Benefits of Synthetically Pre-trained Depth-Prediction Networks for Indoor/Outdoor Image Classification

Kelly X. Lin Irene Cho Amey Walimbe Bryan A. Zamora Alex Rich

Sirius Z. Zhang Tobias Höllerer

University of California, Santa Barbara
Santa Barbara, CA 93106
{lin400, irenecho, amey, bzamoraflores, anrich, zhening, thollerer}@ucsb.edu

Abstract

Ground truth depth information is necessary for many computer vision tasks. Collecting this information is challenging, especially for outdoor scenes. In this work, we propose utilizing single-view depth prediction neural networks pre-trained on synthetic scenes to generate relative depth, which we call pseudo-depth. This approach is a less expensive option as the pre-trained neural network obtains accurate depth information from synthetic scenes, which does not require any expensive sensor equipment and takes less time. We measure the usefulness of pseudo-depth from pre-trained neural networks by training indoor/outdoor binary classifiers with and without it. We also compare the difference in accuracy between using pseudo-depth and ground truth depth. We experimentally show that adding pseudo-depth to training achieves a 4.4% *performance boost over the non-depth baseline model on DIODE, a large standard test dataset, retaining* 63.8% *of the performance boost achieved from training a classifier on RGB and ground truth depth. It also boosts performance by* 1.3% *on another dataset, SUN397, for which ground truth depth is not available. Our result shows that it is possible to take information obtained from a model pre-trained on synthetic scenes and successfully apply it beyond the synthetic domain to real-world data.*

1. Introduction

Depth information is critical in the field of computer vision. However, collecting accurate depth data is a challenging problem as it requires expensive hardware and time, resulting in the lack of expansive RGB and depth datasets for training. Our research focuses on investigating the usefulness of neural networks pre-trained on synthetic scenes to obtain depth information. Specifically, we tackle the task of indoor-outdoor scene classification using additional depth information from networks pre-trained on synthetic data. Indoor-outdoor classification is a critical part of the scene classification process, with applications including photo tagging, image retrieval [40], robot navigation [7], and color constancy [3].

In recent years, researchers have developed state-of-the-art machine learning models that perform indoor-outdoor classification. Common approaches include Neural Networks [38, 39] and Support Vector Machines [23, 33, 17]. Non-parametric techniques involve performing classification directly on the data without learning any parameters, such as K-Nearest Neighbors [22] and Bayesian methods [34]. Most of these processes involve extracting low-level information such as color, edges, and textures embedded in image pixels.

We use a pre-trained neural network trained on synthetic data to generate depth information used as an additional cue for indoor-outdoor classification. We choose the indoor-outdoor classification task because intuitively, outdoor images are likely to have a higher range of depth values than indoor images. However, estimating accurate absolute depth requires calibrated stereo image pairs, and collecting accurate absolute depth requires expensive sensors. This issue with collecting depth is compounded by a lack of ready-made indoor and outdoor datasets containing ground truth depth, making the training of classification models a difficult task. Existing outdoor datasets such as KITTI [12] are usually collected for specific use cases (self-driving cars) and are acquired using monocular cameras or LiDAR sensors. Although depth sensors have a high sample rate, they have relatively low spatial resolution and lack dense depth

imaging at far ranges.

Using an off-the-shelf single-view depth estimation network, our method provides an alternative to the lack of ground-truth depth data in the form of synthetic depth. Specifically, we utilize Omnidata [10], which is pre-trained on synthetic scenes, in order to generate relative depth of pixels in RGB images. While not as accurate as ground truth depth maps, the ease of use and lack of a need for additional equipment offer a valuable trade-off for obtaining depth information. Additionally, using Omnidata gives us insights into how information transfer from a network trained from synthetic data to real-life data.

Our experiments demonstrate that a ResNet50 convolutional neural network [15] trained with a 4-channel input of RGB and relative depth is able to correctly classify indoor and outdoor scenes to a higher degree than a network trained solely on RGB input. In addition, our analysis shows that a CNN trained on RGB and relative depth retains 63.8% of the performance boost achieved from training a CNN on RGB and ground truth depth. We offer insights into failure cases and why depth information acts as a strong cue for classification by analyzing the depth distributions of indoor and outdoor scenes. Furthermore, we experimentally prove that we can transfer knowledge from the synthetic domain to the real domain.

2. Related Work

We review related work on synthetic pre-training, indoor-outdoor image classification, automatic depth estimation, and large-scale ground-truth and synthetic training datasets.

Synthetic Pre-training: It is well-known now that synthetic data can be successfully employed in training networks for computer vision tasks. For high-level vision tasks, previous work has demonstrated that we can train object detectors, semantic segmentation networks, etc., using simulations [26, 11, 12]. For low level vision tasks, DeTone et al. demonstrated that it is possible to learn deep image descriptors [8, 28] using synthetic pre-training. Similar success has also been observed in training optical flow networks [37]. Our work takes advantage of a synthetically pre-trained depth estimation network [10] to better classify indoor-outdoor scenes.

Indoor-Outdoor Classification: Feature extraction is a crucial part of indoor-outdoor scene classification. Widely studied features include edges, color, texture, and shape properties. Payne and Singh proposed a technique for indoor-outdoor classification using the concept that indoor images have a greater proportion of straight edges in comparison with outdoor images [22]. Their method failed in cases when objects overlapped between indoor and outdoor environments. In addition, most classification algorithms extract color since cues such as green grass and blue skies can be highly discriminating to distinguish between indoor and outdoor scenes [20, 34]. However, the existence of colors similar to sky or grass can often yield false positives. Kim and Park combat this deficiency by partitioning images into 5 blocks, which are then represented using edge and color orientation histograms [17]. Many approaches are also based on textures [32, 33, 34].

Previously, relative depth has been used as a novel information source for Support Vector Machines to discriminate between indoor and outdoor images [23]. Ignazio *et al*. use the Make3D monocular depth estimation tool [31] in conjunction with feature sets that have previously attained high performance in classification tasks. They demonstrated that by using an SVM classifier on the existing Gist feature set [21] and Make3D relative depth feature set, indoor-outdoor classification performance was almost always improved. While the Make3D depth estimation tool was pre-trained using real-life images collected with 3D laser scanners, the Omnidata depth tool was pre-trained on purely synthetic data. We extend upon the work in [23] by analyzing the indoor-outdoor classification performance of a ResNet50 neural network when given Omnidata's relative depth as an additional input. This provides greater insight into transferring knowledge from a synthetic domain to a real domain.

Depth Estimation: Extracting depth information from indoor and outdoor scenes is an important task that provides context about the spatial relationships between entities. Viable solutions for depth extraction include robust point-cloud based methods or stereo-based methods [5, 19, 24]. However, these techniques are limited from being widely used due to their requirement of specific equipment such as Kinect cameras and LiDAR sensors. Pulsed LiDAR scanners have high costs and power consumption. Moreover, existing scanning LiDAR systems achieve low spatial resolution at far ranges due to mechanically-limited angular sampling rates [14]. Time-of-flight depth cameras similarly provide high-resolution depth at close ranges indoors [1, 16], but lack dense depth imaging in far range outdoor scenes that go beyond 30 meters.

Alternatively, pre-trained monocular depth estimation systems such as Make3D [31], MiDaS [25], AdelaiDepth [43] and Omnidata [10] can provide the relative depth of pixels in a singular image. MiDaS demonstrated that zero-shot cross-dataset transfer greatly improved monocular depth estimation, indicating the need for large and diverse training sets. Omnidata's dense prediction transformer hybrid was shown to be comparable to, or even outperform MiDaS. Despite the loss in measurement of exact depth, relative depth maps computed using pre-trained

monocular depth estimation networks have been shown to embed useful information for discriminating between indoor and outdoor images [23].

Image Datasets and Simulators: There are two methods for which training data containing depth information for indoor-outdoor classification networks is generally made available: data generation frameworks, and prepared datasets containing images and ground truth scans. Data generation frameworks create synthetic data on the fly from existing 3D scene descriptions. There are several data generation platforms for indoor 3D datasets, such as Habitat [30] and MINOS [29], which can utilize many 3D datasets, most notably SUNCG [36] and Matterport3D [4]. Scanned 3D scene datasets such as Matterport3D come with limitations in scanning accuracy and fixed scene lighting. Datasets built from computer graphics assets, such as Hypersim [27] on the other hand provide more flexibility for view-dependent lighting effects and scene adaptation.

Simulators, such as iGibson [35], AI2-THOR [18], and Kubrik [13] add support for physics interactions between agents and objects in the scene, but don't yet provide large-area outdoor realism. Driving simulators, such as CARLA [9] provide more meaningful realistic outdoor imagery, but mostly focus on urban environments from a driver's perspective. Combinations of synthetic data and 3D point clouds from aerial photography can also be used for believable outdoor scene image generation [6].

We experimented with the Grad-3D tool [2] developed with the Unity game engine, which allows for directing a virtual agent to navigate through an arbitrary Unity 3D scene, while capturing data during each movement. The tool outputs sequences of synthetic RGB images and ground-truth depth maps along flythrough paths. While data generation frameworks allow greater flexibility to users for controlling camera parameters and 3D scenes, the process of importing and setting up virtual datasets to work with these frameworks can be time-consuming.

On the other hand, there is a lack of publicly available outdoor datasets for the training and evaluation of models [44]. Outdoor datasets containing RGB images with their ground truth depth are even more rare. However, there are still several notable datasets which can be used for training. These include the KITTI dataset [12], Make3D dataset [31], DIODE dataset [41], and SUN397 dataset [42]. The KITTI dataset contains outdoor images and true depth maps taken from a moving vehicle. One set of images is for training/testing, with 23,488 pairs for training and 697 for testing. Another set of images is from a crash scene. The Make3D outdoor dataset contains 400 RGB image and depth map pairs for training, and 134 RGB image and depth map pairs for testing. Make3D images primarily depict cityscapes and nature taken during the daytime. DIODE contains 8,574 indoor training images and 16,884 outdoor training images along with their depth maps. Additionally, it contains 325 indoor and 446 outdoor validation images. This dataset contains scenes taken both during the daytime and night. Unlike the other datasets, SUN397 does not provide depth maps. However, it contains 108,754 images of both indoor and outdoor scenes.

3. Methods

Our proposed method for testing the usefulness of depth obtained from a pre-trained neural network utilizes Omnidata [10], a single-view depth prediction neural network trained on synthetic data. While a variety of pre-trained depth prediction networks are available, we choose Omnidata as we hope to study domain transfer from synthetic to real-life data, and Omnidata has shown results outperforming other state-of-art depth estimation networks. We use this pre-trained network to generate depth maps for the data and use the corresponding depth information as an additional channel when training our indoor-outdoor classifier. Lastly, we aim to understand how information learned from the synthetic domain transfers to real-life data.

3.1. Omnidata

The Omnidata depth prediction neural network generates depth maps for RGB images where the depth value for each pixel is relative and normalized 0 to 1. We apply Omnidata to SUN397 and DIODE; some representative RGB images from these datasets and their corresponding depth maps are shown in Figure 2 and Figure 3. Compared to ground truth depth maps, the generated Omnidata depth maps have a smaller range. Additionally, there is less color contrast and sharp edges, making it difficult to identify individual objects. However, the generated depth maps are a good approximation to the ground truth depth maps as there are significant similarities between them.

3.2. Problem Formulation

The binary classification task can be described as

$$y = ArgMax(Softmax(F(x, \omega))))$$

$\omega \in \mathbb{R}^N$ represents the trainable weights, $x \in \mathbb{R}^{H \times W \times D}$ are images from the datasets to be classified with or without depth information included,

$$F \colon \mathbb{R}^{H \times W \times D} \times \mathbb{R}^N \rightarrow R^2$$

is a neural network, and $y \in \{0, 1\}$, which represents the classes indoor and outdoor, respectively. If no depth information is included, $D = 3$, otherwise $D = 4$.

3.3. Binary Classification Methods

3.3.1 RGB Model

To understand the effect of adding depth information in training indoor/outdoor classification models, we first train with RGB images only ($D = 3$ channel dimensions), using ResNet50 described in Figure 1. This serves as the baseline model that estimates the performance of our network on a specific dataset.

3.3.2 RGB-D Model

We hypothesize that adding depth information generated from pre-trained depth-prediction networks trained on synthetic scenes will help improve indoor-outdoor classification accuracy on real data. We test our hypothesis by concatenating depth maps with the RGB images along the channel dimension for all the data ($D = 4$ channel dimensions), training and testing our model with 4 channel inputs. As we keep the network architecture, data transformation, and hyperparameters consistent, the RGB-D model can be compared to the RGB model to measure the effect of adding depth information for our task.

3.3.3 Network Architecture

We choose to use ResNet50, a well-known network that outperforms many other networks on image classification [15]. We follow the implementation from the original paper [15], using identical layers. The only difference is that our images are resized to $H = W = 250$ instead of 224. We find that resizing the images to 250 x 250 retains the maximum amount of information while minimizing the training time. In order to incorporate the depth information into our data, we added a depth channel to the RGB images. Lastly, the network outputs two classes, indoor and outdoor. The architecture of the network is shown in Figure 1.

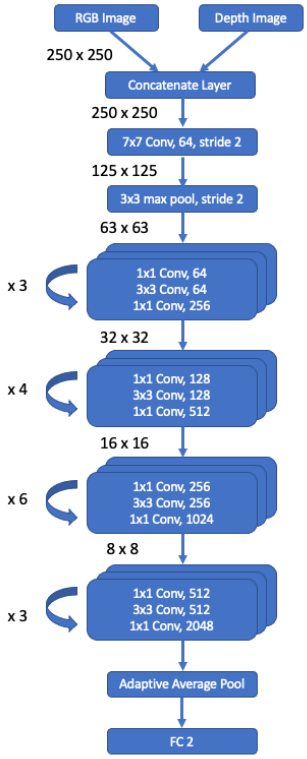

Figure 1: Network Architecture

4. Experiments

With the SUN397 and the DIODE datasets, we conduct two experiments to prove the effectiveness of our proposed method. Each experiment consists of training RGB models and RGB-D models, and the performance of the trained classifiers is analyzed using accuracy, precision, recall, and f1-score on the test set. In order to minimize the number of variables and to isolate the effect of adding depth information, all the training is done with the same network architecture described in Figure 1. Training hyper-parameters and image transformation are also identical for all experiments.

4.1. Datasets

We select the DIODE dataset and SUN397 dataset for the training and evaluation of our model. DIODE provides a large quantity of indoor and outdoor RGB images, as well as their corresponding depth maps and depth map masks. Other options we considered are KITTI and Make3D, but they both face limitations in the diversity of their scenes. KITTI contains images of mostly roads, with little variation in camera height and lighting. Similarly, Make3D only contains images taken during daytime. On the other hand, DIODE contains images taken during different seasons in both daytime and night, with scenes from several cities and various indoor/outdoor environments. Despite the lack of depth, the SUN397 dataset is also useful because of the sheer quantity of indoor and outdoor images which can be used for training. SUN397 is highly diverse, with a total of 397 scenes taken from indoor settings, outdoor natural settings, and outdoor man-made settings.

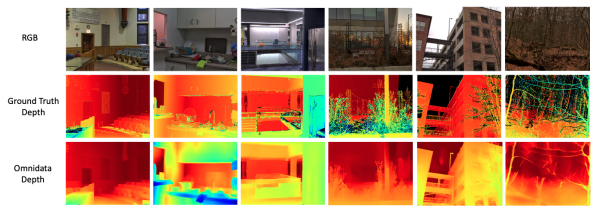

Figure 2: DIODE Dataset

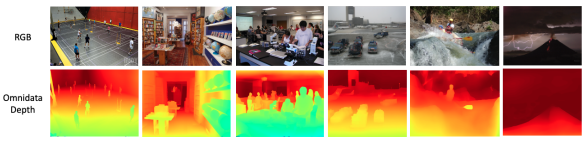

Figure 3: SUN397 Dataset

4.2. Experiment 1: Training on DIODE

The first experiment consists of applying our method of training RGB and RGB-D models on the DIODE dataset. The original partition of the data can be seen in Table 1.

	class	# of scenes	# of scans	# of images
train	indoors	7	80	8574
train	outdoor	12	100	16884
validation	indoors	3	10	325
validation	outdoor	3	10	446

Table 1: DIODE Dataset: Original Partition

Since the test set is not available, we use the entire validation set as the test set. Two scenes from the training set, one indoor and one outdoor, are set aside as the validation set. Our partition for the dataset can be seen in Table 2.

	class	# of scenes	# of scans	# of images
train	indoors	6	75	8108
train	outdoor	11	91	15253
validation	indoors	1	7	466
validation	outdoor	1	10	1631
test	indoors	3	10	325
test	outdoor	3	10	446

Table 2: DIODE Dataset: Our Partition

4.2.1 Using Omnidata Depth

The first part of the experiment studies the effect of adding Omnidata depth for indoor/outdoor classification of the DIODE dataset. Five RGB-D models are trained with corresponding Omnidata depth for each image. Similarly, five RGB models are trained. These models are tested on the test set, and we record the average metrics for RGB and RGB-D models in Table 4. We can see from all the metrics that the RGB-D models outperform the RGB models as all the metrics are significantly greater, especially in the 4.4% improvement in accuracy.

4.2.2 Ablation Study

Since the DIODE dataset contains ground truth depth, we perform an ablation study, where we train RGB-D models with the provided depth.

We quantify the usefulness of depth information provided by the pre-trained Omnidata network by comparing the classification accuracy against models trained with perfect depth information and no depth information. Five models are trained under the same configurations using the ground truth depth. Each model is then tested on the test set, and the results are averaged and recorded in Table 4. We can see that by removing ground truth depth, we lose 6.9% in accuracy and 5.9% in F1-score. More importantly, comparing the metrics between RGB-D models trained with ground truth depth to the ones trained with Omnidata depth show that using the generated depth retains 63.8% of the performance boost in accuracy achieved with ground truth depth. This is an indicator that the information learned by the pre-trained Omnidata network on synthetic scenes can be transferred to real-life data.

4.3. Experiment 2: Training On SUN397

With the experiment on the DIODE dataset as the controlled experiment, we then test the effect of adding Omnidata depth to the SUN397 dataset for indoor/outdoor classification. We hypothesize that similar to experiment 1, adding Omnidata depth will boost the performance of the classifier. However, since this dataset does not have ground truth depth, we cannot study the effect of how much of the missing depth information can be filled in with Omnidata.

SUN397 contains 108,754 images, 47,549 indoor, and 61,205 outdoor. Table 3 shows our partition for the dataset.

	class	# of images
train	indoors	32549
train	outdoor	46205
validation	indoors	5000
validation	outdoor	5000
test	indoors	10000
test	outdoor	10000

Table 3: SUN397 Dataset Partition

Similar to experiment 1, we train 5 RGB models and 5 RGB-D models with Omnidata depth. The models are tested on the test sets, and the average results are recorded in Table 4. We can see that the RGB-D model achieves a 1.3% performance boost in accuracy compared to the RGB model, which is significant and further proves the benefit of using depth from networks pre-trained on synthetic data.

	Acc	Prec	Recall	F1-score
DIODE				
RGB	0.873	0.875	0.913	0.893
RGB-D with Omnidata depth	**0.917**	**0.885**	**0.979**	**0.929**
RGB-D with Ground Truth depth	0.942	0.927	0.978	0.952
SUN397				
RGB	0.919	0.906	0.956	0.930
RGB-D with Omnidata depth	**0.932**	**0.910**	**0.959**	**0.933**
RGB-D with Ground Truth depth	N/A	N/A	N/A	N/A

Table 4: Testing Results

4.4. Image Transformation

The datasets contain images of different sizes. We normalize all the RGB images by the mean and standard deviations of the dataset. Additionally, both the RGB images and the depth maps are resized to 250x250, using the nearest-neighbor interpolation.

4.5. Training Hyper-parameters

Each model is set to train for 80 epochs on a single RTX 3090 and evaluated using the validation set (partitions shown in Table 2 and Table 3) at the end of each epoch. The model with the highest accuracy on the validation set is saved, and the saved model is tested on the test set. We use Cross-Entropy Loss, and the Adam Optimization algorithm to update the weights. We also use a batch size of 128, and a learning rate of 0.001.

4.6. Hardware and Training Time

For the DIODE experiment, each epoch trains for about 350 ± 10 seconds on the RTX 3090, for both with and without depth. With the batch size of 128, about 18 GB, out of 24 GB, of the VRAM is used. SUN397 uses the same amount of VRAM as we keep the batch size consistent. This dataset takes about 900 ± 10 seconds to train on the same GPU. The longer training time comes from the fact that SUN397 is a much bigger dataset.

5. Discussion

While using Omnidata depth in training DIODE is 2.5% less accurate compared to using the ground truth depth, it only takes around 5 minutes (measured on an NVidia RTX 3090 card) to generate depth maps for the entire dataset. The time it takes to train Omnidata can be considered insignificant as it is a one-time process. The ground truth depth for DIODE is collected using sensors that take 11 minutes to complete a single scan [41]. This means that it takes about 36 hours to collect data for the whole 200 scans provided. Using Omnidata to generate the depth takes 99.8% less time compared to using the sensors. Overall, using generated depth from Omnidata retains 63.8% of the performance in terms of accuracy gained by using

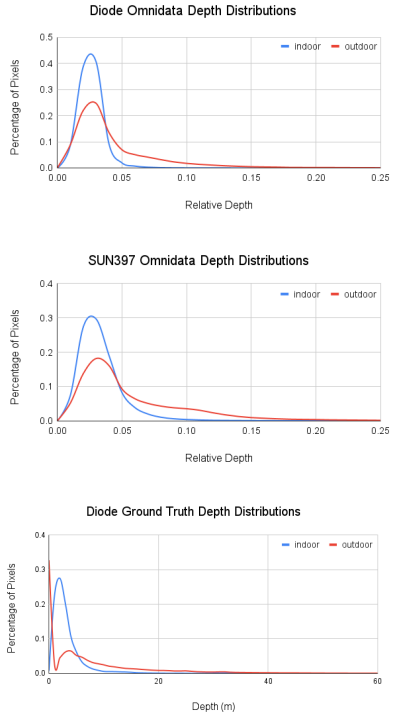

Figure 4: Depth Distributions

the ground truth depth. Specifically, Omnidata boosts the accuracy of the model by 4.4% on DIODE and 1.3% on SUN397. As there is no significant difference in training time between Omnidata and ground truth, the advantages of using a pre-trained network are obvious.

5.1. Depth Distributions

Intuitively, we expect a difference in the depth distributions of indoor and outdoor scenes. Outdoor scenes usually have a larger range of depth, as they are likely to contain objects very far away. Indoor scenes are generally limited by walls and ceilings and as such would have a smaller range in depth. In Figure 4, we show the average depth distributions of all the indoor and outdoor images within the DIODE and SUN397 datasets. We observe that outdoor image depth tends to follow the shape of a long-tailed distribution. On the other hand, the average indoor image depth distribution has a large peak within the smaller ranges, before quickly dropping off. Compared to indoors, the outdoor depth distribution has a smaller peak. As a result, outdoor scene depth distributions are found to have a larger standard deviation than indoor scene depth distributions, and this distinction may be a helpful cue for training indoor/outdoor binary classifiers. These three diagrams show that this trend holds regardless of using pseudo- or ground truth depth, which indicates that information from a network pre-trained on syn-

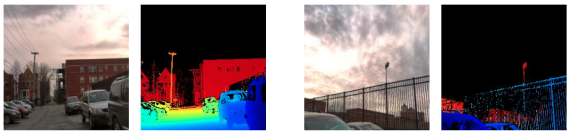

Figure 5: Outdoor Images with High Percentage of Pixels with Invalid Depth (Black Regions)

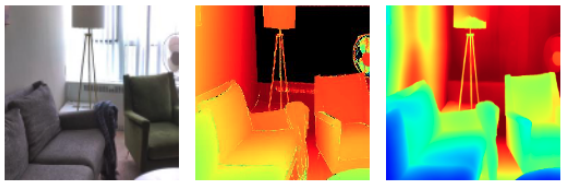

Figure 6: **Left:** RGB; **Center:** Ground Truth Depth; **Right:** Omnidata Depth

thetic scenes generalizes to real data.

Although all three figures follow the same trend, there is still a distinct difference between the Omnidata depth distribution and the ground truth depth distribution for DIODE. The ground truth depth information is provided in the form of validity masks and depth maps, where the validity mask indicates whether the depth at each pixel is valid. The outdoor ground truth distribution has a peak at the depth of zero, meaning there is a high percentage of pixels with invalid depth. It is difficult for physical sensors to detect extreme distances such as the sky. We verify that images with sky yield a high percentage of invalid depth, which acts as a strong cue for classifying an image as an outdoor scene. Figure 5 shows two example images where there is a high percentage of sky and are classified correctly as outdoor scenes. However, using invalid pixels as cues for classification is dangerous as it is biased towards outdoor scenes. Certain indoor images will be incorrectly classified as sensors cannot detect the depth of objects such as mirrors and windows. Figure 6 shows an incorrectly classified indoor image where the window region has invalid ground truth depth, and we can see that the Omnidata depth map is a better representation of the scene. Despite this, our experiments show that RGB-D models trained with the ground truth depth still outperform the ones trained with Omnidata depth. A possible explanation is that there are more outdoor than indoor images in both of the datasets, and models biased towards outdoor scenes will have a higher accuracy overall.

5.2. Failure Cases

On indoor scenes, specifically in the DIODE test set, the RGB-D model trained with Omnidata depth performs approximately 4.1% worse than the one trained on ground truth depth. Specifically, the accuracy of the Omnidata depth model and ground truth depth model is at 81.8% and 85.9%, respectively. Figure 7 displays two indoor images in which our Omnidata-trained model misclassified while the ground truth-trained model made the correct prediction. Comparing the corresponding depth distributions, we notice that the Omnidata pseudo-depth distribution spans a larger range of normalized depth values than the ground truth. This is more similar to the longer-tail average *outdoor* depth distribution observable in Figure 4, partially explaining the classification difference. Regarding outdoor data, classification differences could also be linked to Omnidata's difficulties with capturing the correct depth for vegetation, as exemplified in columns 4,5, and 6 of Figure 2.

In general, we find that classification using Omnidata tends to have difficulty with indoor images that feature larger ranges of depths, such as underground subways, staircases, and hallways. In these cases, the percentage of pixels would be more widely spread along a larger depth range. This results in a longer-tailed depth distribution that resembles an outdoor scene more than an indoor scene.

On the DIODE outdoor scenes, our RGB-D model trained with Omnidata is 0.25% less accurate than the one trained on ground truth depth. We notice that Omnidata generally fails when all objects are relatively the same distance away from the camera, in which a high percentage of pixels would be squeezed to a limited depth range. In this case, the depth distribution would not display the traditional long-tailed shape of outdoor scenes and instead would feature a single large peak of a typical indoor scene. Depth generated by Omnidata is only an approximation, and it is not accurate enough for these cases.

5.3. Edge Cases

Within the SUN397 dataset there are many edge-case images, as seen in Figure 8. Images that were taken underwater are very difficult for our model to handle. Further analysis shows that depth maps produced by Omnidata seem to treat the water as one object, leading to a consolidation of pixels in limited depth range. This indicates an indoor classification. Within the data set, we also noticed mixed-environment images that combined both indoor architecture and distinct outdoor features such as the sky. This conflict is represented in the depth distribution, in which there are rising peaks over a large range of depths. This distribution is unlike the typical outdoor or indoor depth distribution, and classification is therefore challenging. It is likely that Omnidata fails to predict accurate depth maps for images like these because it has not seen them during training.

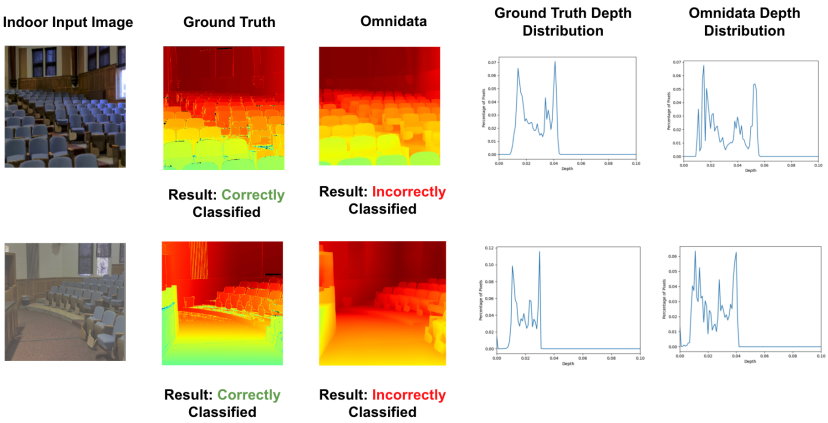

Figure 7: Omnidata Versus Ground Truth Indoor Classifications, Normalized Depth Distribution Cropped to [0-0.1]

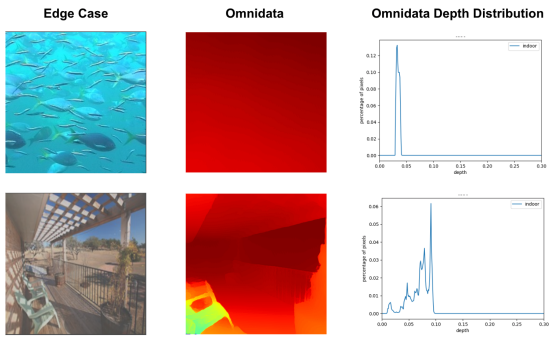

Figure 8: SUN397 Edge Cases

6. Conclusions and Future Work

In this work, we investigate the usefulness of pre-trained depth prediction neural networks in scene classification by using predicted depth as an additional cue when training indoor-outdoor classifiers. Our method tackles the problem of the lack of diverse, expansive indoor-outdoor datasets containing ground truth depth and physical limitations with existing depth sensors. Specifically, we combat this issue by utilizing a pre-trained, single-view depth estimation network in order to generate predicted relative depth. Our results with DIODE show that using synthetically generated depth maps as an additional input for training a ResNet50 CNN achieves a 4.4% performance boost over the non-depth baseline model, retaining 63.8% of the performance boost achieved from training a classifier on RGB and ground truth depth. For the SUN397 dataset, RGB-D trained with Omnidata depth similarly showed a 1.3% performance boost over the non-depth baseline. Although pseudo-depth obtained from Omnidata contains less information overall than ground truth depth, we find evidence that using relative depth maps provided by pre-trained networks still provides a strong cue for indoor-outdoor classification.

We show that adding Omnidata depth improves the performance of our network due to the differences in depth distributions between indoor and outdoor images. Depth distributions for outdoor images tend to have a long tail, while indoor depth distributions have a higher peak. This distinct difference is consistent for both ground truth depth and Omnidata-generated depth across the SUN397 and DIODE datasets. This further demonstrates how pre-trained networks such as Omnidata can extend their applicability to real-life data.

Additionally, despite losing around 2.5% in accuracy compared to using the ground truth depth, using depth generated from pre-trained models is still the more viable option. Ground truth depth for DIODE took 36 hours to collect, while Omnidata generated depth maps for the entire dataset in five minutes — using a pre-trained depth-prediction network is 99.8% faster. Moreover, the entire process does not require any expensive sensors or cameras.

Since our experiments show that depth information transfers from the synthetic domain to real-life data, possible future work includes investigating what other synthetically generated knowledge can be transferred using pre-trained models. Omnidata also provides a surface normal estimation neural network, so we can perform an experiment using surface normals as an additional cue for indoor-outdoor classification. The DIODE dataset also contains ground truth surface normal, so a similar experiment could be conducted to learn about what kind of information is useful and transferable from synthetic data. Additionally, we expect that other image classification tasks would also benefit from additional information provided by networks pre-trained on synthetic scenes. These tests would further explore the limits of information transfer from the synthetic domain to real-life data.

Acknowledgments

This work was supported in part by the Office of Naval Research, under grants N00014-19-1-2553, N00174-19-1-0024, and N00014-20-1-2719, as well as NSF awards IIS-2211784 and IIS-1911230.

References

[1] Supreeth Achar, Joseph R. Bartels, William L. 'Red' Whittaker, Kiriakos N. Kutulakos, and Srinivasa G. Narasimhan. Epipolar time-of-flight imaging. *ACM Trans. Graph.*, 36(4), jul 2017.

[2] Pranav Acharya, Daniel Lohn, Vivian Ross, Maya Ha, Alexander Rich, Ehsan Sayyad, and Tobias Hollerer. Using synthetic data generation to probe multi-view stereo networks. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*.

[3] Simone Bianco, Gianluigi Ciocca, Claudio Cusano, and Raimondo Schettini. Improving color constancy using indoor–outdoor image classification. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 17:2381–92, 01 2009.

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[5] Hui Chen, Yan Feng, Jian Yang, and Chenggang Cui. 3d reconstruction approach for outdoor scene based on multiple point cloud fusion. *Journal of the Indian Society of Remote Sensing*, 47:1761 – 1772, 2019.

[6] Meida Chen, Qingyong Hu, Thomas Hugues, Andrew Feng, Yu Hou, Kyle McCullough, and Lucio Soibelman. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv preprint arXiv:2203.09065*, 2022.

[7] Jack Collier and Alejandro Ramirez-Serrano. Environment classification for indoor/outdoor robotic mapping. CRV '09, page 276–283, USA, 2009. IEEE Computer Society.

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

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

[10] Ainaz Eftekhar, Alexander Sax, Roman Bachmann, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans, 2021.

[11] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[12] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[13] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–3761, 2022.

[14] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *Proceedings - 2019 International Conference on Computer Vision, ICCV 2019*, Proceedings of the IEEE International Conference on Computer Vision, pages 1506–1516. Institute of Electrical and Electronics Engineers Inc., Oct. 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[16] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *UIST '11 Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568. ACM, October 2011.

[17] Wonjun Kim, Jimin Park, and Changick Kim. A novel method for efficient indoor—outdoor image classification. *J. Signal Process. Syst.*, 61(3):251–258, dec 2010.

[18] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[19] Ziquan Lan, Zi Jian Yew, and Gim Lee. Robust point cloud based reconstruction of large-scale outdoor scenes. 04 2019.

[20] Yang Liu and Xueqing Li. Indoor-outdoor image classification using mid-level cues. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–5, 2013.

[21] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[22] Andrew Payne and Sameer Singh. Indoor vs. outdoor scene classification in digital photographs. *Pattern Recogn.*, 38(10):1533–1545, oct 2005.

[23] Ignazio Pillai, Riccardo Satta, Giorgio Fumera, and Fabio Roli. Exploiting depth information for indoor-outdoor scene classification. In Giuseppe Maino and Gian Luca Foresti, editors, *Image Analysis and Processing – ICIAP 2011*, pages 130–139, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[24] Karthik Pujar, Satyadhyan Chickerur, and Mahesh S. Patil. Combining rgb and depth images for indoor scene classification using deep learning. In *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–8, 2017.

[25] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular

depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2019.

[26] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[27] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10912–10922, 2021.

[28] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017.

[30] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[31] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Depth perception from a single still image. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1571–1576. AAAI Press, 2008.

[32] Raimondo Schettini, Carla Brambilla, Claudio Cusano, and Gianluigi Ciocca. Automatic classification of digital photographs based on decision forest. *International Journal of Pattern Recognition and Artificial Intelligence*, 18:819–845, 08 2004.

[33] N. Serrano, A. Savakis, and A. Luo. A computationally efficient approach to indoor/outdoor scene classification. In *2002 International Conference on Pattern Recognition*, volume 4, pages 146–149 vol.4, 2002.

[34] Navid Serrano, Andreas E. Savakis, and Jiebo Luo. Improved scene classification using efficient low-level features and semantic cues. *Pattern Recognition*, 37(9):1773–1784, 2004.

[35] Bokui Shen*, Fei Xia*, Chengshu Li*, Roberto Martín-Martín*, Linxi Fan, Guanzhi Wang, Shyamal Buch, Claudia D'Arpino, Sanjana Srivastava, Lyne P Tchapmi, Kent Vainio, Li Fei-Fei, and Silvio Savarese. igibson, a simulation environment for interactive tasks in large realistic scenes. *arXiv preprint arXiv:2012.02924*, 2020.

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

[37] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] Waleed Tahir, Aamir Majeed, and Tauseef Rehman. Indoor/outdoor image classification using gist image features and neural network classifiers. In *2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET)*, pages 1–5, 2015.

[39] Li Tao, Yeong-Hwa Kim, and Yeong-Taeg Kim. An efficient neural network based indoor-outdoor scene classification algorithm. In *2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE)*, pages 317–318, 2010.

[40] Aditya Vailaya, Mário Figueiredo, and Anil Jain. Image classification for content-based indexing. ieee trans image process. *Image Processing, IEEE Transactions on*, 10:117 – 130, 02 2001.

[41] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. Diode: A dense indoor and outdoor depth dataset, 2019.

[42] Jianxiong Xiao, Krista Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119, 08 2014.

[43] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, 2021.

[44] Keyang Zhou, Kaiwei Wang, and Kailun Yang. Padenet: An efficient and robust panoramic monocular depth estimation network for outdoor scenes. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, 2020.