PanoDepth: A Two-Stage Approach for Monocular Omnidirectional Depth Estimation

Yuyan Li
University of Missouri
yl235@umsystem.edu

Zhixin Yan
BOSCH Research China
zhixin.yan2@cn.bosch.com

Ye Duan
University of Missouri
duanye@umsystem.edu

Liu Ren
BOSCH Research North America
liu.ren@us.bosch.com

Abstract

Omnidirectional 3D information is essential for a wide range of applications such as Virtual Reality, Autonomous Driving, Robotics, etc. In this paper, we propose a novel, model-agnostic, two-stage pipeline for omnidirectional monocular depth estimation. Our proposed framework PanoDepth takes one 360 image as input, produces one or more synthesized views in the first stage, and feeds the original image and the synthesized images into the subsequent stereo matching stage. In the second stage, we propose a differentiable Spherical Warping Layer to handle omnidirectional stereo geometry efficiently and effectively. By utilizing the explicit stereo-based geometric constraints in the stereo matching stage, PanoDepth can generate dense high-quality depth. We conducted extensive experiments and ablation studies to evaluate PanoDepth with both the full pipeline as well as the individual modules in each stage. Our results show that PanoDepth outperforms the state-of-the-art approaches by a large margin for 360 monocular depth estimation. Our code is available at https://github.com/yuyanli0831/PanoDepth_3dv.

1. Introduction

Omnidirectional 3D information is essential for a wide range of applications (e.g. Virtual Reality [2], augmented reality [4], autonomous driving [1], and robotics [44]). Quick and reliable omnidirectional data acquisition can facilitate many use cases, such as user interaction with the digital environment, robot navigation, and object detection for autonomous vehicles. Another relevant application is remote working/shopping/education [18], which has become ubiquitous due to the pandemic. To obtain high-quality omnidirectional 3D information, devices such as omnidirectional LiDARs are widely used in autonomous driving and indoor 3D scans. However, LiDARs are either very expensive or can only produce sparse 3D scans. Compared with LiDARs, cameras are much cheaper and already frequently used for capturing the visual appearance of the scenes. The cost can be significantly reduced if high-quality omnidirectional 3D can be generated directly from camera images.

Deep learning techniques, coupled with a growing accessibility of large-scale datasets, have largely improved the performance of many computer vision tasks including depth estimation [63]. Depth estimation often uses either a monocular input or a stereo pair. For the monocular methods, a common practice is to train a single network to map RGB pixel to real-value depth [17, 20, 28, 31], mostly by learning from various monocular cues such as shape, lighting, shading, object type, etc. Stereo matching methods [47, 34, 8], on the other hand, learn the disparity by matching image patches from stereo pairs and later convert disparity to depth. Despite the significant improvement in monocular estimation methods, there is still a large gap between monocular and stereo depth accuracy [48].

To apply stereo matching for monocular depth estimation, Luo et al. [39] proposed a two-stage pipeline that decomposes monocular depth estimation into two stages, view synthesis and stereo matching respectively. In their approach, a second view is first synthesized and fed together with the original view to the stereo matching stage to compute the disparity. Stereo matching can leverage more geometric constraints into the network training, thus reducing the demand for ground truth depth. More recent studies [52, 57] improved upon this idea and achieved promising performances. These two-stage methods [39, 52, 57], are mainly designed for perspective images. In 360 domain, most of the recent studies [67, 49, 15, 55, 14] still follow the same single-stage monocular estimation procedure with adaptation to 360 [1] geometry.

In this paper, we propose a novel, model-agnostic, two-

[1] We use the terms 360, omnidirectional, equirectangular, spherical interchangeably in this paper

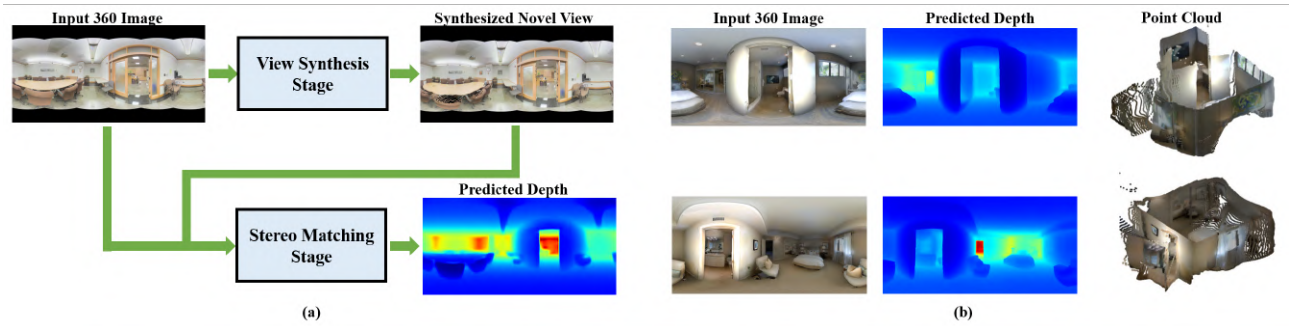

Figure 1. (a) Illustration of our *PanoDepth* framework. *PanoDepth* takes one 360 image as input to generate one or more novel views in the first view synthesis stage. The original and synthesized 360 images are then fed into the second multi-view stereo matching stage to predict final dense depth map. (b) Two examples (top and bottom row) of *PanoDepth* on 360D dataset [67] with 360 image (left) as input, and output depth (middle) and point cloud (right).

stage pipeline (see Figure 1) for solving the problem of 360 monocular depth estimation. Our proposed framework *PanoDepth* takes one equirectangular projection (ERP) image as input, produces one or more synthesized views in the first stage, and feeds the original image and the synthesized images to the subsequent stereo matching stage to predict the final depth map. In the stereo matching stage, we propose a novel differentiable Spherical Warping Layer to handle omnidirectional stereo geometry efficiently and effectively. We conducted extensive experiments and ablation studies to evaluate *PanoDepth* with both the full pipeline and the individual networks in each stage on several public benchmark datasets. Our results demonstrated that our model-agnostic approach *PanoDepth* outperforms the one-stage method by a large margin despite the combinations of coarse estimation and stereo matching networks. Moreover, by adjusting these networks, *PanoDepth* can be adapted to the target computation constraints and performance requirements.

Our contributions can be summarized as follows:

- We propose a novel, model-agnostic, two-stage framework *PanoDepth*, including view synthesis and stereo matching, to fully exploit the synthesized 360 views and spherical stereo constraints.
- *PanoDepth* outperforms the state-of-the-art monocular omnidirectional depth estimation approaches by a large margin.
- We propose a novel differentiable Spherical Warping Layer (SWL) which adapts regular stereo matching networks to 360 stereo geometry, and enables advanced features such as multi-view stereo and cascade mechanism for stereo performance boost.

2. Motivation

In this section, we explain the motivation of formulating the 360 monocular depth estimation problem as two separate stages, namely, a view synthesis stage based on coarse depth estimation, and a multi-view stereo matching stage for final depth output.

2.1. Why Two-Stage?

One main advantage of monocular depth estimation is its potential in dramatically reducing the hardware cost for 3D depth acquisition. Motivated by this, many studies have been proposed to solve this problem. The basic idea of supervised monocular estimation methods is to train a network that directly learns the mapping from the input RGB pixels to the real-value output depth in a single stage. For example, Laina et al. [35] proposed FCRN which uses ResNet-50 [26] as backbone, followed by multiple up-projection modules. Hu et al. [28] leveraged SENet-154 [29] as encoder together with multi-scale fusion module.

On the other hand, deep learning based stereo matching networks [8, 34, 23] utilize the stereo constraints to improve efficiency and the output depth quality. These methods simulate the traditional stereo matching process by learning and optimizing the matching cost across the input image pairs in a deterministic manner. Unlike monocular depth methods which directly map RGB into depth by considering all the monocular cues, stereo matching methods focus on estimating disparity by developing image patch correspondence [17]. Given the predefined baseline and 1D search space along the epipolar line for image patch matching [25], stereo matching produces more accurate depth maps in comparison with monocular methods in general [48].

A typical stereo matching network requires at least two images as input, which is not directly applicable for a monocular input setting. However, if one or more novel views can be synthesized with high quality, these additional views can be utilized to train a stereo matching network. Luo et al. [39] first proposed the two-stage pipeline for perspective images where a novel right view is synthesized at the first stage and paired with the original view to the second stereo stage. Later, two-stage approaches [52] mostly fol-

lowed this work to generate a coarse disparity/depth, and to synthesize novel views via image warping or Depth-Image-Based Rendering (DIBR). Taking the original and the synthesized views as input, the final depth generated from the later stereo matching stage of these approaches shows a significant improvement over the one-stage counterparts.

2.2. Can we successfully synthesize novel view 360 images?

Recent two-stage approaches [39, 52] have shown promising capabilities in improving depth quality on perspective images. However, it remains unclear whether the two-stage approach will be applicable for 360 images, as there are many fundamental differences between the perspective images and 360 images, such as camera projection model, image distortion, and field of view (FoV).

The difference in camera projection model, equirectangular projection (see Figure 3(a)) vs. perspective projection, can be resolved by integrating spherical geometry into the disparity calculation and cost volume fusion procedure. In this paper, we propose a novel Spherical Warping Layer specifically designed for spherical geometry as a solution (Section 4.2). Moreover, the distortion issue can be addressed by applying distortion-aware convolutions [15, 13, 50, 49, 12, 64, 9]. Hence, in this section we will mainly discuss the difference in FoV settings. Comparing with perspective images, 360 images have much larger FoVs (360° horizontally, 180° vertically). A 360 image encodes almost every piece of visual information of the scene except occluded areas, while perspective images suffer from information loss near the image boundaries in addition to occlusions. This could be a great advantage for novel view synthesis of 360 images.

To validate this observation, we conducted various experiments regarding the correlations between image FoV, baseline and synthesized view quality (more details in the Appendix). Our experiment confirms that i) with greater FoV, the synthesized views are less sensitive to large baselines, and ii) synthesized 360 images have the least error and artifacts. According to Gallup et al. [21], depth error of stereo matching comes from both the disparity error (proportional) and the baseline (inversely proportional). Thus, with higher quality synthesized novel views and larger baselines, we expect less error in the final depth output from stereo matching. This also indicates that the two-stage pipeline is well-suited for 360 monocular depth estimation.

3. Related Work

3.1. Monocular Depth Estimation

Monocular depth estimation [46] has seen significant improvements [35, 28] since the first adoption of deep learning by Eigen et al. [17]. To further improve the performance, researchers explored many strategies such as multi-task learning with normal estimation [45] and semantic segmentation [16, 32] along with depth estimation, incorporating CRF [6, 40], integrating attention modules [31, 36], utilizing planar constraints [37, 31], conducting unsupervised learning using constrainsts such as left-right consistency [22, 5, 42], as well as two-stage approaches where stereo constraints are leveraged [39, 52, 57].

As 360 cameras become more affordable, researchers start to explore the possibility to use 360 images for depth estimation [15, 51, 11, 67, 56, 66, 62]. To advance the performance of 360 monocular depth estimation, Eder et al. [14] proposed joint training of surface normal, boundary, and depth. Zeng et al. [62] trained a network which combines 3D layout and depth. Jin et al. [33] took advantage of the correlation between depth map and geometric structure of 360 indoor images. Cheng et al. [11] proposed a low-cost sensing system which combines an omnidirectional camera with a calibrated projective depth camera. The 360 image and the limited FoV depth are used together as input to a CNN. Meanwhile, distortion-aware convolution filters [51, 64, 15, 9] are designed to handle spherical geometric distortion.

3.2. Multi-View Stereo Matching

Besides monocular depth estimation, Multi-View Stereo (MVS) is another group of methods for predicting depth. Given a set of images with known camera poses, MVS approaches [30, 61] can produce highly accurate depth estimates with multi-view geometric constraints. One example is MVSNet [61], in which the variance-based cost volume is presented to fuse multiple features maps from source images into one unified cost volume. Stereo matching can be treated as a special case of MVS. Conventionally, stereo-based depth estimation methods [47, 27] relied on matching pixels across stereo images. Many recent stereo matching approaches [34, 8] leveraged CNNs for feature extraction, cost matching, and aggregation. For example, PSMNet [8] incorporated spatial pyramid pooling (SPP) module and multi-scale 3D hourglass modules to further boost the performance. To improve the efficiency and accelerate training on high-resolution images, Gu et al. [23] presented a cascade cost volume design to gradually retrieve finer hypothesis plane ranging over multiple steps.

In 360 stereo domain, SweepNet [60] and OmniMVS [59] estimated depth from multiple wide-baseline fisheye cameras. Another recent work is 360SD-Net [56] which predicted disparity/depth from a pair of ERP images that are taken by a top-bottom camera pair. They [56] incorporated polar angles to solve distortions and proposed learnable shifting filters to adjust the step size in disparity cost volume construction. However, the learnable shifting filters create extra overhead during training. In this paper, we

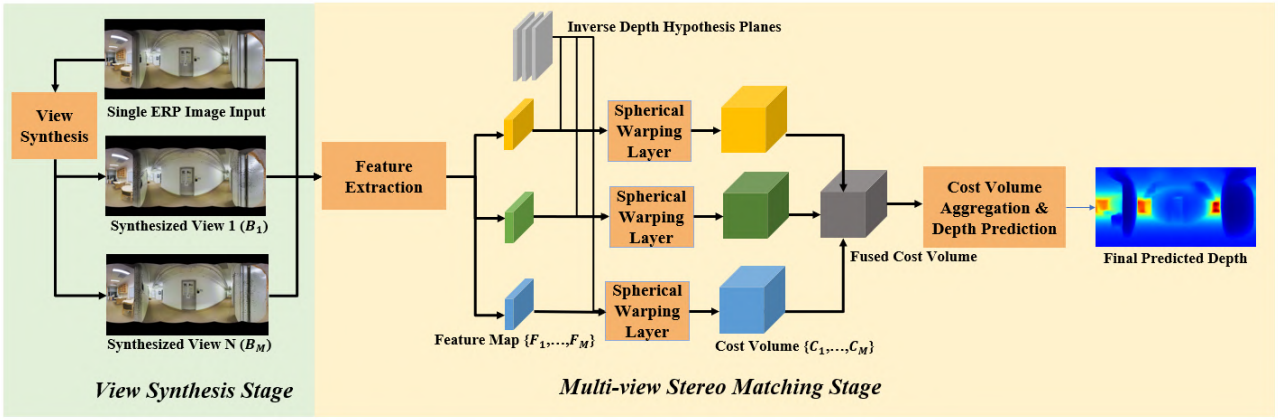

Figure 2. Illustration of our *PanoDepth* framework which consists of a view synthesis stage and a multi-view stereo matching stage. In the view synthesis stage, a total of M synthesized views are generated. In the multi-view stereo matching stage, the synthesized 360 views together with the original input view, are sent to the multi-view stereo matching network to produce the final depth estimation. To better adapt to 360 stereo geometry, we directly sample hypothesis plane on the inverse depth, and use a Spherical Warping Layer (SWL) to warp reference views to the target view (Section 4.2).

propose a closed-form solution, Spherical Warping Layer (SWL), that does not require additional training overhead. Our experiments (Table 2) show that SWL can significantly improve the performance of 360 stereo matching.

4. Approach

We propose an end-to-end framework, *PanoDepth*, that takes a single ERP image as input and produces a high-quality omnidirectional depth map. *PanoDepth* consists of two stages: i) a view synthesis stage that conducts coarse depth estimation followed by a differentiable DIBR module for novel view synthesis, and ii) a stereo matching stage with a customized Spherical Warping Layer for efficient and high-quality 360 depth estimation. A full framework of *PanoDepth* is illustrated in Figure 1.

4.1. View Synthesis Stage

To synthesize high-quality novel views, the coarse depth map is usually estimated first followed by a Depth-Image-Based Rendering (DIBR) module [39, 65]. Based on our empirical observations (see the Appendix for details), such a procedure also works well for 360 novel view synthesis with different configurations of coarse depth estimation networks. Considering both performance and computation cost, in this paper we suggest using a lightweight network: CoordNet [66] for doing the task. CoordNet utilizes coordinate convolution [38] to enforce 360 awareness. We append an atrous spatial pyramid pooling module (ASPP) [10] to the end of the encoder to better aggregate multi-scale context information. Note that *PanoDepth* is model-agnostic, thus any depth estimation network can be used here to fulfill specific requirement. The estimated coarse depth map and the original ERP image are then used to render multiple synthesized views of predefined baselines via a differentiable DIBR operation [53]. In this paper, we choose to use vertical baselines instead of horizontal ones. The analysis of this choice can be found in the Appendix.

4.2. Stereo Matching Stage

The second stage of our *PanoDepth* framework is stereo matching. Again, as *PanoDepth* is model-agnostic, any stereo matching network can be plugged in here. Experimental results that show the performance of different stereo matching network settings is discussed in Section 5.4.

The stereo matching network we used in this paper follows a similar pipeline as PSMNet [8], with several key modifications. The network consists of five main modules: feature extraction, spherical warping layer, cost volume construction, cost aggregation, and depth prediction. Comparing with the original PSMNet [8], our unique contribution is the Spherical Warping Layer (SWL) which is specifically designed for the 360 stereo geometry.

Feature Extraction After the view synthesis stage, the input image along with all the M synthesized novel views will be passed to a weight-sharing neural network to extract features. We use the same layer setting and keep the SPP module as the original PSMNet [8].

Spherical Warping Layer (SWL) The extracted feature maps of all views are then used to build a cost volume at multiple depth hypothesis planes for cost matching. An essential step of cost volume construction is to determine the coordinate mapping, which is reflected as disparity, that warps reference view to the target view. Unlike perspective images where the disparity is proportional to the inverse depth [43, 8, 54], disparity of 360 stereo pairs is related to both inverse depth and spherical latitudes. Our SWL per-

forms direct depth sampling instead of disparity sampling which is commonly used in perspective image pair stereo matching. Compared with the learnable shifting filters proposed in [56], our SWL makes the disparity computation adjustable to the pixel latitudinal value, without introducing additional computation overhead in training. Moreover, SWL directly samples on the absolute depth domain, thereby enabling horizontal 360 stereo (which has both horizontal and vertical disparity, more details in the Appendix), the adaptation of 360 multi-view, and the usage of cascade mechanism.

Figure 3 shows an illustration of the Spherical Warping Layer. We first sample the inverse depth to cover the whole depth range:

$$\frac{1}{d_j} = \frac{1}{d_{max}} + (\frac{1}{d_{min}} - \frac{1}{d_{max}})\frac{v \times j}{D-1}, j = 0, 1, ..., D-1 \tag{1}$$

where D is the total number of hypothesis planes, d_j is the j^{th} depth plane, d_{min} and d_{max} are the minimum and maximum of the depth image, v is the plane interval. The Spherical Warping Layer then transforms depth hypothesis d_j to displacement in spherical domain C_j, to map pixels from the reference synthesized view to the target view. The displacement C_j is defined as:

$$C_{x,j} = 0,\ C_{y,j} = \frac{cos(\theta) \times b}{d_j} \times \frac{H_f}{\pi} \tag{2}$$

where θ refers to the pixel-wise latitudinal values, b represents the baseline, and H_f is the height of the feature map.

Cost Volume Construction The SWL transforms reference view feature maps into the target view domain at the individual hypothesis plane, and then a total of $M+1$ feature volumes are generated. The variance-based cost volume formation method from MVSNet [61] is used for the fusion of these feature volumes into a compact one. Moreover, we adopt a cascade design from Gu et al. [23] to further improve the final depth quality. Specifically, at level $l (l > 1)$, d_{min} and d_{max} is recalculated based on the prediction of level $l-1$, then the new depth range and the new number of planes D_l is used to determine the new intervals. Depth hypothesis for level l is then updated using equ (1). The corresponding displacements are calculated via the same spherical coordinate mapping procedure.

Cost Aggregation and Depth Prediction After the construction of the cost volume, multi-scale 3D CNN is used to aggregate different levels of spatial context information through the hourglass-shape encoding and decoding network. It has been shown this kind of cost aggregation module helps to regularize noises in ambiguous regions caused by occlusions, textureless surfaces, and to improve final prediction quality [34, 8, 24]. Finally, we regress the depth

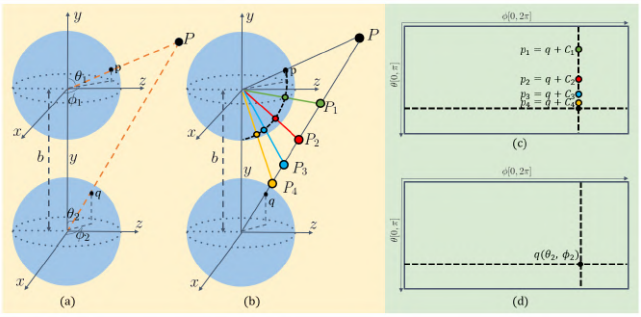

Figure 3. Visualization of our spherical warping method. (a) Vertical 360 stereo model. b is the baseline displacement of two cameras. P is a real-world point. The projection of P on two camera space is represented as $p(\phi_1, \theta_1)$ and $q(\phi_2, \theta_2)$. (b) Spherical epipolar geometry. P_i is the points sampled at different depths. (c) Projection of sampled inverse depth on the ERP image. p_i is the projection of P on the top view at sampled points P_i. C_i is the vertical disparity, it equals to C_y in Equ (2). (d) Projection on the reference image. q is the projection of P at bottom view.

value at each level l:

$$\frac{1}{d_{pred,l}} = \frac{1}{d_{min,l}} + (\frac{1}{d_{min,l}} - \frac{1}{d_{max,l}})\frac{k_l}{D_l - 1} \tag{3}$$

$$k_l = \sum_{j=0}^{D_l - 1} \sigma(p_j) \times (v_l \times j) \tag{4}$$

where k_l is the sum of each plane level weighted by its normalized probability, $\sigma(\cdot)$ represents softmax function, p_j denotes the probability of j^{th} plane value, v_l is the interval for level l.

4.3. Loss Function

PanoDepth is trained in an end-to-end fashion, supervision is applied on both stages. The final loss function is defined as follows,

$$\mathcal{L}_{total} = \omega_1 \mathcal{L}_{coarse} + \omega_2 \mathcal{L}_{stereo} \tag{5}$$

where ω_1 and ω_2 are the weights of coarse depth estimation loss and stereo matching loss respectively. For the optimization on the first stage coarse estimation, we use inverse Huber (berHu) loss as proposed in [35]:

$$\mathcal{L}_{coarse} = \frac{1}{\Omega} \sum_{i \in \Omega} \mathcal{L}_{berHu}(d_i,\ d_i^*) \tag{6}$$

where Ω is a binary mask that is used to mask out missing regions (pixels that have depth values smaller than d_{min} or greater than d_{max}), d_i and d_i^* are the ground truth and the predicted depth value of a valid pixel i respectively. For stereo matching, we calculate berHu loss [35] on all outputs from each level l and then compute the weighted sum-

mation. The stereo matching loss is defined as:

$$\mathcal{L}_{stereo} = \frac{1}{\Omega}\sum_{i\in\Omega}\sum_{l=1}^{N}\lambda_l \mathcal{L}_{berHu}(d_i,\ \ d_i^*) \tag{7}$$

where λ_l is the level l stereo loss weight.

5. Experiments

5.1. Datasets

We train and evaluate our network on three panorama RGBD benchmark datasets including, Stanford2D3D [3], 360D [67] and the omnidirectional stereo dataset [66].

Stanford2D3D Stanford2D3D [3] dataset consists of 1413 real-world panorama images from six large-scale indoor areas. We follow the official train-test split which uses the fifth area for testing, and other areas for training. We resize the images to 256×512 to reduce computation time.

360D 360D [67] is a RGBD panorama benchmark provided by Zioulis et al. [67]. It is composed of two other synthetic datasets (SunCG and SceneNet), and two real-world datasets (Stanford2D3D and Matterport3D). There are 35,977 panorama RGBD images in the 360D that are rendered from the aforementioned four datasets. We again follow the default train-test splits.

Omnidirectional Stereo Dataset. The omnidirectional stereo dataset [66] consists of 7964 stereo pairs of panorama RGBD images rendered from two real-world datasets, Matterport3D [7] and Stanford2D3D [3]. We use the train-test split that removes 3 complete buildings from Matterport3D [7] and 1 complete area from Stanford2D3D [3] for test. Each set of data consists of left-down, right, and up view 360 RGBD images in a triangular fashion with size 256×512. We only use images from the left-down view in our single view depth estimation experiments. Up-down stereo pairs are only used for the ablation study of the stereo matching network.

To investigate the impact of baselines and FoV on the quality of view synthesis, we create a new dataset that is rendered from the mesh of Stanford2D3D [3]. More details of this new dataset as well as the experiments with various baseline configurations are included in the Appendix.

5.2. Implementation Details and Metrics

For parameter settings, we use a default of $N = 2$ levels, with $D_1 = 48$ and $D_2 = 24$ hypothesis planes respectively. The minimum and maximum depth d_{min} and d_{max} for the first level is set to be $0.2m$ and $8m$. We use a default of $M = 3$ synthesized views rendered at vertical baseline placements $-0.24m, +0.24m, +0.4m$. The loss weights ω_1 and ω_2 are set to 1 and 0.02. We train our framework from scratch using Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with a batch size of 8. Initial learning rates for the first and the second stage are set to 0.0002 and 0.0005. We separately train the coarse network for 10 epochs and then train the entire framework end-to-end for 200 epochs. Both of the learning rates decay by a factor of 0.5 every 30 epochs. Performances are evaluated based on commonly used depth quality measures [17]: absolute relative error (Abs Rel), square relative error (Sq Rel), linear root mean square error(RMSE) and its natural log scale (RMSE log) and inlier ratios ($\delta_i < 1.25^i, i \in \{1, 2, 3\}$).

5.3. Overall Performance Comparison with the State-of-the-art Algorithms

Table 1 lists quantitative comparison between *PanoDepth* and other state-of-the-art omnidirectional monocular depth estimation methods [35, 67, 55, 11] on both Stanford2D3D [3] and 360D [67] datasets. As shown in Table 1, our method is able to reduce Abs Rel error by 19.60% on Stanford2D3D and 25.85% on 360D compare to the current leading 360 monocular depth estimation approach BiFuse [55]. Note that BiFuse [55] uses a network architecture with more than 200M parameters and has a large computation complexity for sharing information between CubeMap [41] and ERP formats. The framework with distortion-aware module proposed in [9] outperforms ours but it has more than 60M parameters. Our framework has only around 16M parameters with a smaller computation overhead. Comparing the performance with ODE-CNN [11] on 360D, our approach achieves comparable results while ODE-CNN requires additional depth sensor input. Figure 4 shows the qualitative comparison with the state-of-the-art approaches. As we can see, our method generates high-quality depth with a detailed surface, sharp edges, and precise range.

5.4. Ablation Studies

Spherical Warping Layer In order to evaluate the performance of the *PanoDepth* stereo matching module with the novel Spherical Warping Layer (SWL), we compare it to the state-of-the-art stereo matching approaches, PSMNet [8] and 360-SD Net [56], with ground truth up-down 360 stereo pair as input on the Omnidirectional Stereo Dataset [66]. In the experiment, we use the officially released code of both approaches, and convert the output disparity into depth for evaluation. We set the number of depth hypothesis planes to 64 to ensure fair and consistent comparison. We can see from Table 2 that our proposed stereo matching method with SWL outperforms PSMNet [8] and 360-SD Net [56], even with one-level setting. Moreover, by adding SWL, our one-level setting outperforms the one without SWL (identical to PSMNet[8]) by 47% in terms of Abs Rel. with the same 64 sampling planes. Qualitative illustrations of the effectiveness of SWL is shown in Figure 5.

Model-agnostic Evaluations In Table 3, we further test the performance of the full *PanoDepth* pipeline given

Datasets	Methods	Abs Rel↓	RMSE↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
Stanford2D3D [3]	FCRN [35]	0.1837	0.5774	0.7230	0.9207	0.9731
	RectNet [67]	0.1409	0.4568	0.8326	0.9518	0.9822
	BiFuse with fusion [55]	0.1209	0.4142	0.8660	0.9580	0.9860
	Joint wth layout and semantics [62]	0.0680	0.2640	0.9540	0.9920	0.9980
	PanoDepth(Ours)	0.0972	0.3747	0.9001	0.9701	0.9900
360D [67]	FCRN [35]	0.0699	0.2833	0.9532	0.9905	0.9966
	RectNet [67]	0.0702	0.2911	0.9574	0.9933	0.9979
	Mapped Convolution [15]	0.0965	0.2966	0.9068	0.9854	0.9967
	Distortion-aware [9]	0.0406	0.1769	0.9865	0.9966	0.9987
	BiFuse with fusion [55]	0.0615	0.2440	0.9699	0.9927	0.9969
	ODE-CNN [11]	0.0467	0.1728	0.9814	0.9967	0.9989
	PanoDepth(Ours)	0.0456	0.1955	0.9830	0.9957	0.9984

Table 1. A quantitative comparison with the state-of-the-art approaches on Stanford2D3D [3] dataset and 360D [67] dataset (↓ represents lower the better, ↑ represents higher the better). We report the results based on the original papers [67, 55, 11] using the same evaluation metrics. Note that ODE-CNN [11] requires additional depth sensor input besides the 360 image used by other methods listed in the table. Additional supervision signals including layout and semantics are used in [62]. For our *PanoDepth*, we use the default stereo network setting with three synthesized views and a two-level cascade design.

Methods	Abs Rel↓	Sq Rel↓	RMSE↓	RMSElog↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
(1) PSMNet [8], sample on disparity, $D = 64$	0.0433	0.0252	0.2541	0.1340	0.9722	0.9833	0.9900
(2) 360SD-Net [56], sample on disparity, $D = 64$	0.0387	0.0198	0.2286	0.0955	0.9776	0.9900	0.9940
(3) PanoDepth (ours), one-level, $D = 32$	0.0253	0.0222	0.2268	0.0686	0.9756	0.9874	0.9976
(4) PanoDepth (ours), one-level, $D = 64$	0.0229	0.0087	0.1731	0.0606	0.9900	0.9969	0.9987
(5) PanoDepth (ours), two-level, $D_1 = 48, D_2 = 24$	**0.0178**	**0.0064**	**0.1415**	**0.0519**	**0.9928**	**0.9976**	**0.9990**

Table 2. A quantitative comparison between the PanoDepth stereo matching and existing stereo matching networks on the Omnidirectional stereo dataset [66] where up-down stereo pairs are used as input and output the depth of bottom view. Our proposed stereo matching module (3,4,5) outperforms both (1) PSMNet [8] and (2) 360SD-Net [56]. The two cascade level setting achieves the best performance.

	1st Stage	2nd Stage w/ 1 synthesize view	#params	Abs Rel↓	Sq Rel↓	RMSE↓	RMSElog↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
(1)	CoordNet	N/A	6.1M	0.1264	0.0888	0.4456	0.2084	0.8533	0.9588	0.9813
(2)	RectNet	N/A	10.8M	0.1409	0.0859	0.4568	0.2124	0.8326	0.9518	0.9822
(3)	CoordNet	PSMNet [8], D=32	13.0M	0.1206	0.0833	0.4293	0.2150	0.8671	0.9548	0.9790
(4)	CoordNet	w/ SWL, one-level, D=32	13.0M	0.1132	0.0686	0.4077	0.1869	0.8757	0.9652	0.9863
(5)	RectNet	w/ SWL, one-level, D=32	17.6M	0.1192	0.0775	0.4202	0.1960	0.8655	0.9607	0.9846
(6)	CoordNet	w/ SWL, two-level, D_1=32, D_2=16	16.6M	**0.1040**	**0.0645**	**0.3918**	**0.1827**	**0.8865**	**0.9676**	**0.9875**
(7)	RectNet	w/ SWL, two-level, D_1=32, D_2=16	21.3M	0.1138	0.0761	0.4274	0.1961	0.8711	0.9577	0.9837

Table 3. An ablation study of the impact of various combinations of coarse estimation network and stereo matching network on the final performance. The experiments are trained on Stanford2D3D [3]. We use two types of coarse estimation networks, (1) CoordNet, and (2) RectNet [67]. We can see that even with one synthesize view, our proposed two-stage *PanoDepth* pipeline (3,4,5,6,7) is able to outperform the one-stage-only methods (1,2). Adding Spherical Warping Layer (SWL) (4,5,6,7) and two cascade levels (6,7) further improves the performance. The experimental results indicate that our two-stage pipeline is model-agnostic under various network settings.

different variations of stereo matching networks on Stanford2D3D dataset [3]. Comparing with the single-stage coarse depth estimation, all two-stage configurations show better performances. By adding a light-weight one-level stereo matching network with 32 depth plane in the second stage, *PanoDepth* can already reach comparable performance to BiFuse [55]. The performance can be further improved by introducing SWL, adding more cascade levels, and using more sophisticated coarse depth estimation.

In addition, by comparing the performance of two-stage approaches with two different backbones, CoordNet and RectNet, we can also postulate that coarse depth, which has an impact on the synthetic image quality, is positively correlated with the final depth prediction.

More ablation studies regarding the number of synthesized views, the number of hypothesis depth planes, and the comparison between one-stage alternatives (e.g., multi-tasking and adding depth refinement) and our two-stage method can be found in the Appendix.

6. Conclusion and Future Work

In this paper, we demonstrate a technique that leverages view synthesis and stereo constraints to advance monocular depth estimation performance that can be applied on 360 images. We propose a novel model agnostic two-stage framework *PanoDepth* for generating dense high-

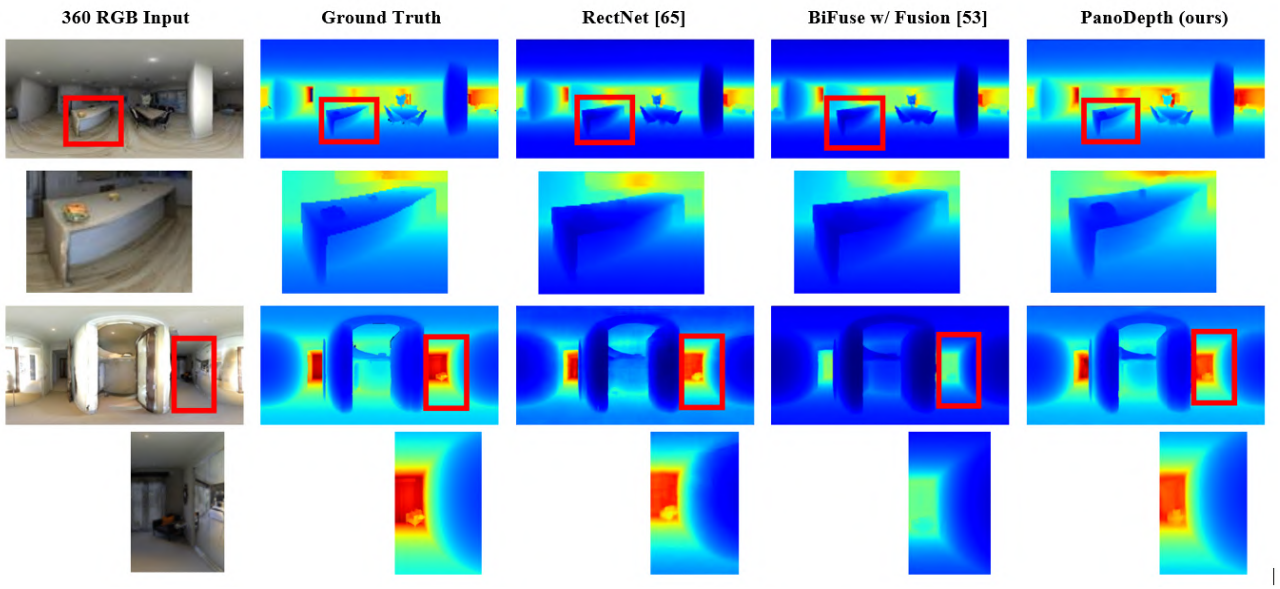

Figure 4. A qualitative comparison between RectNet [67] (3rd column), BiFuse [55] (4th column), and our method (5th column) on 360D [67]. We highlight and zoom in some areas that distinguish the performance of three methods. We can see that our *PanoDepth* is able to produce sharp edges, predict depth range accurately, and recover surface detail.

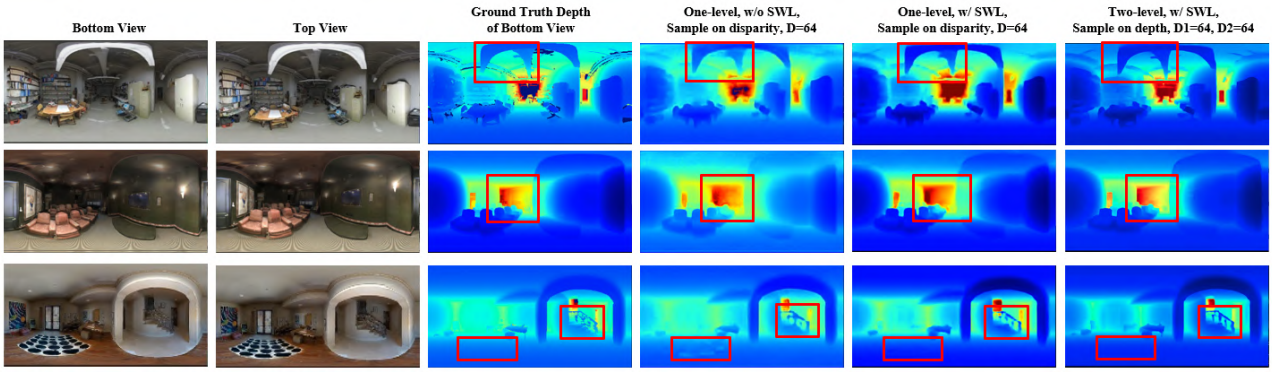

Figure 5. A qualitative comparison to show the effectiveness of SWL in the PanoDepth stereo matching module. We compare between one-level stereo matching method without SWL(4th column), one-level with SWL (5th column), and two-level with SWL (last column). The experiments are trained on Omnidirectional Stereo Dataset [66]. Our stereo matching module with SWL recovers clear details and shows fewer artifacts than the one without SWL (see highlighted areas).

quality depth from a monocular 360 input. Extensive experiments show that *PanoDepth* outperforms state-of-the-art approaches by a large margin. Our stereo matching sub-network in the later stage adapts to the 360 geometry and achieves top-ranking performance in 360 stereo matching. We believe the good performance of PanoDepth could draw more interests from both the industry and academia to 360 images for its still under-explored capability in tasks such as depth estimation. We hope our work can motivate more research and applications in 360 images. There are several research venues we would like to further explore in the future, such as alternative view synthesis methods like [19, 58], and 360 depth estimation in outdoor scenarios for applications like autonomous driving.

Acknowledgments

The research of Yuyan Li and Ye Duan were partially supported by the National Science Foundation under award CNS-2018850, National Institute of Health under awards NIBIB-R03-EB028427 and NIBIB-R01-EB02943, and U.S. Army Research Laboratory W911NF2120275. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U. S. Government or agency thereof.

References

[1] Naveen Appiah and Nitin Bandaru. Obstacle detection using stereo vision for self-driving cars. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 926–932, 2011.

[2] Lemonia Argyriou, Daphne Economou, and Vassiliki Bouki. Design methodology for 360 immersive video applications: the case study of a cultural heritage virtual tour. *Personal and Ubiquitous Computing*, pages 1–17, 2020.

[3] 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.

[4] Matthias Berning, Takuro Yonezawa, Till Riedel, Jin Nakazawa, Michael Beigl, and Hide Tokuda. parnorama: 360 degree interactive video for augmented reality prototyping. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 1471–1474, 2013.

[5] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems*, pages 35–45, 2019.

[6] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2017.

[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.

[8] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[9] 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.

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[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.

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

[13] Marc Eder and Jan-Michael Frahm. Convolutions on spherical images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–5, 2019.

[14] 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.

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

[16] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[17] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[18] Michael S. Feurstein. Towards an integration of 360-degree video in higher education. In *DeLFI Workshops*, 2018.

[19] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019.

[20] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[21] David Gallup, Jan-Michael Frahm, Philippos Mordohai, and Marc Pollefeys. Variable baseline/resolution stereo. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[22] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[23] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. *arXiv preprint arXiv:1912.06378*, 2019.

[24] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019.

[25] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

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

[27] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814. IEEE, 2005.

[28] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019.

[29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018.

[30] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[31] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkila. Guiding monocular depth estimation using depth-attention volume. *arXiv preprint arXiv:2004.02760*, 2020.

[32] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European conference on computer vision (ECCV)*, pages 53–69, 2018.

[33] 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 (CVPR)*, June 2020.

[34] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[35] 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.

[36] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust depth prediction. In *Asian Conference on Computer Vision*, pages 663–678. Springer, 2018.

[37] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2588, 2018.

[38] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems*, pages 9605–9616, 2018.

[39] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018.

[40] Faisal Mahmood and Nicholas J Durr. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. *Medical image analysis*, 48:230–243, 2018.

[41] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic. Salnet360: Saliency maps for omni-directional images with cnn. *Signal Processing: Image Communication*, 69:26–34, 2018.

[42] Greire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P Breckon. Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 789–807, 2018.

[43] Marc Pollefeys, Reinhard Koch, and Luc Van Gool. A simple and efficient rectification method for general motion. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 496–501. IEEE, 1999.

[44] Gyula Pudics, Miklós Zsolt Szabó-Resch, and Zoltán Vámossy. Safe robot navigation using an omnidirectional camera. In *2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI)*, pages 227–231. IEEE, 2015.

[45] 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.

[46] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006.

[47] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[48] Nikolai Smolyanskiy, Alexey Kamenev, and Stan Birchfield. On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1007–1015, 2018.

[49] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. In *Advances in Neural Information Processing Systems*, pages 529–539, 2017.

[50] Yu-Chuan Su and Kristen Grauman. Kernel transformer networks for compact spherical convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9442–9451, 2019.

[51] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[52] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9799–9809, 2019.

[53] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In

Proceedings of the European Conference on Computer Vision (ECCV), pages 302–317, 2018.

[54] Julien Valentin, Adarsh Kowdle, Jonathan T Barron, Neal Wadhwa, Max Dzitsiuk, Michael Schoenberg, Vivek Verma, Ambrus Csaszar, Eric Turner, Ivan Dryanovski, et al. Depth from motion for smartphone ar. *ACM Transactions on Graphics (ToG)*, 37(6):1–19, 2018.

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

[56] Ning-Hsu Wang, Bolivar Solarte, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 360sd-net: 360 stereo depth estimation with learnable cost volume. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 582–588. IEEE, 2020.

[57] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. *arXiv preprint arXiv:2008.01484*, 2020.

[58] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020.

[59] Changhee Won, Jongbin Ryu, and Jongwoo Lim. Omnimvs: End-to-end learning for omnidirectional stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8987–8996, 2019.

[60] Changhee Won, Jongbin Ryu, and Jongwoo Lim. Sweepnet: Wide-baseline omnidirectional depth estimation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6073–6079, 2019.

[61] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[62] 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.

[63] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *arXiv preprint arXiv:2003.06620*, 2020.

[64] Qiang Zhao, Chen Zhu, Feng Dai, Yike Ma, Guoqing Jin, and Yongdong Zhang. Distortion-aware cnns for spherical images. In *IJCAI*, pages 1198–1204, 2018.

[65] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[66] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, Federico Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360 depth estimation. In *2019 International Conference on 3D Vision (3DV)*, pages 690–699. IEEE, 2019.

[67] 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.