Depth Estimation from a Single Omnidirectional Image using Domain Adaptation

Yihong Wu
School of Electronics and Computer Science, University of Southampton
UK
yihongwu@soton.ac.uk

Yuwen Heng
School of Electronics and Computer Science, University of Southampton
UK
y.heng@soton.ac.uk

Mahesan Niranjan
School of Electronics and Computer Science, University of Southampton
UK
mn@ecs.soton.ac.uk

Hansung Kim
School of Electronics and Computer Science, University of Southampton
UK
H.Kim@soton.ac.uk

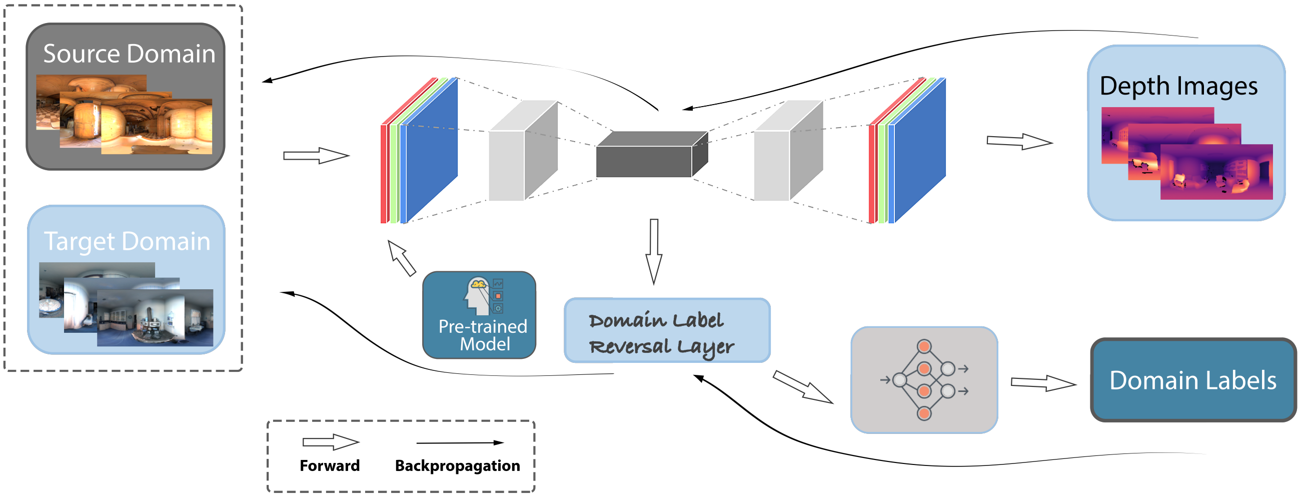


Figure 1: Proposed architecture for omnidirectional single image depth estimation. The input data in the dashed box on the left are RGB omnidirectional images, including the source domain with labels and unlabelled target domain, given with domain labels as 0 and 1, respectively. The encoder-decoder of the upper part of the image is an improved model from [Alhashim and Wonka 2018]. We changed the backbone of the encoder as a ResNet50 model pre-trained on the ImageNet. The right side shows the model outputs, including depth maps and domain labels. The lower part shows the domain label reversal layer and the domain classifier with two fully connected layers.

ABSTRACT

Omnidirectional cameras are becoming popular in various applications owing to their ability to capture the full surrounding scene in real-time. However, depth estimation for an omnidirectional scene is more difficult than normal perspective images due to its different system properties and distortions. It is hard to use normal depth estimation methods such as stereo matching or RGB-D sensing. A deep-learning-based single-shot depth estimation approach can be a good solution, but it requires a large labelled dataset for training. The 3D60 dataset, the largest omnidirectional dataset with depth labels, is not applicable for general scene depth estimation because it covers very limited scenes. In order to overcome this limitation, we propose a depth estimation architecture for a single omnidirectional image using domain adaptation. The proposed architecture gets labelled source domain and unlabelled target domain data together as its input and estimated depth information of the target domain using the Generative Adversarial Networks (GAN) based method.

CVMP '21, December 6–7, 2021, London, United Kingdom
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-9094-1/21/12…$15.00
https://doi.org/10.1145/3485441.3485649

The proposed architecture shows >10% higher accuracy in depth estimation than traditional encoder-decoder models with a limited labelled dataset.

CCS CONCEPTS

• **Computing methodologies → Computer vision representations**.

KEYWORDS

single image, depth estimation, omnidirectional image, domain adaptation

ACM Reference Format:
Yihong Wu, Yuwen Heng, Mahesan Niranjan, and Hansung Kim. 2021. Depth Estimation from a Single Omnidirectional Image using Domain Adaptation. In *European Conference on Visual Media Production (CVMP '21), December 6–7, 2021, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485441.3485649

1 INTRODUCTION

In the past decades, 3D scene reconstruction and representation have been essential tasks in computer vision and robot vision. Depth estimation, predicting or measuring the distance to visible surfaces from the sensor, is one of the most important modules in 3D scene reconstruction [Steger et al. 2018]. Depth sensors, such as LiDAR and Time-of-Flight cameras, can generate relatively accurate depth maps, but these sensors usually have limitations, e.g., lack of texture, short sensing range, expensive reconstruction process [Alhashim and Wonka 2018], low resolution [Zioulis et al. 2018], etc. Depth estimation techniques from visual inputs can be an alternative solution.

There have been many different approaches for depth estimation from visual inputs, such as using motion parallax in videos [Lei et al. 2015], multi-view geometry from multiple cameras [Steger et al. 2018], and depth cues from a single image [Bhoi 2019]. The single-shot approach has more flexibilities in its applications, while the stereo or multi-view approaches require more constraints in system configuration, such as camera calibration and synchronisation between cameras. A human can perceive depth even from one eye through various monocular depth cues about the scene, e.g., shadow, motion parallax, relative size, etc., based on prior knowledge and experiences [Howard 2012]. Estimating depth from single images is a challenge in artificial intelligence (AI) that may take computer vision beyond more widely considered simpler tasks such as object recognition and localisation segmentation to scene understanding. A comparative experiment has been conducted on various aspects, such as object size and camera pose, to reveal exactly how the network learns depth from a single image input [Dijk and Croon 2019].

Depth estimation has a wide range of applications. Augmented reality (AR) applications require depth information of a scene to naturally integrate virtual objects into the given scene [Lee et al. 2011]. Depth information is also useful in surveillance applications [Asif and Soraghan 2009; Lamża et al. 2013]. Predicting distance is essential for autonomous cars, as the vision-based autonomous driving system needs to measure the distance between the current vehicle with surrounding vehicles, pedestrians, and obstacles [Janai et al. 2020; Luo et al. 2018; Wang et al. 2019].

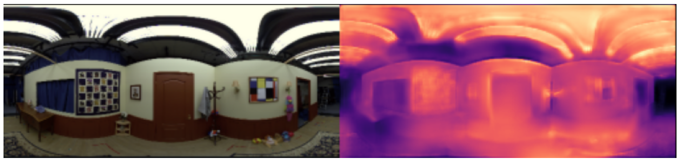


Figure 2: Depth estimation result for a different indoor scene with different light condition, building type and omnidirectional camera

There is one more problem in scene representation and understanding using vision sensors. Normal perspective cameras with a limited field-of-view (FoV) provide only a partial observation of the scene. Observation of the whole surrounding 3D environment requires multiple calibrated and synchronised sensors. Omnidirectional cameras provide a good solution, as they capture the full surrounding scenes in one image [Kim and Hilton 2013]. Omnidirectional images should be processed in slightly different ways from the processes of normal perspective images due to the distortions in the images. There have been several studies on omnidirectional single image depth estimation. For example, there was an end-to-end U-Net shape model that took the lead for depth estimation of omnidirectional images [Zioulis et al. 2018]. Bi-projection based depth estimation architecture using both the equirectangular image and its cube map projections has also been proposed [Wang et al. 2020].

These methods require a large labelled dataset for training, but it is difficult to collect a large depth-labelled dataset because a synchronised RGB-D sensor for omnidirectional capture is not generally available. Even the 3D60 dataset [Zioulis et al. 2018], the largest omnidirectional images set with depth labels, contains only three types of scenes: office, home and synthetic indoor scenes, which is not enough for general scene depth estimation. Figure 2 shows an example. The Stanford2D3D set in the 3D60 dataset consists of 898 images with depth labels captured from 6 office buildings. It was trained with RectNet model [Zioulis et al. 2018] and shows 95% depth accuracy in the given dataset. However, it shows very poor results when we apply it to a different indoor scene image we capture in our studio. We can easily observe that the estimated depth map includes lots of errors, especially on planar regions such as the walls, ceiling and floor, where a more smooth transition of depth field is expected. The model was over-fitted to the specific scenes in the dataset.

In this paper, we propose to use domain adaptation for omnidirectional single image depth estimation in order to overcome this problem. In the proposed domain adaption method, the labelled source domain data and unlabelled target domain data are taken as input to infer depth information of the target domain leveraging the features in the source domain. It provides a good solution for the problem of single omnidirectional image depth estimation when the limited labelled set is available for the source data scene. We evaluated the proposed architecture on existing datasets by limiting the number of labels. The result showed that the proposed architecture outperforms a traditional encoder-decoder model by over 10% in terms of depth accuracy when the labelled set is very limited.

2 RELATED WORK

Single-view Depth Estimation for Normal Perspective Images. Most single image depth estimation methods are based on convolutional neural networks, and they are often seen as a regression from an RGB image to a depth map [Fu et al. 2018]. Eigen *et al.* [Eigen et al. 2014] proposed an end-to-end model concatenating two networks for depth estimation of normal perspective images. These two networks are coarse and fine networks, and they are based on AlexNet. The output of the coarse network is concatenated as part of the input of the fine network. In order to get higher performance, Alhashim & Wonka [Alhashim and Wonka 2018] proposed an end-to-end model with a deeper encoder and a shallow decoder to estimate depth maps with RGB images as input. Although depth estimation of normal perspective images may get good performance, most of these works aim at normal perspective images [Abuowaida and Chan 2020; Alhashim and Wonka 2018; Fu et al. 2018; Hambarde and Murala 2020]. Our experiments show that they do not perform well for omnidirectional images due to the different FoV and distortions.

Single-view Depth Estimation for Omnidirectional Images. Release of omnidirectional image datasets with depth labels such as Matterport3D and StanFord2D3D [Karakottas et al. 2018], and neural network models for omnidirectional images enabled depth estimation of omnidirectional images. Similar to the depth estimation of normal perspective images, there was an end-to-end networked neural network based on U-Net shape to train the RGB images and predict the depth maps [Zioulis et al. 2018]. Wang *et al.* [Wang et al. 2020] proposed to combine two networks with the equirectangular image and its corresponding cubic projection map to avoid the distortion problem of omnidirectional images. Although these models show good performance with the given labelled datasets, the reality is that omnidirectional imaging is suffering a serious lack of labelled datasets, as mentioned in the Introduction. One way to overcome this problem is to use domain adaptation. It can solve the problem of differences in data distribution of different scenes by mapping information in different fields to a feature space [Pan and Yang 2009].

Transfer learning. Transfer learning is the application of the knowledge or patterns learned in a particular field or task to different but related fields. It allows the model to be transferred from labelled data in the source domain to data in the target domain [Pan and Yang 2009]. Alhashim & Wonka [Alhashim and Wonka 2018] utilised the pre-trained DenseNet model on ImangeNet to train and fine-tune with the NYU image dataset [Silberman et al. 2012] to predict the corresponding depth map from the RGB image. Yeh *et al.* [Yeh et al. 2020] used transfer learning and ordinal regression to achieve the depth estimation of NYU and KITTI [Geiger et al. 2013] datasets.

Domain Adaptation. Domain adaptation is a method to map the data distributed in different source domains and target domains to a feature space [Pan and Yang 2009]. Ganin & Lempitsky [Ganin and Lempitsky 2015] proposed a depth estimation model based on GAN including feature mapping network, label classification network, and domain discrimination network to recognise unlabelled digits dataset. Similarly, Ren & Lee [Ren and Lee 2018] proposed a domain adaptation based architecture for depth estimation of normal perspective RGB images by training computer graphics (CG) images based on a generative adversarial network to predict depth maps.

However, most state-of-the-art models aim at normal perspective images, and only a few of them focus on omnidirectional images. Omnidirectional images contain more distortion [Zioulis et al. 2018] than normal perspective images. This distortion makes the models that aim at normal perspective pictures cannot be directly used to estimate omnidirectional depth maps. A novel architecture is needed for depth estimation for a single omnidirectional image. To solve this problem, we proposed an architecture based on an encoder-decoder model with domain adaptation.

3 METHOD

3.1 Problem Specification

As briefly mentioned in the Introduction, Figure 2 demonstrates that even with a 95% accuracy performance model trained with Stanford2D3D dataset, the performance on an unlabelled different indoor image set that has similar semantic structures is still poor. The picture on the wall and the door in the middle should have the same distance to the camera, and the depth for these parts should be smooth, but it is recognised as a different distance. The model did not get high performance because the existing training dataset covers only a few types of scenes, which leads to overfitting the model in the training process. In addition, it is much more difficult to generate ground-truth depth maps for omnidirectional images than normal perspective images because there is no omnidirectional depth sensor available. A depth sensor takes a lot of time to scan and capture a high-resolution depth map, and manual depth map generation is also hard due to its image distortion and wide FoV. Therefore, the lack of a training depth label set is a serious problem in single omnidirectional image depth estimation.

In order to overcome the poor performance with a new dataset from a different domain and the difficulty of getting a large number of labelled images from new scenes, we propose an architecture based on domain adaptation. By adding the unlabelled target domain image set to the training process, we can achieve better results than the traditional encoder-decoder model, even with limited labelled images.

3.2 Proposed Architecture

The overview of our proposed architecture is illustrated in Figure 1. We propose an architecture based on GAN that not only allows the model to accurately predict the depth of the input RGB images but also cannot distinguish their domain labels. This architecture leverages the domain adaptation technique for omnidirectional depth estimation with input images from different domains, including unlabelled images. In this architecture, the input images are omnidirectional RGB images with given domain labels 0 (source) and 1 (target).

The architecture can be divided into three parts, the encoder, decoder and domain classifier. We improved the end-to-end model from [Alhashim and Wonka 2018] by replacing the original DenseNet169 backbone in the encoder with a pre-trained ResNet50, as our experimental results show that a better performance with it. The encoder with a pre-trained model transforms the RGB images into embedded features, while the decoder predicts the depth maps based on these embedded features. The encoder-decoder is called a depth predictor, and the training process tries to make the predictor's loss as small as possible. The green part in Figure 3 shows the reverse-gradient layer. The domain classifier predicts domain labels based on the reverse features outputted from this reverse-gradient layer and makes gradient descent towards the direction of loss increase. It is used to obfuscate domain labels in the training process so that different domains can be mapped to the same feature space with similar feature distribution, resulting in the domain-invariant features.

Therefore, there are two directions of gradient descent during the training process, the loss of the encoder-decoder model is expected to be as low as possible, while the loss of the domain classifier is expected to be as high as possible. By adding a domain classifier to the end-to-end model, it makes the model unable to identify which domain the images come from [Ganin and Lempitsky 2015]. By loading the model with depth labelled images as the source domain and unlabelled images in the target domain, the model can predict the depth maps of the target domain images.

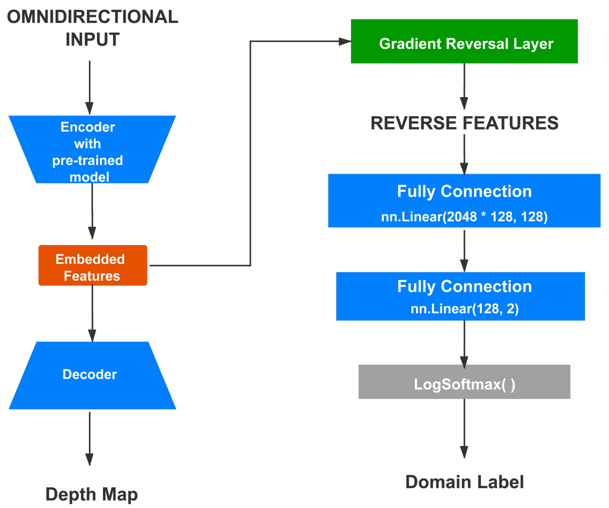


Figure 3: Structure of domain adaptation

3.3 Loss Function

The training loss fuction (Equation 1) is defined as a combination of four loss fucntions, indluding depth loss (Equation 2), Structural Similarity (SSIM) [Wang et al. 2004] loss (Equation 3), and two domain label losses for source domain and target domain, respectively. SSIM is a good loss function for depth estimation tasks [Alhashim and Wonka 2018; Godard et al. 2017].

$$\begin{aligned} L(GT, Output) = \lambda L_{depth}(GT, Output) + L_{SSIM}(GT, Output) \\ + L_{label_s}(GT, Output) + L_{label_t}(GT, Output) \end{aligned} \tag{1}$$

λ is a weight parameter and set as 0.1 according to empirical result [Alhashim and Wonka 2018]. 'GT' represents ground truth depth maps, while 'Output' demonstrates the output depth map from the network, and 'point' means the pixel in the image.

$$L_{depth}(GT, Output) = \frac{1}{n}\sum_{point}^{n}\left|GT_{point} - Output_{point}\right| \tag{2}$$

$$L_{SSIM}(GT, Output) = \frac{1 - SSIM(GT, Output)}{2} \tag{3}$$

The domain label losses, L_{label_s} and L_{label_t}, are calculated with Negative Log-Likelihood Loss (NLLLoss). Note that the losses of them are reverse-gradient.

3.4 Evaluation

In order to quantify and accurately describe the performance of the model, the six metrics about accuracy and loss of models are often used as evaluation indicators as they are all correlated to the performance of models [Alhashim and Wonka 2018; Eigen et al. 2014; Zioulis et al. 2018]. In this section, we introduce these six evaluation metrics: δ_1, δ_2, δ_3, rel, rms, and log_{10}.

3.4.1 Accuracy. Following [Eigen et al. 2014], for accuracy, we use three thresholding accuracy with thresholds 1.25, 1.25^2, and 1.25^3. As shown in Equation 4, it indicates the difference between the two images by comparing the ground truth and the depth map output of the model.

$$\max\left(\frac{GT_{points}}{Output_{points}}, \frac{Output_{points}}{GT_{points}}\right) = \delta < threshold \tag{4}$$

The mean of accumulated δ represents the 'accuracy with thresholding'.

3.4.2 Loss. There are three loss functions to evaluate the robustness of the model: Abs Relative Difference (Equation 5), Linear RMSE(Equation 6) and Log10 RMSE referred from [Alhashim and Wonka 2018] (Equation 7). They are shown as rel, rms, and log_{10}, respectively, in the result tables. 'T' represents the total number of pixels in an image. The Abs Relative Difference metric is to reduce the impact of large distance errors by normalising the error between output and ground truth depth maps. Linear RMSE is a traditional method for measuring regression error, while Log10 RMSE is to reduce the impact of large distance errors as the logarithm makes the errors relative. The smaller the loss values, the better the performance.

$$\frac{1}{|T|}\sum_{GT_{\in}T}\left|\, GT_{points} - Output_{points} \,\right| \,/\, Output_{points} \tag{5}$$

$$\sqrt{\frac{1}{|T|}\sum_{GT\in T}\left\|GT_{points} - Output_{points}\right\|^2} \tag{6}$$

$$\frac{1}{T}\sum_{point}^{T}\left|\log_{10}\left(GT_{point}\right) - \log_{10}\left(Output_{point}\right)\right| \tag{7}$$

4 IMPLEMENTATION

In this section, the proposed architecture is trained and tested at the pixel level and regarded as a depth regression problem. In order to prove that the depth prediction proposed in this research can be used for unlabelled omnidirectional images, we implemented the architecture for omnidirectional image depth estimation and with a house-scene-based dataset and an office-scene-based dataset from 3D60 dataset [Zioulis et al. 2018].

The proposed architecture and models are trained on NVIDIA RTX 3090, with 24GB of CUDA memory. To support research in this area, the code and dataset used in this work were made available at https://github.com/MinisculeDust/single_omnidirectional_image_depth_estimation

4.1 Data Exploration

3D60 dataset [Zioulis et al. 2018] was released with three omnidirectional image datasets, including Matterport3D, Standford 2D3D, and SUNCG. SUNCG is a computer graphic dataset, while Matterport3D and Standford 2D3D are real-world captures. StanFord2D3D is divided into six areas, as they are taken from 6 different office buildings.

Figure 4 shows one area of the Matterport3D dataset, presenting house scenes, while the Stanford2D3D dataset demonstrates the scenes in office rooms dataset in Figure 5.

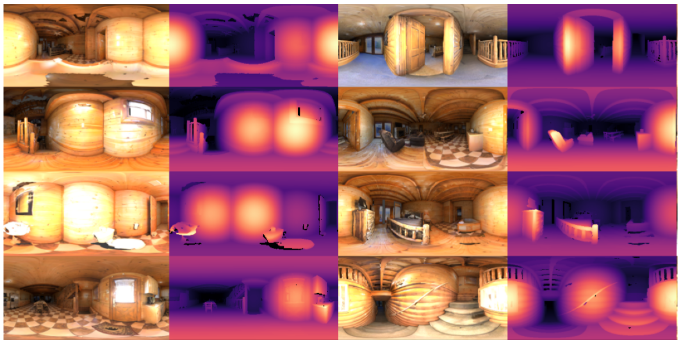


Figure 4: Samples of Matterport3D. The left is original RGB image and right is its corresponding depth map. In the depth map, the brightness represents its depth (the brighter, the closer)

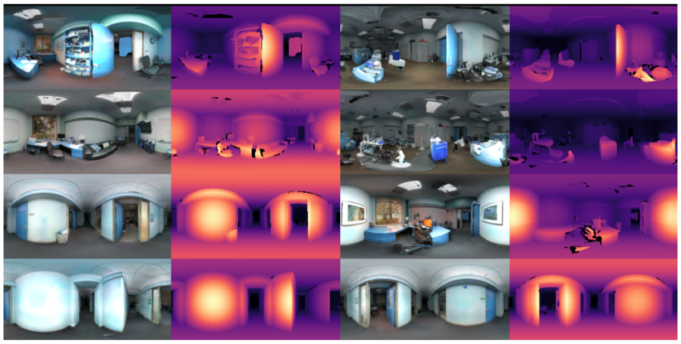


Figure 5: Samples of StanFord2D3D

It should be noted that these sets contain outliers even though we use them as the ground truth. They were captured by RGB cameras and LiDAR sensors. These sensors have limitations of scanning density and also false (or missing) depth in transparent or reflective surface areas. Due to these hardware limitations, there are some missing depth areas. These pixels are recorded as 1,000,000 meters and marked as outliers [Zioulis et al. 2018]. There are still false depth regions, such as the area behind glass or windows, and it is difficult to filter them out.

Stanford2D3D dataset contains 898 images divided into six parts as they are taken in 6 different office buildings. Among them, we selected Area1 with 190 images as a training dataset in our experiments. For data preprocessing, we removed those scenes that contain more than 5% of outliers. After that, the source domain of Stanford2D3D Area1 contains 128 images. The Matterport3D dataset contained 1280 images. We chose one area (88 images after removing scenes containing more than 5% outliers) of this house-scene dataset for the target domain, called 'Matterport3D Area2'. The distribution of depth maps show that depth in the scene is between 0.5 metres and 10 metres apart, and very few areas are above 10 metres. In order to compensate for the inherent problem with the loss terms [Huang et al. 2018; Ummenhofer et al. 2017], we set the maximum distance of depth maps as 10 meters and normalised all depth fields considering the reciprocal of the depth[Alhashim and Wonka 2018].

These datasets were acquired in different circumstances with different cameras but with some similar objects, such as doors and chairs.

4.2 Implementation Details

In all experiments, the input image resolution was 256 × 512, and the batch size was 16. The learning rate was set as 0.0001, and the number of the epoch was set as 100. The Adam-optimizer method was adopted, with parameter $\beta 1 = 0.9$, $\beta 2 = 0.999$.

We did not crop any part of the input images, even though they contained missing points or outliers due to correction preprocessing. We also did not crop the output images before computing the accuracies and losses. This is because, in practice, the image contains different amounts of outliers, which affects the output of the model to some extent. In addition, the purpose of our work is not to simply improve the accuracy of the predicted depth map but to verify that the proposed semi-supervised architecture based on the domain adaptation method can outperform the traditional supervised model with limited labelled data.

5 EXPERIMENTS

5.1 Baseline

In order to simulate the situation of limited labelled images in different scene types, we tested the performance of depth estimation according to the size of the labelled training dataset. We trained the end-to-end models with StanFord2D3D Area1 as the training dataset and Matterport Area2 as the testing set. We gradually reduced the proportion of the training set to simulate the scenario in which a limited amount of data is used to train and predict depth maps of unlabelled RGB images.

Table 1: Performance of ResNet50 backbone encoder-decoder model with different size of dataset

Training Dataset	Testing Dataset	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$rel \downarrow$	$rms \downarrow$	$log_{10} \downarrow$
Whole Stanford2D3D Area1 (128 images)	Matterport3D Area2	0.6576	0.8986	0.9585	0.1918	1.8300	0.0908
40% Stanford2D3D Area1 (51 images)		0.6494	0.8871	0.9587	0.2077	1.9669	0.0935
20% Stanford2D3D Area1 (25 images)		0.6135	0.8394	0.9390	0.2376	2.2732	0.1033

Table 2: Performance of proposed architecture with different size of dataset (uncertainty of the model is shown in Figure 9)

Source Domain	Target Domain	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$rel \downarrow$	$rms \downarrow$	$log_{10} \downarrow$
Whole Stanford2D3D Area1 (128 images)	Matterport3D Area2	0.7259	0.8994	0.9557	0.2189	1.7223	0.0839
40% Stanford2D3D Area1 (51 images)		0.7191	0.9063	0.9546	0.1805	1.6025	0.0827
20% Stanford2D3D Area1 (25 images)		0.7181	0.9252	0.9709	0.1871	1.5431	0.0799

Table 1 shows that the accuracy of estimated depth by the ResNet50 backbone encoder-decoder model decreased to 61.35% of first thresholding accuracy when only 20% of the training set (25 images) were used.

5.2 Domain Adaptation

We chose the Resnet50-backbone model for our domain adaptation architecture as it showed the best performance. The source domain is Stanford2D3D Area1, and the testing dataset is Matterport3D Area2. Table 2 shows the output of the proposed domain adaptation architecture with decreasing number of labelled training images. Overall the proposed architecture shows higher accuracy in depth estimation than the baseline method. One more important observation is that the proposed method kept a similar level of performance (72.59% to 71.81% for δ_1) when the number of the labelled training set is reduced, while the performance of the baseline method decreased from 65.76% to 61.35%. Other metrics also showed similar performance even though the size of the training set had been decreased. They sometimes showed even slightly better performance with less training set (source domain). We guess this is because the training has been less biased to the source domain.

Obtained results of proposed architecture with 20% data are shown in Figure 6. The 1st and 2nd ones are close to the ground truth though they are a bit blurry. It can be observed in the 3rd image that if a wall has some patterns, it influences the depth map and make it bumpy. The 4th one shows some depth errors around the stairs region, but even the ground truth map also have errors in the region. The 5th and 6th ones show errors in the window regions as they are transparent.

We tested the proposed methods for our own dataset captured in various indoor scenes: studio, corridors, and building reception areas. They were captured with Spheron VR[1] and Ricoh Theta S[2] omnidirectional cameras. Figure 7 shows the comparison of depth estimation results of the proposed domain adaptation architecture against the encoder-decoder architecture. Only subjective evaluation can be provided as their ground-truth depth maps are not available. The test scenes are different from the training set. We can see that the proposed method predicted roughly accurate depth maps for the test images. It can be observed that the output generated by domain adaptation architecture has a smoother texture on the object with the same depth plane in the real world. The estimated depth by the proposed model with domain adaptation is closer to the real distance.

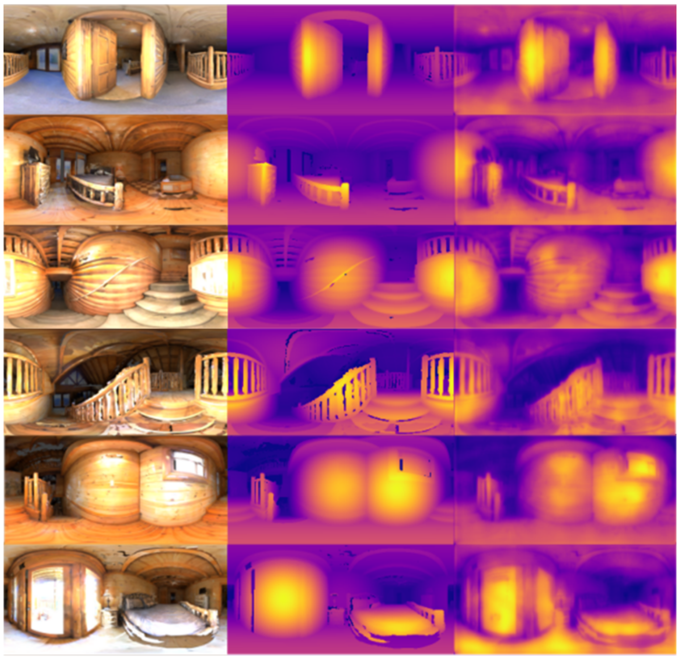


Figure 6: Depth estimation results with the proposed domain adaptation architecture. (Left: Original image, Middle: Ground-truth depth map, Right: Estimated depth map)

In conclusion, the results show that the performance of the proposed architecture outperforms the traditional end-to-end models when the labelled omnidirectional images are limited.

[1] https://www.spheron.com/home.html
[2] https://theta360.com/uk/about/theta/s.html

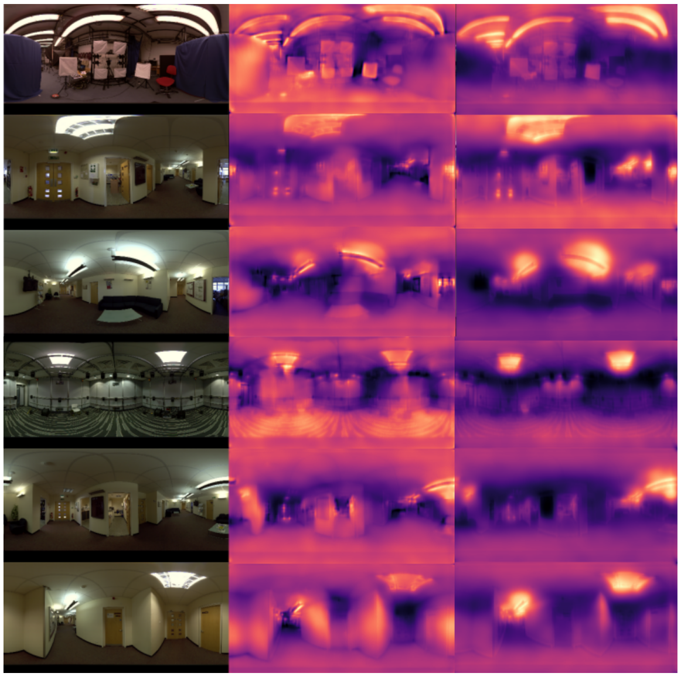


Figure 7: Depth estimation results on our own dataset. (Left: Original image, Middle: Depth map by the encoder-decoder model, Right: Depth map by the proposed domain adaptation model)

5.3 Error Analysis and Discussion

To further analyse the performance of the model, we demonstrate the δ maps of several samples, representing the difference against the ground-truth depth map in the form of a heat map. The delta map in Figure 8 shows errors calculated by the first thresholding accuracy evaluating formula mentioned in Section 3.4.1.

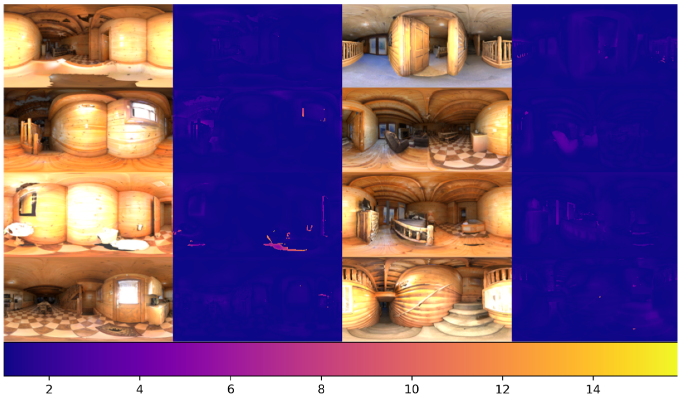
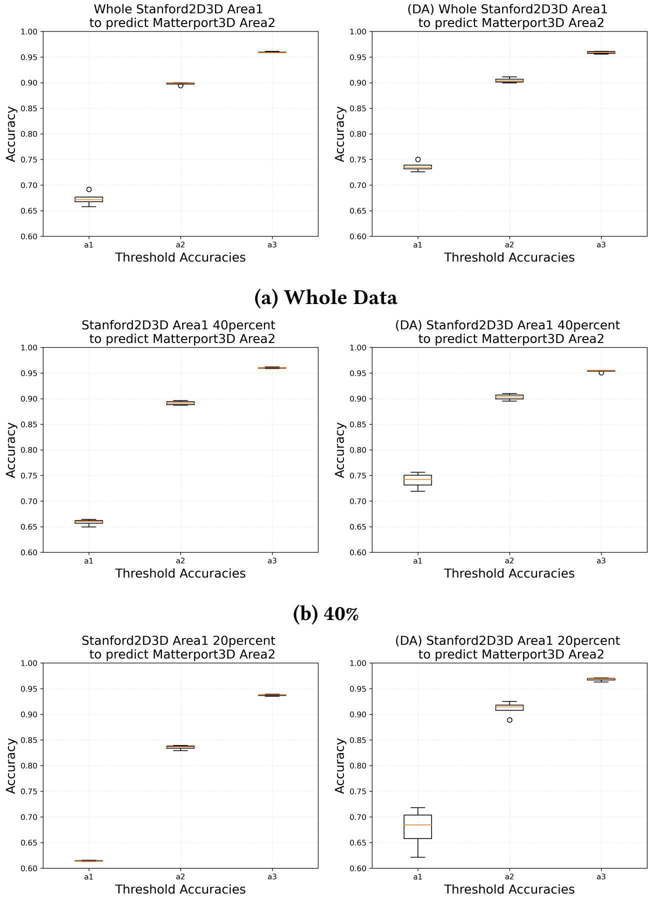


Figure 8: δ Maps of the ResNet50-based model for domain adaptation

For the uncertainty of the results, Figure 9 demonstrate the encoder-decoder model and domain adaptation architecture's performance with different sizes of the source domain, respectively.



(a) Whole Data

(b) 40%

(c) 20%

Figure 9: Different threshold accuracies of depth estimation under different dataset sizes. Uncertainty in estimates displayed as boxplots. (Left: Encoder-Decoder model, Right: Proposed Domain Adaptation model)

Each box contains five values, representing the accuracies on the epoch 80, 85, 90, 95 and 100. It can be observed that although the stability of domain adaptation is not as good as the traditional end-to-end model when the dataset is small, the accuracy is significantly higher than the traditional model.

As previously mentioned, the ground-truth depth maps for training have incomplete regions due to hardware limitations. These false labels may cause the wrong prediction of the model. Figure 10 shows an example showing serious depth errors in the regions with large glass walls. If we consider those glasses as a solid structure, the depth map should show planar depth at the locations of the walls. Even if we ignore glasses, considering the limitation of the sensors, the ground truth for the regions beyond the glasses are still wrong. Most depth sensors, including LiDAR, cannot properly detect and measure transparent or reflective surfaces. This is another reason for the low accuracy of our model as those wrong depths were also considered as ground-truth for training and even for evaluation.

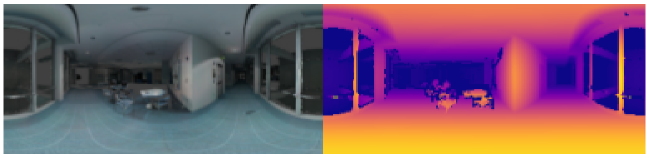


Figure 10: Example of false ground truth. (Left: Original image, Right: Given depth labels)

For practical applications which need to detect even glasses and mirrors, additional modalities, such as acoustic sensors, can be considered to overcome these problems[Kim et al. 2020]. However, we don't consider these issues in this work. Our research focus was to verify the application and efficiency of domain adaptation in the field of omnidirectional images. Therefore we did not focus on improving the performance of the model itself.

6 CONCLUSIONS AND FUTURE WORK

We proposed the depth estimation architecture with domain adaptation to predict scene depth for unlabelled omnidirectional image sets when the labelled training sets are limited. The experiments show that the performance of domain adaptation architecture outperforms the traditional end-to-end model for omnidirectional depth estimation in the situation of a limited number and variety of data. Furthermore, this performance shows that an end-to-end model with domain adaptation can predict the reasonably good quality of depth maps for the omnidirectional images in a different scene without labels. This result means that our work creates a potential direction for depth estimation of unlabelled omnidirectional scenes with limited labelled data.

There are some future works for omnidirectional single image depth estimation, such as getting better performance and solving the problem of scenes with transparent and reflective objects. In order to get better performance, the domain adaptation technique can be combined with more efficient models for distortion, such as deformable convolution [Dai et al. 2017] to consider the distortion of omnidirectional images due to the projection of the spherical domain to the equirectangular domain. In addition, the domain adaptation technique that we used in our proposed architecture depends on representations learned in the source domain that can also be useful in the target domain. For this to be effective, we may need to extract generic features and not fine details from the source. This has been explored in the context of human activity recognition with cascade learning [Du et al. 2019; Marquez et al. 2018] in which a model trained layer-by-layer is shown to extract features in a coarse to fine manner. We expect to adopt this framework to explore better extraction of transferable features. Furthermore, we may find a better method to recognise windows and leverage the idea of masking the window parts to solve this problem during the training process for the images containing windows.

ACKNOWLEDGMENTS

This work was supported by the EPSRC Programme Grant Immersive Audio-Visual 3D Scene Reproduction Using a Single 360 Camera (EP/V03538X/1).

REFERENCES

Suhaila FA Abuowaida and Huah Yong Chan. 2020. Improved Deep Learning Architecture for Depth Estimation from Single Image. *Jordanian Journal of Computers and Information Technology (JJCIT)* 6, 04 (2020), 434–445.

Ibraheem Alhashim and Peter Wonka. 2018. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941* (2018).

Muhammad Asif and John J Soraghan. 2009. Depth estimation and implementation on the DM6437 for panning surveillance cameras. In *2009 16th International Conference on Digital Signal Processing*. IEEE, 1–7.

Amlaan Bhoi. 2019. Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402* (2019).

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 764–773.

Tom van Dijk and Guido de Croon. 2019. How do neural networks see depth in single images?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2183–2191.

Xin Du, Katayoun Farrahi, and Mahesan Niranjan. 2019. Transfer learning across human activities using a cascade neural network architecture. In *Proceedings of the 23rd international symposium on wearable computers*. 35–44.

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283* (2014).

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

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 270–279.

Praful Hambarde and Subrahmanyam Murala. 2020. S2dnet: Depth estimation from single image and sparse samples. *IEEE Transactions on Computational Imaging* 6 (2020), 806–817.

Ian P Howard. 2012. *Perceiving in depth, volume 1: basic mechanisms*. Oxford University Press.

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

Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. 2020. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision* 12, 1–3 (2020), 1–308.

Antonis Karakottas, Nikolaos Zioulis, Dimitrios Zarpalas, and Petros Daras. 2018. 360D: a dataset and baseline for dense depth estimation from 360 images. In *1st Workshop on 360o Perception and Interaction, European Conference on Computer Vision (ECCV), Munich, Germany*. 8–14.

Hansung Kim and Adrian Hilton. 2013. 3d scene reconstruction from multiple spherical stereo pairs. *International journal of computer vision* 104, 1 (2013), 94–116.

Hansung Kim, Luca Remaggi, Sam Fowler, Philip Jackson, and Adrian Hilton. 2020. Acoustic Room Modelling using 360 Stereo Cameras. *IEEE Transactions on Multimedia* (2020).

Aleksander Lamża, Zygmunt Wróbel, and Andrzej Dziech. 2013. Depth estimation in image sequences in single-camera video surveillance systems. In *International Conference on Multimedia Communications, Services and Security*. Springer, 121–129.

Wonwoo Lee, Nohyoung Park, and Woontack Woo. 2011. Depth-assisted real-time 3D object detection for augmented reality. In *ICAT*, Vol. 11. 126–132.

Jianjun Lei, Jianying Liu, Hailong Zhang, Zhouye Gu, Nam Ling, and Chunping Hou. 2015. Motion and Structure Information Based Adaptive Weighted Depth Video Estimation. *IEEE Transactions on Broadcasting* 61, 3 (2015), 416–424. https://doi.org/10.1109/TBC.2015.2437197

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

Enrique S Marquez, Jonathon S Hare, and Mahesan Niranjan. 2018. Deep cascade learning. *IEEE transactions on neural networks and learning systems* 29, 11 (2018), 5475–5485.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

Zhongzheng Ren and Yong Jae Lee. 2018. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 762–771.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*. Springer, 746–760.

Carsten Steger, Markus Ulrich, and Christian Wiedemann. 2018. *Machine vision algorithms and applications*. John Wiley & Sons.

Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. 2017. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5038–5047.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. 2020. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 462–471.

Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. 2019. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8445–8453.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Chia-Hung Yeh, Yao-Pao Huang, Chih-Yang Lin, and Chuan-Yu Chang. 2020. Transfer2Depth: Dual Attention Network With Transfer Learning for Monocular Depth Estimation. *IEEE Access* 8 (2020), 86081–86090.

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