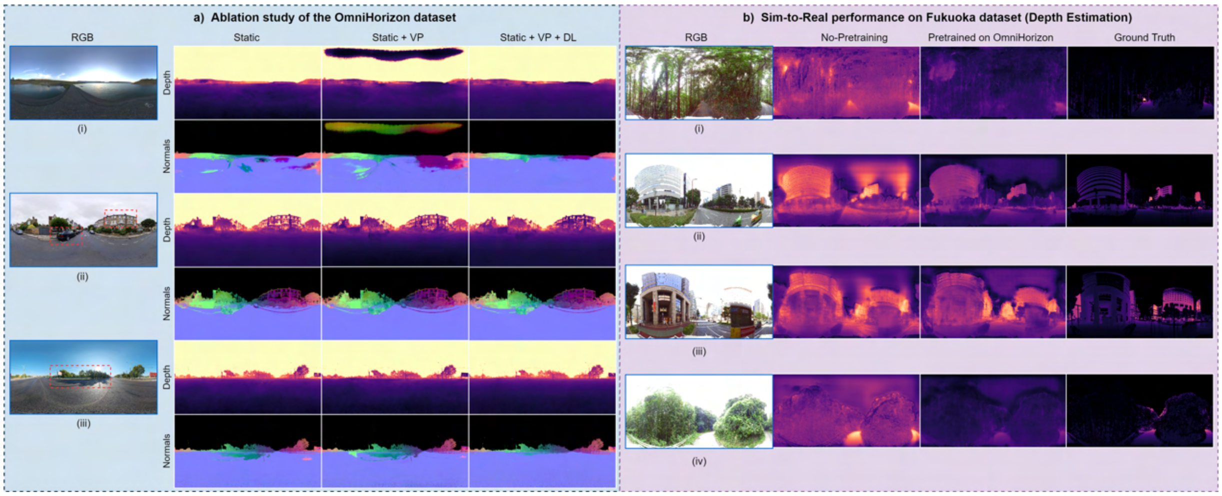

Figure 7. a) *Ablation study of the OmniHorizon dataset.* Comparison for the depth and normal estimation between the various versions of the dataset: Static, Static + VP, and Static + VP + DL. b) *Sim-to-Real performance on Fukuoka dataset.* We compare the performance of depth estimation between the network pre-trained on OmniHorizon and fine-tuned on Fukuoka against the network trained from scratch.

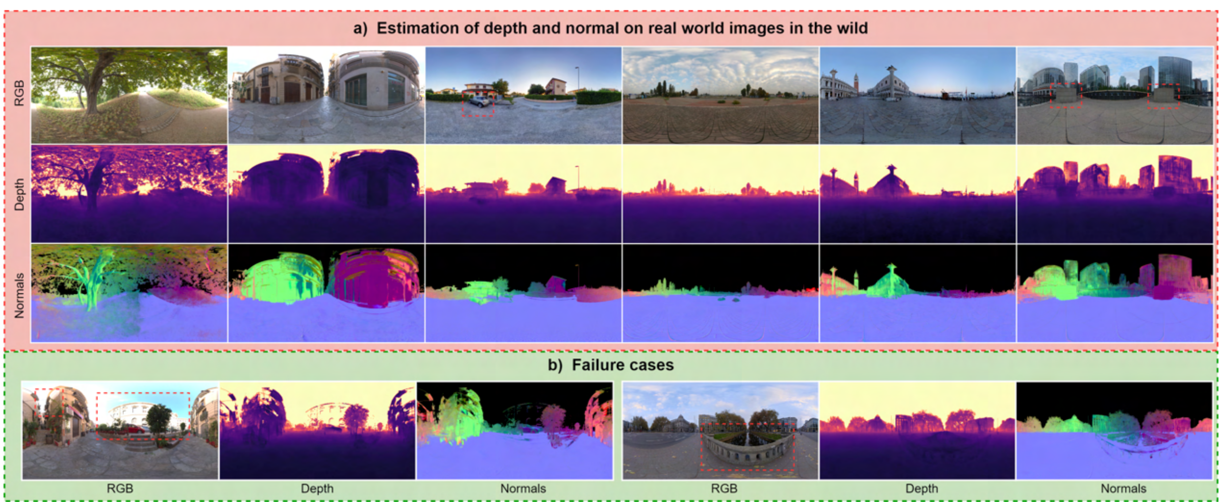

Figure 8. a) *Predictions on the real-world images in the wild.* Depth and normals estimated from real-world images representing the diverse outdoor scenarios. b) *Failure cases.* Network fails to estimate depth and normals in scenarios with overexposed regions. It also fails to recognize vertical upright structure such as the bridge railing.

5.2. Ablation Study

Our ablation study aims to address a key question: *Does context matter in outdoor scenarios?* It assesses the impact of dynamic components, specifically vehicles and pedestrians (VP), and dynamic lighting (DL), on the dataset. We create two additional dataset versions for comparison: a static version with only static meshes and no dynamic components, and a version with pedestrians and vehicles but without dynamic lighting. Note that the full OmniHorizon dataset includes all dynamic components. Table 3 shows the comparison between the various versions of the dataset using the UBotNet architecture. We observed incremental gain in the performance with the addition of dynamic components, specifically for depth accuracy and normal metrics. Figure 7 shows the visual results from the ablation study. For the first image, Static version struggle with the normal estimation for water surface while the Static+VP version faces issues with lighting and normal estimation. Differences in the vehicle on the left and building are observed in the second image between the static version and others. The last image highlights shadow artifacts present in normal maps for the static and static+VP versions, absent in the results from the full dataset. The above visual differences between the different versions of the dataset demonstrate the importance of context and dynamic elements in the outdoor scenarios. Moreover, the absence of these elements could impair the capacity of neural networks to make accurate predictions.

Table 4. *Quantitative results for Sim-to-Real performance on Fukuoka dataset after pre-training on OmniHorizon.*

	Depth Error ↓			Depth Accuracy ↑		
Model	RMSE	MRE	RMSE log	$\delta 1$	$\delta 2$	$\delta 3$
UBotNet*	0.036	0.633	0.307	0.265	0.497	0.664
UNet	0.036	0.615	0.301	0.422	0.647	0.771
Bifuse	0.034	0.638	0.289	**0.447**	**0.659**	0.773
UBotNet	**0.029**	**0.611**	**0.271**	0.424	0.655	**0.782**

*trained only on Fukuoka dataset

5.3. Sim-to-Real Transfer

We evaluate the simulation-to-real domain transfer performance of our method on a real-world dataset - Fukuoka [34]. To achieve this task, we pretrain the UBotNet on our dataset and fine-tune it on Fukuoka for the task of depth estimation. Note that Fukuoka dataset does not provide ground truth for normal data and therefore we only evaluate the depth estimates. Additionally, Fukuoka only provides 650 images for training compared to the OmniHorizon with 24,335 samples. Hence, we pre-train the network only on 2K samples ($<$ 10% of our dataset). Table 4 summarizes the performance comparison between the networks pre-trained on our dataset and that trained on Fukuoka from scratch. We noted better performance of the pretrained network specifically for depth accuracy, where we see a gain of 12.2%. We also observed more accurate depth maps estimated from the test images when compared to training from scratch on Fukuoka as shown in Figure 7. When trained from scratch, the network struggles notably with vegetation. On the other hand, it benefits from a better understanding of scenes with complex vegetation when it was pre-trained on OmniHorizon. Interestingly, we also observe a similar trend for other networks that were trained first on OmniHorizon.

5.4. Testing on the Real-world Images In-the-wild

The real-world omnidirectional images have been curated from the Polyhaven website [36] for testing the trained network on the images in the wild. We selected images that represent diverse outdoor scenarios cluttered with various objects and captured during different time of day settings. Figure 8 shows depth and normals estimated by UBotNet from the images. The images illustrate the ability of the network to estimate depth at a large range in various settings. Our network learns high level details from the vegetation (images 1, 3 and 4). This is reflected in the image 1 where the network was able to recognize the large tree in the foreground along with the walking path. It also captures the details from the cars in image 3. The network was able to identify sky region in cases with full clouds (image 2) and clean sky with no clouds(image 3 and 5). This demonstrates the advantage of the including various cloud formations and time of day settings in the dataset. The final image which shows a skating area is a good example of the ability of UBotNet to estimate normals of two upright structures (highlighted in red) in front of the buildings with a texture similar to the concrete floor. It highlights the capacity of the network to learn information in a global context to understand the orientation of normal surfaces. Overall, the network demonstrates promising results for the estimation of depth and normal on real-world images. We show additional results in the supplementary material.

5.5. Limitations

There are specific scenarios where sunlight may overexpose parts of a scene while underexposing others. In such instances, the network struggles to correctly estimate depth and normals for the overexposed parts of the scene. Additionally, the network occasionally misinterprets vertical elements like handrails and bridge supports. Figure 8 shows both such challenging scenarios where our method compromised. We discuss the assumptions of our dataset in supplementary material.

6. Conclusion

We presented a new dataset called OmniHorizon and a hybrid architecture called UBotNet for depth and normal estimation in diverse outdoor scenarios. Firstly, our dataset includes diverse outdoor spaces and also dynamic scene participants such as pedestrians and vehicles. Secondly, our UBotNet, based on U-Net and Bottleneck transformer, trained on the OmniHorizon dataset demonstrated significantly improved and scene-consistent normal estimation against the vanilla U-Net architecture. Furthermore, we presented UBotNet Lite, a smaller version of the network that retains respectable depth and normal accuracy while having only 30% of the network parameters. We outlined the benefits of pre-training network on OmniHorizon and fine-tuning it on Fukuoka dataset. Finally, we demonstrated the application of our model trained on OmniHorizon for estimating the depth and normals of real-world outdoor omnidirectional images in-the-wild.

References

[1] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3d: A holistic benchmark and a solid baseline for 360° depth estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3722–3732, 2021. 1, 3, 4, 5

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

[3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13137–13146, 2021. 5

[4] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. In *British Machine Vision Conference (BMVC)*, 2022. 3

[5] Jay Bhanushali, Achsah Steffi John, and Manivannan Muniyadi. Attention score: Objective measure of attentiveness in immersive omnidirectional videos. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 163–170, 2022. 2

[6] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 1

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 1, 3

[8] Hong Xiang Chen, Kunhong Li, Zhiheng Fu, Mengyi Liu, Zonghao Chen, and Yulan Guo. Distortion-aware monocular depth estimation for omnidirectional images. *IEEE Signal Processing Letters*, 28:334–338, 2021. 1

[9] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021. 5

[10] Weifeng Chen, Donglai Xiang, and Jia Deng. Surface normals in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1557–1566, 2017. 1, 5

[11] Xinjing Cheng, Peng Wang, Yanqi Zhou, Chenye Guan, and Ruigang Yang. Omnidirectional depth extension networks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 589–595. IEEE, 2020. 3

[12] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

[13] 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. 2, 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 4

[15] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 1

[16] Marc Eder, Pierre Moulon, and Li Guan. Pano popups: Indoor 3d reconstruction with a plane-aware network. In *2019 International Conference on 3D Vision (3DV)*, pages 76–84. IEEE, 2019. 3

[17] Marc Eder, True Price, Thanh Vu, Akash Bapat, and Jan-Michael Frahm. Mapped convolutions. *arXiv preprint arXiv:1906.11096*, 2019. 3

[18] Epic Games. Unreal engine. 3

[19] Cheng Sun et al. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, 2021. 1

[20] Giovanni Pintore et al. SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *CVPR*, 2021. 1, 5, 6

[21] Brandon Yushan Feng, Wangjue Yao, Zheyuan Liu, and Amitabh Varshney. Deep depth estimation on 360 images with a double quaternion loss. In *2020 International Conference on 3D Vision (3DV)*, pages 524–533. IEEE, 2020. 3

[22] Epic Games. Good sky plugin, 2022. https://www.unrealengine.com/marketplace/en-US/product/good-sky. 4

[23] Epic Games. Metahumans, 2022. https://www.unrealengine.com/en-US/metahuman. 4

[24] Ravi Garg, Vijay Kumar B. G, and Ian D. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. *CoRR*, abs/1603.04992, 2016. 1

[25] Andrew Svanberg Hamilton. Rural australia, 2022. https://www.unrealengine.com/marketplace/en-US/product/rural-australia. 4

[26] Haruya Ishikawa. Pyequilib: Processing equirectangular images with python, 2021. http://github.com/haruishi43/equilib. 5

[27] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526, 2021. 1, 3, 5

[28] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2020. 3

[29] Antonios Karakottas, Nikolaos Zioulis, Stamatis Samaras, Dimitrios Ataloglou, Vasileios Gkitsas, Dimitrios Zarpalas, and Petros Daras. 360 surface regression with a hyper-sphere

loss. In *2019 International Conference on 3D Vision (3DV)*, pages 258–268. IEEE, 2019. 1

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[31] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 5

[32] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2801–2810, 2022. 1

[33] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J.Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? 2017. 3

[34] Oscar Martinez Mozos, Kazuto Nakashima, Hojung Jung, Yumi Iwashita, and Ryo Kurazume. Fukuoka datasets for place categorization. *The International Journal of Robotics Research*, 38(5):507–517, 2019. 1, 2, 3, 8

[35] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018. 5

[36] Polyhaven. Hdri polyhaven, 2022. https://polyhaven.com/hdris. 8

[37] PurePolygons. Downtown west modular pack, 2022. https://www.unrealengine.com/marketplace/en-US/product/6bb93c7515e148a1a0a0ec263db67d5b. 4

[38] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 1

[39] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. 3

[40] Renderpeople. Scanned 3d people pack, 2022. https://www.unrealengine.com/marketplace/en-US/product/9c3fab270dfe468a9a920da0c10fa2ad. 4

[41] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360MonoDepth: High-resolution 360 monocular depth estimation. In *CVPR*, 2022. 1, 2, 3

[42] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360deg monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3762–3772, June 2022. 2

[43] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Winter Conference on Applications of Computer Vision*, 2020. 5

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 2, 4

[45] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. 1

[46] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360 depth estimation. In *European Conference on Computer Vision*, pages 195–211. Springer, 2022. 5, 6

[47] SilverTm. City park environment collection, 2022. https://www.unrealengine.com/marketplace/en-US/product/city-park-environment-collection. 4

[48] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[49] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16519–16529, June 2021. 2, 4, 5

[50] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[51] 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, 5, 6

[52] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 707–722, 2018. 3

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[54] Fu-En Wang, Hou-Ning Hu, Hsien-Tzu Cheng, Juan-Ting Lin, Shang-Ta Yang, Meng-Li Shih, Hung-Kuo Chu, and Min Sun. Self-supervised learning of depth and camera motion

from 360° videos. In *Asian Conference on Computer Vision*, pages 53–68. Springer, 2018. 2, 3

[55] 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 *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 5, 6

[56] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020. 3

[57] Changhee Won, Hochang Seok, Zhaopeng Cui, Marc Pollefeys, and Jongwoo Lim. Omnislam: Omnidirectional localization and dense mapping for wide-baseline multi-camera systems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 559–566, 2020. 2

[58] Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3

[59] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020. 1

[60] Wei Zeng, Sezer Karaoglu, and Theo Gevers. Joint 3d layout and depth prediction from a single indoor panorama image. In *European Conference on Computer Vision*, pages 666–682. Springer, 2020. 3

[61] Richard Zhang. Making convolutional networks shift-invariant again. In *ICML*, 2019. 5

[62] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pages 519–535. Springer, 2020. 1, 3

[63] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018. 1, 3, 5, 6