Foreground-aware Dense Depth Estimation for 360 Images

Qi Feng

Waseda University
Tokyo, Japan
fengqi@ruri.waseda.jp

Hubert P. H. Shum

Northumbria University
Newcastle upon Tyne, UK
hubert.shum@northumbria.ac.uk

Ryo Shimamura

Waseda University
Tokyo, Japan
s-ryo@akane.waseda.jp

Shigeo Morishima

Waseda Research Institute for
Science and Engineering
Tokyo, Japan
shigeo@waseda.jp

ABSTRACT

With 360 imaging devices becoming widely accessible, omnidirectional content has gained popularity in multiple fields. The ability to estimate depth from a single omnidirectional image can benefit applications such as robotics navigation and virtual reality. However, existing depth estimation approaches produce sub-optimal results on real-world omnidirectional images with dynamic foreground objects. On the one hand, capture-based methods cannot obtain the foreground due to the limitations of the scanning and stitching schemes. On the other hand, it is challenging for synthesis-based methods to generate highly-realistic virtual foreground objects that are comparable to the real-world ones. In this paper, we propose to augment datasets with realistic foreground objects using an image-based approach, which produces a foreground-aware photorealistic dataset for machine learning algorithms. By exploiting a novel scale-invariant RGB-D correspondence in the spherical domain, we repurpose abundant non-omnidirectional datasets to include realistic foreground objects with correct distortions. We further propose a novel auxiliary deep neural network to estimate both the depth of the omnidirectional images and the mask of the foreground objects, where the two tasks facilitate each other. A new local depth loss considers small regions of interests and ensures that their depth estimations are not smoothed out during the global gradient's optimization. We demonstrate the system using human as the foreground due to its complexity and contextual importance, while the framework can be generalized to any other foreground objects. Experimental results demonstrate more consistent global estimations and more accurate local estimations compared with state-of-the-arts.

Keywords
Depth Estimation, Scene Understanding, Data Augmentation, 360 images

Figure 1: A demonstration of incorrect representations of dynamic objects (e.g. a running person) captured with an omnidirectional RGB-D scanning device.

Figure 2: The previous approach of inserting human models introduces the problem of severe domain bias. This is demonstrated by comparing synthetic data (left) with captured data (right).

1 INTRODUCTION

As 360 cameras have become more popular and efficient, the need for image processing algorithms applicable to omnidirectional images increases. The ability to estimate depth from a monocular omnidirectional image can greatly benefit a wide range of applications including navigation in robotics [7], stereoscopic rendering in graphics [10], augmenting virtual objects [14].

Existing omnidirectional approaches produce sub-optimal estimations on real-world scenarios due to their lack of consideration of dynamic foreground objects. For captured-based approaches, using a stereo setup of two 360 cameras will inevitably include the other camera in the captured data [4]. While recent 360-capable scanning devices [3] can acquire paired RGB and ground truth depth of scenes with improved quality, they are incapable of including any dynamic object as a result of scanning and stitching scheme (Figure 1) [1]. For synthesis-based approaches [25], although researchers attempt to solve this problem by inserting 3D models into the scene to improve the
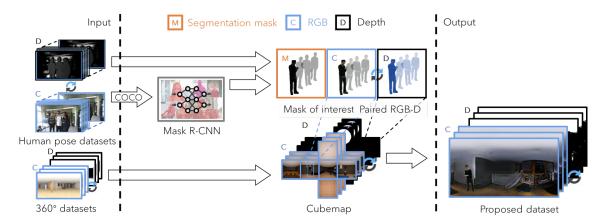
Figure 3: The pipeline of the proposed data synthesizing system. The left section shows the input datasets, the middle section shows the intermediate results, and the right section shows the output. We generate masks of the interested region with mask R-CNN and corresponding RGB-D batches from the input 2D dataset. The batches are then composited to the input 360 dataset with regard to the depth information.

prediction (Figure 2), it is challenging to efficiently generate highly-realistic virtual foreground objects that resemble real-world ones [2], and non-photorealistic data often lead to undesirable and inaccurate outputs.

In this paper, we tackle the problem of foreground by first augmenting datasets with realistic foreground representations. We observe that given the same object with a determined distance, its scale in spherical images should remain consistent. Taking advantage of it, we effectively composite color data of abundant and easily obtainable 2D datasets and rendered omnidirectional images according to ground truth depth maps to ensure correct occlusion representations. To preserve correct distortions in equirectangular images, we project the data to cube maps before and after compositions.

We then propose a novel auxiliary deep neural network that estimates both the mask of the foreground objects and regresses the depth of the omnidirectional images. With the depth and segmentation estimations, we design a new local depth loss of dynamic foreground objects to achieve more consistent depth predictions. This solves the problem that small areas with steep local gradients often got minimized when regressing the global gradient of the prediction, resulting in areas of interest that are frequently smoothed out in existing work.

In this paper, we choose humans as the dynamic foreground object to show the efficacy of our approach. As a foreground object, human shares both a high complexity in deformation and non-uniform depths, and great importance being one of the most interested and common subjects to deliver the context of the image. By showcasing accurate estimations of human, we demonstrate the ability of our method to be generalized to other foreground objects.

Experimental results show that the proposed method yields more consistent global estimations and more accurate local estimations against contemporary state-of-the-art models quantitatively and qualitatively. This research is best applied in fields including occlusion-aware augment reality, stereoscopic rendering.

Our contributions are summarized as follows:

1. We propose a method to synthesize an RGB-D omidirectional dataset with dynamic foreground objects to tackle the challenge of estimating the depth of them in the context of spherical images. The dataset is offered to promote future research.

2. We employ the proposed auxiliary network that estimates depth and segmentation masks to calculate a new local depth loss of dynamic foreground objects. This can resolve the issue of steep local gradient getting smoothed out during optimization and improve the estimation results of local regions. The source code is publicly offered online.

The rest of the paper is organized as follows: we revisit learning-based monocular depth estimation methods and methods for synthesizing training data in Section 2. In Section 3, we explain the novelty of our dataset and describe the generation framework. In Section 4, we describe the network architecture and the proposed loss function to leverage the dataset. Details of experiments are presented in Section 5 along with qualitative and quantitative evaluations. Finally, Section 6 concludes this work.

2 RELATED WORK

2.1 Learning-based Monocular Depth Estimation for Omnidirectional Images

Estimating the depth given a monocular RGB image is one of the most fundamental capabilities in understanding the 3D geometry of the scene [13]. A wide range of applications in robotics, graphics, virtual reality, etc.

can benefit from more accurate depth predictions. Owing to more established machine learning algorithms, learning an implicit relation between color and depth has seen significant progress recently.

A variety of algorithms [19] [17] have been proposed by training a model with collected color and ground truth depth images in a supervised fashion. Lately, numerous strategies have been proposed to achieve a more coherent and accurate monocular depth estimation. Multi-scale networks [5] make coarse global depth prediction and refine the local prediction. Multitask learning [15] [23] with multiple regression and classification objectives is also prevalent in understanding scene geometry and semantics due to their complementarity. A fully convolutional network architecture [17] that endows novel up-sampling blocks achieved impressive accuracy and efficiency.

Unsupervised methods focused on a stereo correspondence framework to cope with the need for an expensive secondary supervisory signal. This is either accomplished by synthesizing stereoviews with left/right consistency [10] to produce intermediary disparity map [6] [27], or multi-view consistency with structure-from-motion (SfM) [28] to learn a dense disparity prediction.

To yield accurate estimations of both global and local objects in the context of the omnidirectional domain, lacking paired data with dynamic foreground objects and distortion introduced by equirectangular projection will result in poor outputs for supervised approaches. On the other hand, while some unsupervised approaches do not explicitly require paired datasets, issues like distortions and occlusions still persists.

Therefore, predicting the depth of 360 contents with the aforementioned 2D approaches often yields suboptimal results [29]. Failing to learn feature representations in the equirectangular domain inevitably leads to inferior accuracy and coherency. To improve the performance of prediction in 360 contents, cubemap projection is one of the most popular choices. By projecting spherical signals onto faces of a cube, six non-distorted square patches can still be processed with existing convolution techniques. However, while such an issue may not be critical in certain tasks such as stylization and classification, the lack of consistency between the output of each patch is more pronounced in depth regression. Recently, methods for enabling rotation-equivariance in CNNs were proposed by Cohen [4]. However, since such equivariant architectures provide a lower network capacity, only single variable regression problems were demonstrated. Inspired by [26], the state-of-the-art method [29] incorporated distorted CNN filters to improve the performance of fully convolutional networks with skip connections and showed impressive predictions of equirectangular images. However, without any consideration on fore-

ground objects, the network will penalize small areas with a steep local gradient when regressing the global gradient of the prediction, resulting in areas of interest such as humans are frequently missing in the output.

2.2 Synthesizing Omnidirectional Datasets

Since the standard method to approach monocular depth estimation is to train a model directly from paired RGB images and ground truth depth, the performance of such supervised approaches cannot produce better results than the limits of its training data. With advanced imaging devices and depth sensors, high-quality datasets consisting of traditional perspective images are easily obtainable, for instance, KITTI [8], NYUv2 [24], Make3D [22], etc. However, obtaining paired 360 data is not as straightforward as using traditional imaging devices with calibrated color and depth sensors such as Kinect to capture 2D contents. Using a stereo setup of two 360 cameras to calculate disparity is challenging due to the presence of occluded regions [20]. In the case of omnidirectional images, both cameras will inevitably include the other camera in the captured data. Recent scanning devices are capable of acquiring datasets that consist of paired 360 RGB and ground truth depth of static scenes with improved quality, such as Stanford 2D-3D [1] and Matterport3D [3].

However, existing methods fail to include any dynamic object in the scene. As a result of a scanning and stitching scheme, trying to include dynamic foreground objects in the captured data [1] [3] will lead to distorted and incorrectly composited images, as shown in Figure 1. [29] repurposed 3D model datasets, SunCG [25] and SceneNet [11], to render 360 synthesis-based RGB-D images with virtual cameras. However, a model trained with synthetic data does not necessarily generalize well to real-world scenarios, due to dataset bias. As observable in Figure 2, a previous attempt to resolve this issue by inserting human models into the existing synthetic dataset suffers from a severe domain bias from the real-world scenarios. However, because of the inability to include realistic human representation in the existing omnidirectional RGB-D dataset, the performance of all previous methods is greatly limited when applied to real-world scenarios with humans.

3 FOREGROUND-AWARE PHOTOREALISTIC 360 DATASET

To produce a foreground-aware photorealistic dataset for machine learning algorithms, we explain our method of augmenting datasets with realistic foreground objects using an image-based approach in this section. The pipeline of our method is visualized in
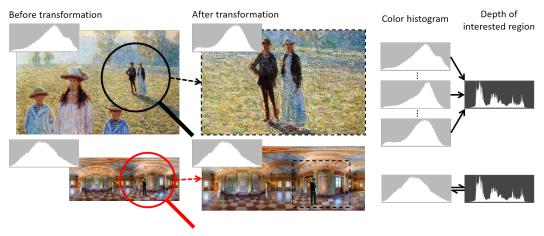
Figure 4: Since traditional 2D images can be processed with perspective transformations, it is difficult to establish a correspondence between color and depth information. In comparison, 360 images cannot be cropped or zoomed and hence have a scale-invariant RGB-D correspondence. As shown on the right side, in the 2D plane, different color representations map to the same depth map of interested regions. One the other hand, 360 images share a one-to-one mapping between color and depth, and every object has a fixed scale.

Figure 3. As shown in Figure 4, based on the observation that 360 images can circumvent challenges brought by perspective transformations in the traditional 2D plane, we effectively composite color data of abundant and easily obtainable 2D datasets and rendered omnidirectional images with z-buffer. We employ a Mask R-CNN network to predict pixel-perfect masks of the dynamic foreground objects. With the acquired masks of interest, we can obtain perspective paired color and depth batches. With cubemap projections done before and after compositions, we can composite with correct occlusions and distortions.

3.1 Scale-invariant RGB-D Correspondence in 360 Images

In this section, we explain the novelty and feasibility of compositing existing 2D RGB-D datasets onto equirectangular images.

We observe that it is difficult to establish a correspondence between color and depth in the traditional 2D domain. We take perspective transformations as an example and demonstrate with Figure 4. During the process of "zooming in" onto the target region (dashed box), the global color data changes continuously while the depth of the target area stays the same, forming a many-to-one mapping. It is particularly true in the real-world: when we use binoculars to observe the same object, even given the prior knowledge of an object's average size, it is inherently harder to estimate the distance without knowing the magnification.

One the other hand, the relation between color and depth in 360 images is scale-invariant. While some perspective transformations such as cropping will make 360 images no longer spherical, rotation and zoom will not affect the global color representation of the original

image after down-scaling. Therefore, given the same object with a determined distance, the appearance of the target region in 360 images should remain consistent.

Based on this observation, we exploit such an advantage of omnidirectional images by inversely composite local regions onto them with regard to the depth information. In this work, we choose z-buffer to composite owing to its simple implementation, high efficiency and compatibility of occlusions.

3.2 Synthesizing RGB-D Foreground Batches

3.2.1 General Foreground Synthesis

To automatically acquire paired color and depth maps of a dynamic foreground object, we can either capture with sophisticated RGB-D sensors or take advantage of abundant and easily obtainable existing datasets in the traditional 2D domain. In order to efficiently acquire highly accurate segmentation masks of the input data, we adopt a Mask R-CNN model with a backbone of ResNet-101, trained with the COCO dataset to predict per-pixel label masks. The strengths of per-instance prediction and less complex post-processing are the main reason we choose Mask R-CNN over a simpler U-Net network. During prediction, our implementation predicts per-person-instance masks in a near-real-time speed (5 fps) with high accuracy. Some examples are shown in Figure 10 and Figure 11. With acquired masks for areas of interest, we crop batches from the input RGB-D data accordingly.

3.2.2 Human Batch Synthesis

Since human as a dynamic foreground object shares both a high complexity in deformation and non-uniform
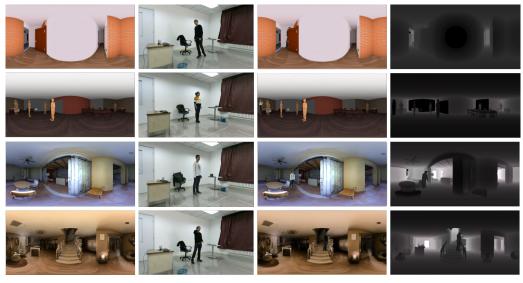
Figure 5: Generated examples with the proposed method. From left to right: rendered color images with original omnidirectional datasets, samples from an input human pose dataset, generated omnidirectional images with humans, and corresponding depth maps.

depths, we choose humans to show the efficacy of our approach. At the same time, humans have great importance being one of the most interested and common subjects to deliver the context of the image. By showcasing accurate estimations of human, we demonstrate the ability of our method to be generalized to other foreground objects. In this work, we repurpose the PKU-MMD dataset [18], which contains calibrated and synchronized RGB-D video sequences. This large-scale dataset includes motions of 51 categories performed by 66 distinct subjects. It contains different views, sufficient intra-class variations and adequate classes of motions to ensure a robustness prediction result.

3.3 Synthesizing RGB-D Omnidirectional Data with Foregrounds

3.3.1 Omnidirectional Background Synthesis

Since paired real-world 360 RGB-D datasets with humans are not available to our knowledge, to alleviate the difficulty of evaluating the accuracy between ours and the-state-of-the-art approaches, we use a similar strategy matching with [29] to render paired and realistic omnidirectional RGB-D images from the Stanford 2D-3D dataset and the Matterport3D dataset captured with professional 360-capable scanning devices. Specifically, a path tracing renderer with a virtual omnidirectional camera is used to generate the samples. The light source is positioned identically with the virtual camera. Omnidirectional depth maps with linear distances of each pixel are generated with Z depth. To show the effectiveness of our method across different domains and to benchmark the accuracy with synthetic 360 datasets, identical processes are brought out with the SunCG [26] and the SceneNet [11] as well.

3.3.2 Compositing Foregrounds and Backgrounds

Since the RGB-D local batches are captured in the traditional 2D domain, a direct composition will lead to distorted and unrealistic appearances in the 360 context. To cope with this challenge, both RGB and depth map of each rendered omnidirectional sample is projected onto a cube map through cubic projection. With ground truth depth information of both foreground batches and background faces, the composition is done through highly efficient and effective Z-buffer, preserving correct depth annotations and in-scene occlusions. To simulate real-world scenarios, batches are randomly composited to lower halves of 4 surrounding cube faces, while faces of the ceiling and the floor are not used during composition. Finally, a reverse cubemap projection is done to generate high-quality RGB-D equirectangular samples with dynamic foreground objects.

In this work, our proposed dataset consists of 25,000 realistic and 25,000 synthetic equirectangular samples with synchronized color information and depth annotations. Abundant variation is achieved through a sufficiently wide range of indoor scenes as backgrounds, and a large human batch pool acquired in the previous step as foregrounds.

4 DEPTH ESTIMATION FOR OMNIDIRECTIONAL IMAGES

This section presents our proposed end-to-end learning model to estimate a depth map from an equirectangular image. As shown in Figure 6, We use two fully-convolutional encoder-decoder structured networks,
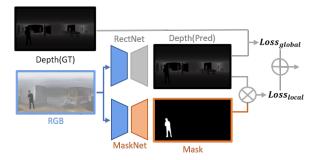
Figure 6: An overview of the proposed depth estimation network. The weight of the auxiliary MaskNet is fixed when training the depth estimation model RectNet [29].

RectNet and MaskNet to regress depth and predict masks of local regions respectively from a given RGB input. The RectNet that resembles the design in the literature can regress depth with changing filters in an omnidirectional context. To take advantage of the generated dataset with dynamic foreground objects, we leverage the generated masks of interested areas to train the auxiliary MaskNet. By calculating both local depth loss and global loss, our network further improves the consistency in local predictions.
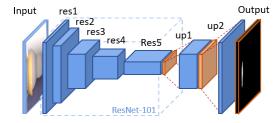
4.1 Network Structure

Figure 7: The architecture of the fully convolutional auxiliary MaskNet. The encoder of our network shares the same structure of ResNet-101 [12], followed by a decoding process with two upsampling layers to predict the mask of the target object.

The proposed network approaches dense depth estimation from monocular RGB images shares an encoder-decoder design that progressively downscales and upscales to the target representation through regression. Skip connections similar to ResNet structures can help to preserve the information from a higher level during regression while preventing vanishing gradient. When applied to equirectangular images, inspired by [26], we incorporate rectangular filters with changing sizes according to rows of the input to cope with the characteristic that the density of information, or namely the distortion level changes along the vertical axis but invariant along the horizontal axis. In addition to L2 depth loss to regress the prediction, a neighborhood smoothness regularization term [29] is also calculated to improve the global consistency of the output.

However, small regions with steep gradient changes usually got smoothed out during the regression and missing in the prediction. This can be observed in Figure 10. Predictions of human severely suffer from this issue. To tackle this limitation, we introduce an auxiliary network, MaskNet, to calculate the local depth loss of humans. The MaskNet network that predicts masks of foreground objects from equirectangular RGB inputs has the architecture shown in Figure 7. It is trained with the COCO dataset and finetuned with generated equirectangular RGB images with foreground objects and corresponding segmentation masks to minimize a cross-entropy loss. The weight is fixed during training the depth estimation model.

4.2 Loss Function

We train the depth estimating network in a completely supervised fashion with input of the generated foreground-aware RGB-D dataset. To address the problem of vanishing local gradients for areas of interest while keeping the desirable properties of the original RectNet like consistent global predictions, the total loss of our model consists of three different terms:

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

while the α, β and γ are the weights for each loss term. Since the loss is calculated under different scales i, the estimations of lower scales are interpolated with nearest neighbors are concatenated together to form the final output. The depth loss L_{depth} is regressed by minimizing the least square errors between the groudtruth depth maps D_{gt} and the predicted depth maps D_{pred}:

$$L_{depth} = \|D_{gt} - D_{pred}\|^2.$$

The smoothness loss is calculated by $\|\nabla D_{pred}\|^2$ to minimize the gradient of the prediction. In order to calculate the local depth loss, we pass the equirectangular color image C_{input} through the trained auxiliary network $\mathbf{M}$ to obtain the mask of human instances $M_{human} = \mathbf{M}(C_{input})$, so we can calculate the local depth loss with

$$L_{local} = \|D_{pred} \otimes M_{human}\|^2.$$

By minimizing the local depth loss, we can ensure that spatially closer pixels within the same area of interest would have closer depth values.

5 EXPERIMENTS

In this section, we first evaluate our data augmentation method by presenting quantitative comparisons between models trained with existing omnidirectional datasets and our generated datasets. We then verify the performance of the proposed network by comparing it to the state-of-the-art omnidirectional depth estimation algorithm. Finally, to evaluate the effectiveness of our method in real-world scenarios with human objects,

Table 1: Quantitative results of different training datasets. Error metrics are calculated on a global basis.

Dataset	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Synthetic	0.4918	0.4133	0.8944	0.6550	0.4083	0.6806	0.8212
Proposed Synthetic	**0.3789**	**0.2893**	**0.6878**	**0.5225**	**0.4245**	**0.7926**	**0.9257**
Realistic	0.3765	0.3540	0.8864	0.5230	0.5907	0.7500	0.8926
Proposed Realistic	**0.3190**	**0.2180**	**0.5993**	**0.4788**	**0.6988**	**0.8454**	**0.9150**

For four error metrics, absolute relative difference (Abs Rel), squared relative difference (Sq Rel), root mean square error (RMSE) and RMSE log, lower values are better. For percentage of inliers under threshold $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$, higher values are better. Same for tables below.

Table 2: Quantitative evaluation against other models. Error metrics are calculated on a global basis.

Model	Training Set	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
RectNet [29]	Proposed Syn	0.3789	0.2893	0.6878	0.5225	0.4245	0.7926	0.9257
Proposed	Proposed Syn	**0.2895**	**0.2354**	**0.5957**	**0.4272**	**0.7440**	**0.8805**	**0.9284**
RectNet [29]	Proposed Real	0.3190	0.2180	0.5993	0.4788	0.6988	0.8454	0.9150
Proposed	Proposed Real	**0.1984**	**0.0817**	**0.3286**	**0.2608**	**0.7298**	**0.8984**	**0.9727**

we offer comparative qualitative results of estimating unseen images by different methods.
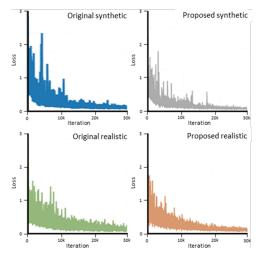
5.1 Training Details

Figure 8: Learning curves of models respectively trained with original synthetic, original realistic, proposed synthetic and proposed realistic datasets.

For fair comparisons, we randomly acquired 25,000 samples from existing synthetic omnidirectional datasets to train models as the existing synthetic dataset, and then we acquired 25,000 samples from existing realistic omnidirectional datasets to train models as the existing realistic dataset. We respectively generate 25,000 synthetic samples and realistic samples augmented with human objects to train models as our proposed datasets. Each 512 x 256 sample has color information and corresponding ground truth depth annotation. We randomly split samples from each dataset into training and validation datasets with a ratio of 80% and 20%. All networks in this paper

are implemented with PyTorch [21] on an Nvidia RTX 2080Ti graphic card and trained with Adam optimizer [16], Xavier initialization [9], and a learning rate of 2e-4. Training parameters of our networks are $[\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma] = [0.482, 0.245, 0.121, 0.061, 0.090]$, while parameters of training previous RectNet models are $[\alpha_1, \alpha_2, \beta_1, \beta_2] = [0.535, 0.272, 0.134, 0.068]$. The same quantitative metrics from the literature [10] [29] are used for evaluation. During experiments, predicting a single image approximately costs 100 ms with the same setup.
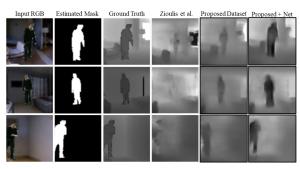

Figure 9: Estimated depth information of local regions with different configurations. An ablation study shows that using our augmented dataset can improve the accuracy of local regions, and the proposed network shows an improved consistency with clearer boundaries.

5.2 Quantitative Results

Table 1 presents the results of the state-of-the-art models respectively trained with existing synthetic and realistic datasets and our proposed datasets. We observe that when tested on unseen samples with human objects, networks trained with our proposed datasets outperform the existing ones. The increased performance in accuracy against previous methods attributes to more
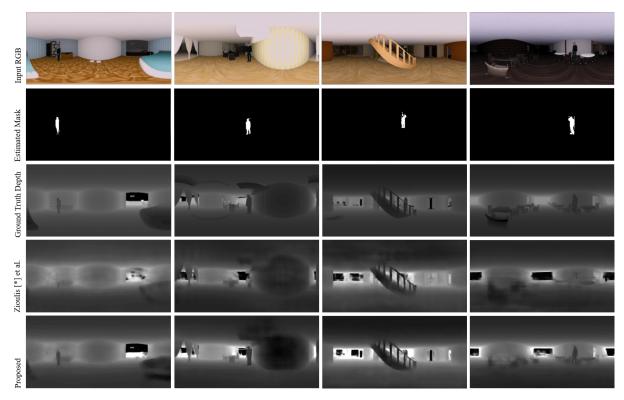
Figure 10: Qualitative comparison between each model when tested on synthetic images.
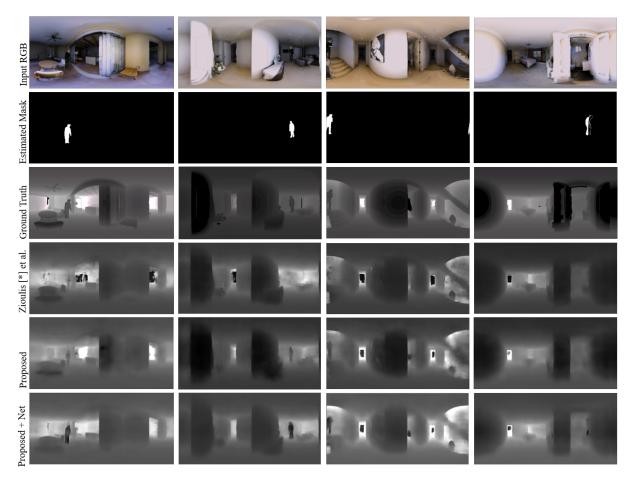

Figure 11: Qualitative comparison between each model when tested on realistic images.

accurate estimations of local human regions, as can be observed in Figure 9. By further quantitatively evaluating the accuracy of estimations between our proposed network and the state-of-the-art models, we can observe the inferior performance of previous approaches as expected in Table 2. Depth estimations of local human regions are further refined with our proposed network.

5.3 Qualitative Results

To qualitatively evaluate our models' ability to generalize to unseen data, we further acquire and augment samples from the SunCG and the Matterport3D that come from other locations different from training datasets. As we can observe in Figure 10 and Figure 11, our models perform better to estimate the depth of both synthetic and realistic scenes with a human. While previous models yield human depth estimations that are blended with the background and have a blurred edge, our models can predict much clearer and human-shaped results. It is worth mentioning that although all omnidirectional samples used in the experiment only cover indoor settings, our method works with outdoor cases as well.

After observing generated samples, we believe there are many challenges left to overcome. First, even our method can augment foreground objects, we do not take lighting into consideration during the process. This unnaturalness may lead to less robust estimation in certain scenarios (e.g. scenes with very high brightness).

5.4 Ablation Study

In Figure 9, we compare the accuracy of depth estimations for local regions under different configurations. Specifically, we compare using original data and proposed data to train only the depth estimation network without the auxiliary MaskNet at first to validate the effectiveness of our data generation method. We then use augmented data to train depth estimation networks with the auxiliary MaskNet, and verified that the local depth loss can successfully improve the consistency of estimated depth within areas of interest. As we can observe in Figure 9, our method significantly outperforms the state-of-the-art in local depth estimation.

6 CONCLUSION

We have presented a data augmentation method to generate high-quality equirectangular datasets with paired color and ground-truth depth annotations by repurposing abundant and easily obtainable 2D RGB-D datasets. With this dataset, we further introduced and implemented an auxiliary network that calculates local depth loss to resolve an issue that small regions of interest are frequently smoothed out during optimizing global gradients. We take human, a crucial subject in 360-degree contents, as an example to show the efficacy of our approach. We showed improved accuracy of our approach

compared to the state-of-the-art technique. We believe that the ability to estimate depth for foreground objects in 360 images can benefit a wide range of applications such as navigation in robotics and augmenting virtual objects with occlusions.

Currently, our data augmentation method is based on the premise that both 2D and 360 data are captured with similar extrinsic parameters (e.g. cameras are aligned horizontally, positioned at average eye-level height) and lighting conditions, while it is true for most data captured in lab conditions, its application for in-the-wild images is limited. Furthermore, our approach works for both indoor and outdoor settings. Nevertheless, for outdoor settings, a higher dynamic range of luminosity and sunlight's ambient IR will render capturing RGB and depth information inherently difficult. For future work, we aim to explore generating samples with different lighting conditions with GANs to improve the robustness of depth estimation.

7 REFERENCES

[1] 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.

[2] A. Atapour-Abarghouei and T. P. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2800–2810, 2018.

[3] 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.

[4] T. Cohen, M. Geiger, J. Köhler, and M. Welling. Spherical cnns. *International Conference on Learning Representations (ICLR)*, 2018.

[5] 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*, pp. 2650–2658, 2015.

[6] 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*, pp. 740–756. Springer, 2016.

[7] J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Transactions on robotics and automation*, 16(6):890–898, 2000.

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. IEEE, 2012.

[9] 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*, pp. 249–256, 2010.

[10] 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*, pp. 270–279, 2017.

[11] 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)*, pp. 5737–5743. IEEE, 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[13] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 654–661. IEEE, 2005.

[14] J. Huang, Z. Chen, D. Ceylan, and H. Jin. 6-dof vr videos with a single 360-camera. In *2017 IEEE Virtual Reality (VR)*, pp. 37–44. IEEE, 2017.

[15] 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*, pp. 7482–7491, 2018.

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

[17] 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)*, pp. 239–248. IEEE, 2016.

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

[19] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5162–5170, 2015.

[20] K. Matzen, M. F. Cohen, B. Evans, J. Kopf, and R. Szeliski. Low-cost 360 stereo photography and video capture. *ACM Transactions on Graphics (TOG)*, 36(4):148, 2017.

[21] 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.

[22] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

[23] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

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

[25] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1746–1754, 2017.

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

[27] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pp. 842–857. Springer, 2016.

[28] 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*, pp. 1983–1992, 2018.

[29] 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)*, pp. 448–465, 2018.