Bi-projection-based Foreground-aware Omnidirectional Depth Prediction

Qi Feng* Hubert P. H. Shum† Shigeo Morishima‡

*Waseda University †Durham University ‡Waseda Research Institute for Science and Engineering

E-mail: *fengqi@ruri.waseda.jp, †hubert.shum@durham.ac.uk, ‡shigeo@waseda.jp,

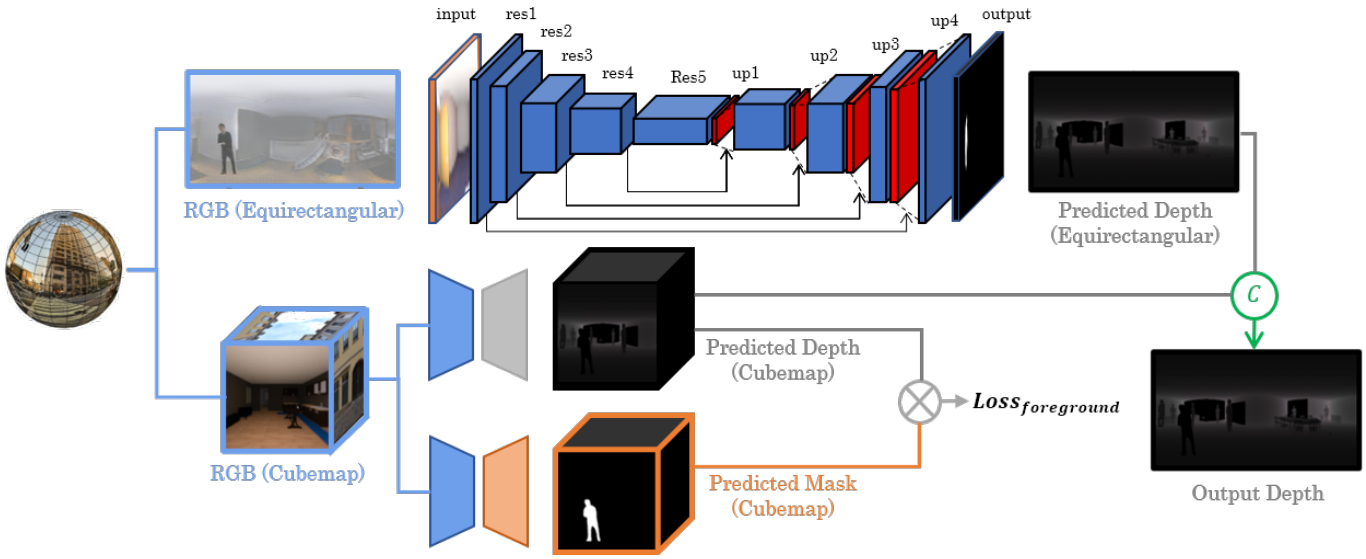

Figure 1 The proposed bi-projection-based foreground-aware dense depth prediction method for omnidirectional images. We first transform spherical contents into equirectangular and cubemap projection. For the equirectangular projection we directly regress depth maps with a distorted CNN kernel. For the cubemap projection, we simultaneously predict the semantic segmentation and depth maps. After calculating an additional local loss for foreground objects, we merge the cubemap depth map with the equirectangular one to achieve consistent global prediction with sharp and detailed local regions.

Abstract

Due to the increasing availability of commercial 360-degree cameras, accurate depth prediction for omnidirectional images can be beneficial to a wide range of applications including video editing and augmented reality. Regarding existing methods, some focus on learning high-quality global prediction while fail to capture detailed local features. Others suggest integrating local context into the learning procedure, they yet propose to train on non-foreground-aware databases. In this paper, we explore to simultaneously use equirectangular and cubemap projection to learn omnidirectional depth prediction from foreground-aware databases in a multi-task manner. Experimental results demonstrate improved performance when compared to the state-of-the-art.

1 Introduction

Commercial omnidirectional cameras have gained increasing research interest in recent years thanks to their capability to capture surrounding environments efficiently with high quality. As it provides great potential applications in the field of medical and education, self-driving vehicles, virtual reality [1], etc., the need for better visual reasoning algorithms in the context of omnidirectional media rises accordingly. One of the most important visual reasoning capabilities is to predict depth information from a single color image as it provides structural clues of the surroundings, and thus facilitate applications such as 3D rendering [2] and robotic navigation [3]. Predicting depth information for omnidirectional images with existing methods does not show satisfying results. Depth prediction is an ill-posed problem: ambiguity in scales, different lighting conditions can be problematic to obtain an accurate prediction. Existing approaches that are capable to predict depth information from a single omnidirectional image do not generalize well to real-world scenarios. Upon experimenting [4], most existing methods are designed for indoor-only static samples without any consideration of foreground objects. The main cause if that currently methods either use a synthetic database (i.e. PanoSunCG [5] and SceneNet [6]) or a captured database (i.e. Stanford 2D-3D [7] and Matterport3D [8]).

Only very few databases include foreground objects into consideration. However, while [9] focuses on solving the foreground problem through an image processing method with an additional loss term that emphasizes the prediction of the foreground objects, it is essentially a trade-off between consistent global depth prediction and local ones. A bi-projection-based method [10] has shown promising improvements, yet they fail to explore consolidating accurate prediction for foreground objects with additional segmentation information.

In this paper, we propose to obtain accurate and sharp foreground depth prediction with consistent global predictions by exploring a bi-projection algorithm that consists of an equirectangular projection that predicts global depth information and a cubemap projection that simultaneously estimates the depth and the semantic segmentation of cube faces. While the equirectangular projection ensures a consistent and smooth global context, the cubemap faces provide insights regarding local details with a smaller FOV. By merging two projections together, we achieve better depth prediction for omnidirectional images with foreground objects. We choose humans as the foreground object to study as it is one of the most interesting subjects with a high appearing frequency in different media. We quantitatively and qualitatively evaluate our method against state-of-the-art and show a better performance. Our major contribution is the bi-projection-based network architecture that facilitates accurate depth prediction for foreground objects. This research is best applied in fields including 3D reconstruction of omnidirectional contents and virtual reality.

The rest of the paper is organized as follows: we revisit learning-based depth prediction in Section 2. In Section 3, we explain the foreground-aware training database we use in this research. In Section 4, we describe the proposed bi-projection network architecture to leverage the database. Implementation and experimental results are presented in Section 5, and Section 6 concludes this work.

2 Related Work

Predicting depth information from a color image is one of the most important tasks when comes to 3D geometry understanding [11], as accurate depth maps can facilitate a wide range of applications such as autonomous driving [12]. While classic depth prediction methods usually rely on probabilistic models and hand-crafted features [13], recent learning-based algorithms have shown greatly improved accuracy and quality. A standard supervised approach of depth prediction often trains on paired color and depth maps, trying to build an implicit relationship between the two through convolution networks. A multi-scale network that refines predictions in a coarse-to-fine fashion showed improved local details [14], and a fully convolutional network structure [15] has further improved the accuracy of the prediction. Multitask learning [16] the semantic segmentation and the depth prediction is also prevalent in understanding scene geometry due to their complementarity. For unsupervised methods, this is usually achieved through geometric models, such as building stereo correspondence from successive frames without the need to acquire ground truth supervisory signal [2], which is computationally expensive and time-consuming. Learning intermediary disparity map for generating depth maps [17] or reconstruct 3D models through multiple scenes and produce pseudo depth afterward [18] are also viable to learn dense depth prediction.

For depth prediction from a single color image in the context of omnidirectional format, the literature is scarcer compared to traditional perspective images. Directly applying perspective-based approaches on omnidirectional images with equirectangular projection usually produce less accurate and coherent results due to heavy distortions introduced during the process of converting spherical contents onto a flat 2D plane [4]. To cope with this problem, rotation equivariant CNNs [19] and graph-based learning [20] try to learn directly from spherical signals. However, such equivariant architectures provide a lower network capacity, rendering generative tasks such as learning depth prediction ineffective. Instead, using cubemap projection is a popular choice for multiple benefits [21]. Since projecting spherical contents onto six faces of a cube can eliminate distortion for each face to a great extent, it is made possible to adopt perspective-based methods with minimal effort. Moreover, as each face has a reduced FOV, cubemap projection puts more focus on local objects compared to equirectangular projection [10]. However, since each face is processed independently, the discontinuity along edges is problematic for many applications. A common approach to alleviate this problem is through padding edges during the process [22] of merging predictions back to a single output. In this work, we try to incorporate both equirectangular and cubemap projection to complement each other, so that while the equirectangular prediction can provide a global context, the proposed network can still yield accurate results for foreground objects.

3 Foreground-aware Omnidirectional Databases

In this section, we explain the method to prepare the foreground-aware database through an image processing

technique described in [9]. Previously, high-quality training samples are generated using a scanning device or rendered through 3D models with virtual cameras. While it is impossible for the scanning device to including any dynamic foreground objects due to its lengthy capturing time, it is also costly to introduce realistic foreground objects that highly resemble real-world appearance. To this end, we use a data augmentation pipeline that takes the advantage of abundant perspective color-depth databases and correctly composites the object of interest onto existing omnidirectional databases through Z-buffer. As shown in Fig. 2, we first employ a Mask R-CNN model to acquire pixel-perfect binary masks of foreground objects. We then crop out perspective paired color and depth batches. We finally composite the batches with correct occlusions and distortions by conducting cubemap projection before and after the processes.

To acquire pixel-perfect binary masks of foreground objects automatically, we propose to take advantage of abundant obtainable existing 2D perspective databases. We efficiently obtain high-quality segmentation masks by adopting a Mask R-CNN network with a backbone of ResNet-101. By training on the COCO database extensively, we can predict per-pixel label masks in a real-time manner. With per-instance prediction, we can cope with scenarios that incorporate multiple foreground objects when compared to using a simpler U-Net network. As explained in the introduction section, human is an important subject with complex deformations and detailed depth with a high appearing frequency in real-world scenarios. Therefore, we demonstrate our method by choosing humans as the foreground object to be generalized to other objects. In this research, we use the PKU-MMD database [23], which contains color and depth videos of a human subject performing a wide range of motions. It contains multiple view-angles and subject appearance.

For omnidirectional background, we adopt existing databases in both synthetic and real domains to showcase the effectiveness of our method. This includes captured databases, the Stanford 2D-3D database, and the Matterport3D database, and also synthetic databases, the SunCG [24] and the SceneNet [6] as well. Since the batches of foreground objects are acquired from traditional perspective images, a direct composition will lead to distortion in the omnidirectional context. Therefore, we first perform cubemap projection before the composition process to obtain different local batches of the background. With depth information of both foreground and background batches for local batches, we can easily solve the occlusions through Z-buffer and preserve correct

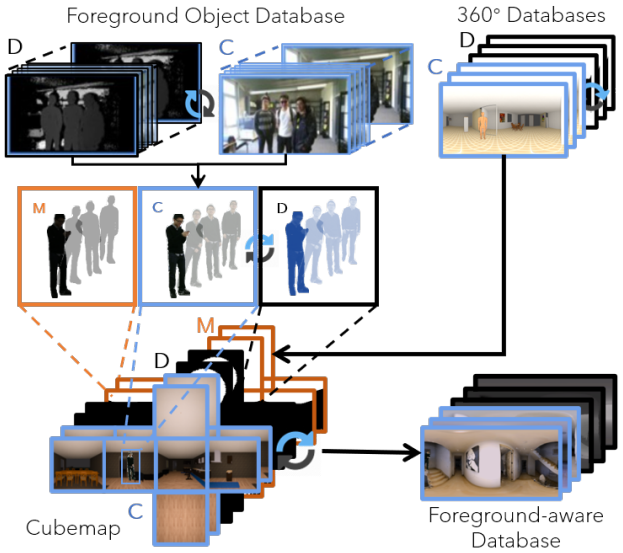

Figure 2 The process of generating foreground-aware omnidirectional database. With foreground object databases, we use a Mask R-CNN to acquire batches of binary mask, color and depth images, and then composite to the cubemap projection of omnidirectional databases through z-buffer. We finally obtain a foreground-aware database through equirectangular projection.

depth annotations even after the composition. Afterward, we reverse the cubemap projection process with equirectangular projection to generate omnidirectional samples with foreground objects. During the experiment, our training data consists of 25,000 synthetic and 25,000 realistic samples with correct depth annotations. Some examples are shown in Fig. 3. It is worth mentioning that more variations can be achieved through a similar process described in the previous step through re-purposing other perspective databases.

4 Bi-projection based Depth Prediction

We explain the proposed foreground-aware bi-projection-based depth prediction method for omnidirectional images in this section. We use a multi-branch end-to-end structure that incorporates two different projections to achieve more consistent global context and detailed local foreground object features. The proposed architecture is shown in Fig. 1. In particular, the first branch learns regressing depth information from a single omnidirectional image in the format of equirectangular, providing surrounding information through a wider FOV. As directly using equirectangular images usually causes blurred prediction for local objects with steep gradient changes, the second branch uses cubemap projection to make it more effective to learn local features. With a narrower FOV,

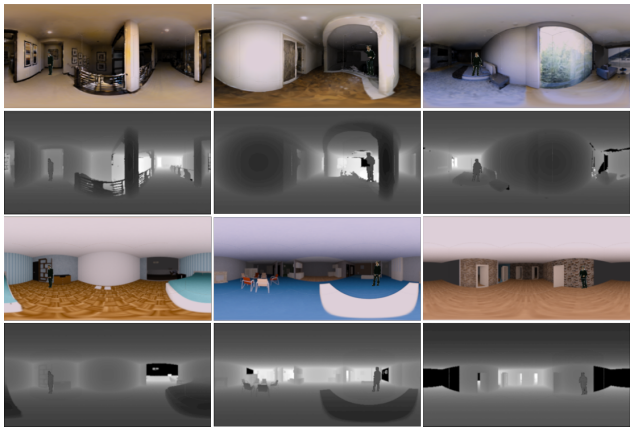

Figure 3 Some examples from the generated foreground-aware database. The first and second rows are generated examples of paired color images and dense depth maps from capturing-based omnidirectional databases and the bottom two rows are generated from synthetic databases.

cube faces provide more insights into shape and boundary for foreground objects. We further utilize the binary mask for foreground objects prepared in the previous step and propose a depth/semantic segmentation multi-task learning scheme for the cubemap branch to strengthen the loss for foreground objects with a foreground object loss.

$$L_{foreground} = \|D_{cubic\,depth} \otimes M_{foreground}\|^2,$$

and thus the overall loss function for the network is

$$L_{total} = \sum_i (\alpha_i L_{output\,depth} + \beta_i L_{smooth} + \gamma L_{foreground}),$$

while the α, β and γ are the weight coefficients for each loss term. Since semantic segmentation and depth prediction are two tasks usually learned together to reveal the scene layout [25] [26] [27], we can improve the accuracy of depth prediction through this foreground-aware network.

For the equirectangular branch, it regresses dense depth information from omnidirectional images with an encoder-decoder structure by progressively downscales and upscales to the depth output. Since skip connections are used to preserve features from higher levels, we adopt Resnet as the encoder of the network. We take advantage of a distorted CNN filter [24] that changes filter sizes with regard to the coordinate on the equirectangular image to improve the effectiveness when training directly on spherical images. We use a traditional L2 loss to calculate the depth loss and a smoothness regularization term [4] to improve the consistency of the output.

We further introduce spherical padding and a convolution module at the end of both branches to ensure a consistent merged output. While cubemap projection does not quite suffer from the distortion when projecting spherical information onto a 2D plane, it instead introduces discontinuity at the boundaries of each face. To alleviate this problem, we adopt a spherical padding technique [10] that increases the FOV when rendering each face, and connects them afterward to address the consistency issue. After two branches produce respective dense depth predictions, we unify both branches by concatenating them together and pass through a convolution module described in [28].

5 Experimental Results

5.1 Implementation details

For generating the foreground-aware database, we randomly selected 25,000 synthetic and 25,000 realistic omnidirectional image pairs from existing databases, and split them into training and validation sets with a ratio of 80% and 20%. We then composite foreground objects (i.e. human) onto the acquired samples with a resolution of 512 x 256. We implement the aforementioned network structure with PyTorch[29], Adam optimizer [30], Xavier initialization [31], and a learning rate of 2e-4. The training process is conducted on an Nvidia RTX 2080Ti graphic card. The parameters used for training are $[\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma] = [0.482, 0.245, 0.121, 0.061, 0.090]$. We use the same metrics from the previous work [2] to evaluate our method. At runtime, predicting images with the same resolution can achieve real-time performance.

5.2 Evaluation

We quantitatively and qualitatively evaluate our proposed method in this section. In Table 1, we present the result of depth prediction when compared to the state-of-the-art omnidirectional method, [4]. The upper column showcases the effectiveness when applied to the synthetic domain, while the bottom column demonstrates its efficacy in real-world scenarios. We can observe that our method shows favorable performance with improved accuracy across the board against the existing method when benchmarking with accuracy metrics. We believe that the increased accuracy attributes to the bi-projection network architecture in addition to the semantic segmentation task in the cubemap projection branch. This is qualitatively verified through Fig. 4, as we can observe that our model generalize to unseen data with foreground objects and yield satisfying depth prediction.

6 Discussion and Conclusion

In this paper, we present a foreground-aware bi-projection-based depth prediction method for omnidirec-

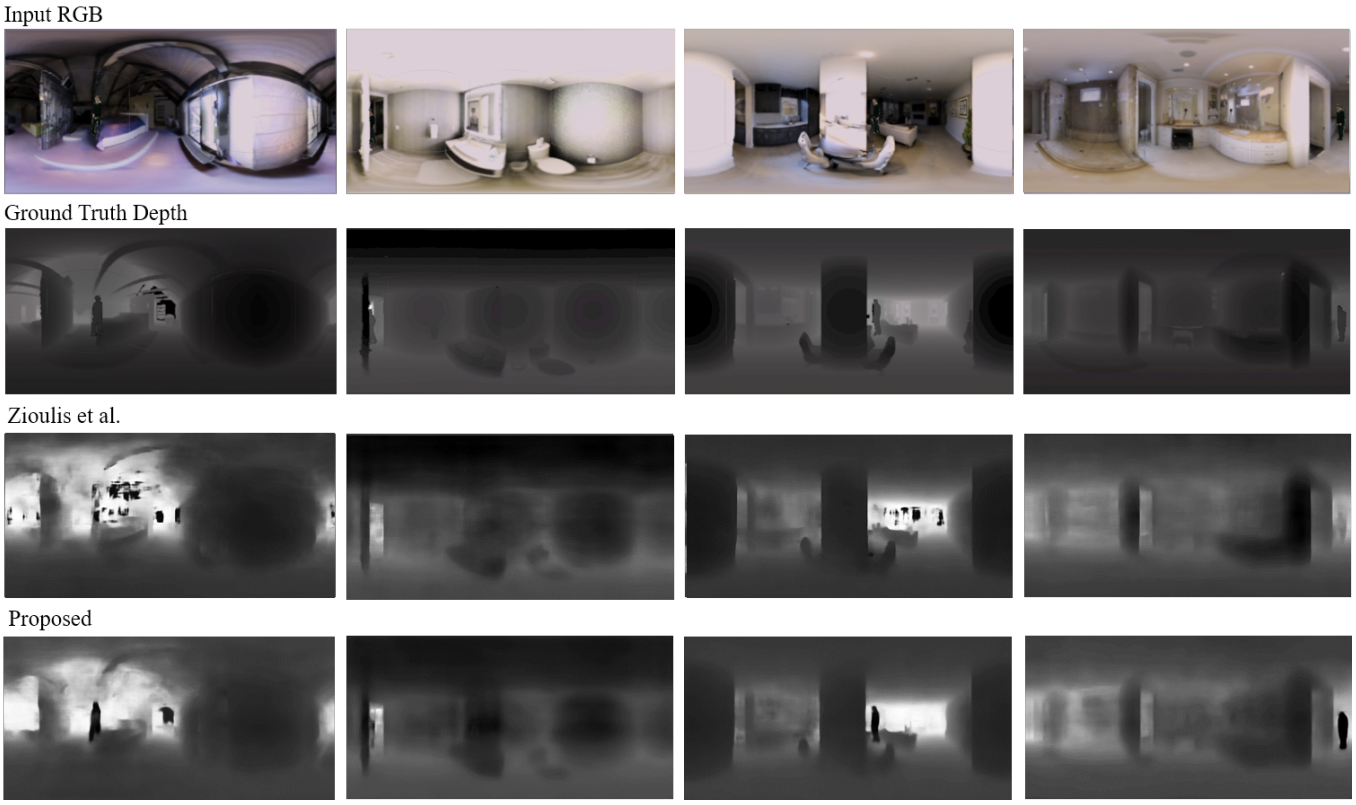

Figure 4 Qualitative comparison against the state-of-the-art method when tested on realistic images.

Table 1 Quantitative comparison against state-of-the-art methods.

Metrics	Database	OmniDepth [4]	Ours
Abs Rel ↓	Synthetic	0.3789	**0.2279**
Sq Rel ↓	Synthetic	0.2893	**0.2134**
RMSE ↓	Synthetic	0.6878	**0.5999**
RMSE log ↓	Synthetic	0.5225	**0.2257**
$\delta < 1.25$ ↑	Synthetic	42.45%	**78.41%**
$\delta < 1.25^2$ ↑	Synthetic	79.26%	**92.85%**
$\delta < 1.25^3$ ↑	Synthetic	92.57%	**97.13%**
Abs Rel ↓	Real	0.3190	**0.2246**
Sq Rel ↓	Real	0.2180	**0.1727**
RMSE ↓	Real	**0.5993**	0.6042
RMSE log ↓	Real	0.4788	**0.2427**
$\delta < 1.25$ ↑	Real	69.88%	**75.37%**
$\delta < 1.25^2$ ↑	Real	84.54%	**91.73%**
$\delta < 1.25^3$ ↑	Real	91.50%	**96.66%**

tional images and explore using image processing methods to generate color/depth databases with dynamic foreground objects. The proposed architecture produces consistent global depth prediction with the equirectangular projection, while enforcing local detailed features through the cubemap projection. An additional foreground loss acquired through a multitask learning approach of semantic segmentation complementarily provides sharper boundaries of predicted foreground objects. With quantitative and qualitative evaluation, we successfully verified the effectiveness of the proposed method. We believe the ability to accurately predict depth information for omnidirectional images can facilitate a wide range of applications such as 3D reconstruction and virtual reality.

Currently, the database generation still requires synthetic/captured omnidirectional databases to composite on, which greatly limits the generalizability of this approach. In the future, we plan to explore self-supervised methods or multi-view-based generation methods to further reduce the need to acquire expensive ground truth data.

References

[1] R. Shimamura, Q. Feng, Y. Koyama, T. Nakatsuka, S. Fukayama, M. Hamasaki, M. Goto, and S. Morishima, "Audio–visual object removal in 360-degree videos," *The Visual Computer*, vol. 36, no. 10, pp. 2117–2128, 2020.

[2] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[3] J. Gaspar, N. Winters, and J. Santos-Victor, "Vision-based navigation and environmental representations with an omnidirectional camera," *IEEE Transactions on robotics and automation*, vol. 16, no. 6, pp. 890–898, 2000.

[4] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 448–465.

[5] F.-E. Wang, H.-N. Hu, H.-T. Cheng, J.-T. Lin, S.-T. Yang, M.-L. Shih, H.-K. Chu, and M. Sun, "Self-supervised learning of depth and camera motion from 360 videos," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 53–68.

[6] A. Handa, V. Pătrăucean, S. Stent, and R. Cipolla, "Scenenet: An annotated model generator for indoor scene understanding," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5737–5743.

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

[8] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[9] Q. Feng, H. P. Shum, R. Shimamura, and S. Morishima, "Foreground-aware dense depth estimation for 360 images," 2020.

[10] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, "Bifuse: Monocular 360 depth estimation via bi-projection fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 462–471.

[11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[12] Y. Chen, J. Wang, J. Li, C. Lu, Z. Luo, H. Xue, and C. Wang, "Lidar-video driving dataset: Learning driving policies effectively," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5870–5878.

[13] A. Saxena, S. H. Chung, A. Y. Ng *et al.*, "Learning depth from single monocular images," in *NIPS*, vol. 18, 2005, pp. 1–8.

[14] D. Eigen and R. 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*, 2015, pp. 2650–2658.

[15] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[16] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[17] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[18] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992.

[19] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical cnns," in *International Conference on Learning Representations*, 2018.

[20] R. Khasanova and P. Frossard, "Graph-based classification of omnidirectional images," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 869–878.

[21] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360 videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1420–1429.

[22] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos and spherical images," *International Journal of Computer Vision*, vol. 126, no. 11, pp. 1199–1219, 2018.

[23] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.

[24] Y.-C. Su and K. Grauman, "Learning spherical convolution for fast features from 360 imagery," in *Advances in Neural Information Processing Systems*, 2017, pp. 529–539.

[25] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[26] S. Kim, K. Park, K. Sohn, and S. Lin, "Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields," in *European conference on computer vision*. Springer, 2016, pp. 143–159.

[27] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3994–4003.

[28] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5038–5047.

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.