MONOCULAR DEPTH PREDICTION IN PHOTOGRAMMETRIC APPLICATIONS

M. Welponer, E. K. Stathopoulou*, F. Remondino

3D Optical Metrology (3DOM) Unit, Bruno Kessler Foundation (FBK), Trento, Italy,
Web: http://3dom.fbk.eu – Email: (welponer, estathopoulou, remondino)@fbk.eu

Commission II - WGII/4

KEYWORDS: monocular, depth prediction, 3D reconstruction, CNN, deep learning, photogrammetry

ABSTRACT

Despite the recent success of learning-based monocular depth estimation algorithms and the release of large-scale datasets for training, the methods are limited to depth map prediction and still struggle to yield reliable results in the 3D space without additional scene cues. Indeed, although state-of-the-art approaches produce quality depth maps, they generally fail to recover the 3D structure of the scene robustly. This work explores supervised CNN architectures for monocular depth estimation and evaluates their potential in 3D reconstruction. Since most available datasets for training are not designed toward this goal and are limited to specific indoor scenarios, a new metric, large-scale synthetic benchmark (ArchDepth) is introduced that renders near real-world scenarios of outdoor scenes. A encoder-decoder architecture is used for training, and the generalization of the approach is evaluated via depth inference in unseen views in synthetic and real-world scenarios. The depth map predictions are also projected in the 3D space using a separate module. Results are qualitatively and quantitatively evaluated and compared with state-of-the-art algorithms for single image 3D scene recovery.

1. INTRODUCTION

Depth estimation from 2D images is a fundamental research topic in photogrammetry and computer vision toward 3D reconstruction and scene understanding with a vast field of applications, including mapping, navigation, and augmented reality. Most scenarios have high requirements for dense and accurate depth estimation for, if possible, every scene pixel to recover the 3D structure reliably. The recent success of deep learning in several image recognition tasks, such as image classification (Krizhevsky et al., 2012; He et al., 2016), object detection (Girshick et al., 2014; He et al., 2017), and semantic segmentation (Long et al., 2015; Chen et al., 2017; Badrinarayanan et al., 2017), motivated their application also in the field of depth estimation and 3D reconstruction, especially for tackling the matching ambiguities and occlusions problem. Depth estimation using deep learning can be applied in stereo, multi-view, or monocular scenarios. Indeed, various supervised or unsupervised methods have been suggested in the literature in recent years (Zbontar and Lecun, 2015; Yao et al., 2018; Huang et al., 2021).

In particular, monocular depth estimation methods aim to recover distances between scene objects and camera parameters from a single image. It is, by definition, an ill-posed problem since redundant 3D scenes can be projected to the same 2D image. Indeed, an efficient depth map recovering from a single image would require rich scene prior cues, commonly used in conventional methods (Saxena et al., 2008). In the deep learning era, monocular depth estimation refers to the task of single image inference during test time, first introduced by Eigen et al. (2014) using a coarse-to-fine approach. Since then, the problem has been broadly studied in the literature as a supervised or unsupervised task. As with all supervised learning methods, supervised monocular depth estimation relies on corresponding ground truth (GT) depth maps for every RGB image. On the contrary, unsupervised methods learn stereo cues or video sequences during training and predict a depth map for single images during testing. Despite the tremendous underlying potential, supervised depth estimation generally requires an enormous amount of training data to generalize in diverse scenarios properly, i.e., indoor, outdoor, and aerial applications; this fact is particularly true in monocular depth estimation. Most state-of-the-art methods achieve their results by training and testing on each benchmark separately; few focus on generalization, commonly assuming ordinal depth relations and only recently investigating affine-invariant depth (Yin et al., 2020). However, we believe that the greatest challenge of monocular depth estimation is the quality of the 3D reconstruction derived from the predicted estimates. The deficiency derives commonly due to the lack of 3D supervision cues and the difficulty in determining the camera intrinsics. Indeed, it is not trivial to enforce geometric constraints from monocular images without additional scene cues. In fact, most methods are limited to depth prediction, and while achieving low depth error values, the actual 3D scene reconstruction mostly fails; 3D structure recovery remains an unexplored topic for state-of-the-art methods. Few recent works discuss this issue, integrating geometric supervision (Yin et al., 2019) or relying on extra modules for training in point cloud level separately from depth estimation (Yin et al., 2021).

The transferability of deep learning depth estimation for real-world photogrammetric scenarios is a challenging problem that has only recently been acknowledged in the community (Madhuanand et al., 2021; Steenbeek and Nex, 2022).

1.1 Aim of the work

This work investigates the potential of integrating learning-based monocular depth estimation in photogrammetric applications. Our contributions can be summarized as follows (Figure 1):

(1) we introduce a novel, large-scale dataset (ArchDepth) of photorealistic outdoor scenes of historic buildings, including high-quality, complete, metric depth maps for every image; (2) we present a straightforward training pipeline following an encoder-decoder network for metric monocular depth estimation to demonstrate the potential of this dataset; (3) we employ a 3D reconstruction module based on our predictions for single-view 3D scene recovery; (4) we evaluate the generalization performance of our trained model and investigate its applicability in real-world photogrammetric scenarios.

* Corresponding author

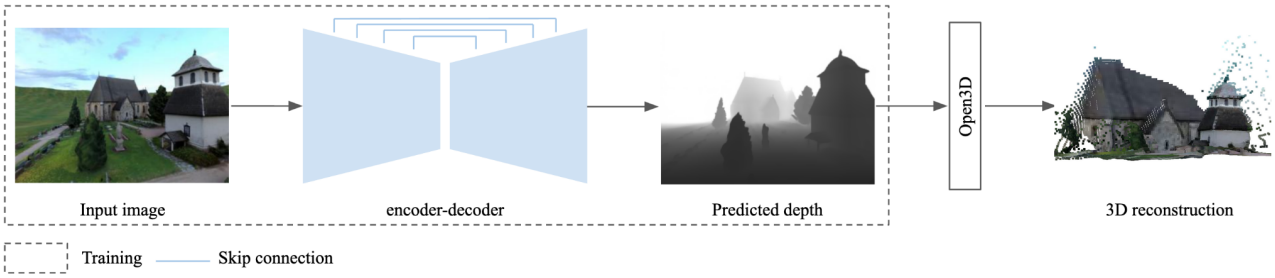

Figure 1. The pipeline of our method is based on an encoder-decoder architecture with skip connections for monocular depth prediction. An additional module for 3D reconstruction is also employed afterward.

2. RELATED WORK

Monocular Depth Prediction. Early methods for monocular depth estimation relied on handcrafted features and used complementary cues to recover the depth since limited information about the scene geometry can be directly extracted from a single image (Saxena et al., 2008). In the deep learning era, the seminal work of Eigen et al. (2014) proposed a scale-invariant loss function in a coarse-to-fine context using a VGG network. The approach was further extended by adding more layers while predicting surface normals and semantic maps (Eigen and Fergus, 2015). Since then, the problem has been studied in the literature as a supervised (Laina et al., 2016; Xu et al., 2018; Fu et al., 2018; Hu et al., 2019) or unsupervised problem (Garg et al., 2016; Godard et al., 2017; Tosi et al., 2019). An architecture often adopted in such methods is the encoder-decoder (e.g., Fu et al., 2018) with RGB images as input and direct regression of pixel-wise depth maps as output. Indeed, most methods perform pixel-wise supervision, yet Conditional Random Fields (CRFs) have also been used to exploit neighbor relations and include a more global context (Liu et al., 2015). The loss function can be formed either as a regression or a classification problem. Skip connections in a ResNet fashion are used to preserve the fine-grained features of the first layers (Laina et al., 2016). Cues such as texture, shading, and structural information are used, while high quality and pixel-aligned GT depth maps are needed. Depending on the available training data, the scene depth can be estimated as ordinal, i.e., relative (Fu et al., 2018) or Euclidean (Eigen et al., 2014; Yin et al., 2019). Local planar priors have also been incorporated as guidance (Lee et al., 2019). Apart from standard CNN models, adversarial training (Chen et al., 2018), attention mechanisms (Chen et al., 2020), and transformer architectures (Ranftl et al., 2021; Yang et al., 2021) have also been recently proposed.

3D scene recovery. Even though achieving excellent results in depth map prediction (e.g., Hu et al., 2019), the respective reconstructions in the 3D space suffer from significant distortions and the presence of artifacts. Only recently, few works have tried to incorporate 3D awareness into the methods. Since most man-made scenes can be decomposed in planar structures, plane detection can be used as a prior for monocular depth estimation (Liu et al., 2019). However, the 3D structure was not explicitly considered until recently; Yin et al. (2019) formulated a joint loss function using virtual normals to enforce high-order geometric consistency between surface patches in a large range. The work was further extended by considering affine-invariant depth (Yin et al., 2020) and adding an extra training module for scene 3D reconstruction (Yin et al., 2021). These state-of-the-art methods, although promising, still suffer from generalization limitations in diverse scenarios.

Datasets. *KITTY Vision* (Geiger et al., 2012) and *NYU Depth v2* (Silberman et al., 2012) are the pioneer efforts and widely-used large-scale datasets regarding the number of images. *KITTY Vision* contains real-world street scenes captured with a LiDaR sensor, while *NYU Depth v2* contains indoor scenes acquired with the Kinect sensor. Indeed, most existing benchmark datasets for depth estimation are video sequences of indoor scenes acquired with such commodity RGB-D sensors. Since then, the increasing demand for training data has led to the release of similar datasets *SUN RGB-D* (Song et al., 2015) and larger-scale ones regarding scene diversity and acquired images such as *Stanford 2D-3D Semantics* (Armeni et al., 2017), *ScanNet* (Dai et al., 2017) and the synthetic *SceneNet* (McCormac et al., 2017). The aforementioned benchmark datasets have established a common baseline for new algorithms to be developed and evaluated. They have contributed significantly to developing new methods and have driven the research in novel directions during the last decade. However, although the scenarios contain a vast number of images, they are mostly similar; that is, constrained by the usage of depth sensors, they are limited to indoor environments. Moreover, depth sensors inevitably introduce errors during acquisition, resulting in noisy training data. To improve the generalization of such methods in random scenes, datasets with crowdsource images from the internet have also been introduced (Li and Snavely, 2018). Yet, they solve the depth estimation only at the ordinal level, prohibiting distortion-free and metric 3D reconstructions. To overcome such limitations, in this paper, we propose a novel, metric, large-scale dataset containing outdoor scenes of historic buildings of varying architectural styles (Figure 2). We aspire that this dataset will enable further research in the field.

3. METHODOLOGY

Most state-of-the-art networks for monocular depth estimation focus on indoor datasets and typically fail to generalize in outdoor, real-world scenarios. Therefore, we introduce a new metric dataset of photorealistic environments. We employ an encoder-decoder architecture for training and an additional single-view 3D reconstruction module to prove its effectiveness.

3.1 The ArchDepth dataset

We introduce a novel dataset, named ArchDepth, consisting of seven photorealistic outdoor scenes of historic buildings of diverse architectural styles (Figure 2). The first six scenes are 3D models of northern European medieval churches retrieved from the web[1], namely *Kuusisto, Liedon, Mietoinen, Nousiainen, Piikkio*, and *Saint Jacobs*. The last scene includes various similarly harvested historic facades rendered in a *virtual Piazza*.

[1] https://sketchfab.com/3d-models

Figure 2. Our synthetic dataset ArchDepth. The first six scenes depict 3D models of churches, while the last scene includes several historic facades. The diverse camera paths for each scene are indicated with a black line.

We have built upon the open-source software Blender[2] for image rendering. For the first six scenes, we designed four camera paths around each of our models and five paths for the *Liedon* model, rendering a total of 24,000 images of 640x480 resolution. The *virtual Piazza* consists of eight camera paths along the facades. Images are generated based on the pinhole camera model, so no distortions were present.

Moreover, we also generated a hybrid dataset *Modena Cathedral*; it contains 88 real-world images acquired for photogrammetric 3D reconstruction. A point cloud collected with a commercial laser scanner was used as a ground truth model. However, since the images also contained areas not acquired with the scanner (due to occlusions, sensor range limits, etc.) or appeared with sparse points, starting from the acquired point cloud, we have generated a 3D model to render complete depth maps.

The generation of a new dataset of outdoor scenes for training purposes was undoubtedly a laborious and expensive task, yet we believe it can be a starting point for further research.

3.2. Network architecture and training

We employ a straightforward encoder-decoder architecture with skip connections based on the network of Alhashim and Wonka (2018). The network has ca. 58M parameters and has been proven to work efficiently and produce high-quality depth maps with clear boundaries. The original method exploits transfer learning, starting from pre-trained weights on other visual recognition tasks, i.e., image classification. Following this idea, we use a pre-trained DenseNet-201 (Huang et al., 2017) on ImageNet (Deng et al., 2009; Krizhevsky et al., 2012) as a backbone. The decoder part consists of successive bilinear up-sampling layers and their skip connections, and the output is half the input resolution (320*240). As in the original implementation (Alhashim and Wonka, 2018), we do not perform batch normalization (Ioffe and Szegedy, 2015).

3.2.1 Loss function. The choice of the loss function is crucial and should be appropriate for the particular problem. For depth regression, a standard approach considers the pixel-wise depth difference between GT depth value y and prediction y^* (Eigen et al., 2014). Following the approach of Alhashim and Wonka (2018), apart from the pixel-wise loss L_{depth}, we use the loss over the image gradient L_{grad} and the loss based on structural similarity L_{SSIM} as defined by Wang et al. (2004). The final total loss is therefore defined as a weighted sum:

$$L_{total} = w_1 L_{depth} + w_2 L_{grad} + w_3 L_{SSIM} \quad (1)$$

For more details on the loss function, we refer the reader to (Alhashim and Wonka, 2018).

3.2.2 Data augmentation. Data augmentation is proven to be beneficial to reduce overfitting, especially when limited data are

[2] https://www.blender.org/

	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$absrel \downarrow$	RMSE↓	$\log_{10} error \downarrow$
experiment 1 - ours	0.981	0.994	0.997	0.036	8.82	0.016
experiment 2 - ours	0.973	0.991	0.994	0.078	5.77	0.033
experiment 1 - LeRes	0.240	0.457	0.575	0.749	53.51	-
experiment 2 - LeRes	0.126	0.250	0.334	3.217	77.51	-

Table 1. 2D metric for experiments 1 and 2. The first two rows refer to depth map inference using our training model, while the last two refer to depth maps predicted using the trained model of Yin et al. (2021).

available (Krizhevsky et al., 2012). In the particular scenario of depth estimation, certain geometric transformations may not be appropriate or meaningful. In this case, we only consider mirroring, while for radiometric transformations, we consider color channel permutations following (Alhashim and Wonka, 2018).

3.2.3 Experimental setup. For the training of the original network, subsets of the *NYU Depth v2* (Silberman et al., 2012) and the *KITTY Vision* (Geiger et al., 2012) datasets were used. We consider the synthetic dataset and the hybrid dataset *Modena Cathedral* under two different experiments.

Experiment 1 - Synthetic dataset. We split the synthetic dataset based on a random shuffling approach on all the seven scenes by keeping 88% for training, 6% for validation, and the rest is kept for testing.

Experiment 2 - Model fine-tuned on Modena Cathedral. We test on the *Modena Cathedral* dataset to demonstrate how well such an architecture, trained on our synthetic dataset, generalizes in other scenes. For the fine-tuning, we use 68 images for training, 42 for validation, and 19 for testing.

3.2.4 Implementation details. In our implementation, we use the open-source TensorFlow[3] library (version 2.3.1) and trained on an NVIDIA GeForce RTX 2070 with 8G RAM. Regarding the network hyperparameters, we use the Adam optimizer with a learning rate $a = 0.0001$ and a decay factor of 0.7. Training is performed for 74 epochs for Experiment 1 using early stopping and 48 epochs for Experiment 2.

3.3 3D reconstruction

In photogrammetric applications, which commonly have high-quality requirements, 3D reconstruction is typically achieved using stereo and multi-view methods. However, monocular estimation can be potentially helpful in cases of low overlap percentage between the images. Given the recent advancements in the state-of-the-art, in this study, we investigate the potential of reliable 3D scene recovery from a single image. We reconstruct the 3D scene explicitly by projecting each depth value to the 3D space using the camera projection matrix. For this module, we use the open-source library Open3D (Zhou et al., 2018). The input is the RGB image along with its respective 16-bit depth maps; a pinhole camera model is adopted.

4. EVALUATION

4.1 Evaluation Metrics

4.1.1. 2D metrics. A standard set of metrics typically used in the literature since the seminal work of Eigen et al. (2014) is adopted to evaluate the quality of the depth maps, namely absolute relative error, root mean square error, logged root mean square error, and accuracy under threshold. These metrics are calculated by comparing all pixel predictions y^* with their GT equivalents y. Predicted depth maps were upscaled to the original resolution (640x480) using bilinear sampling.

$$absrel = \frac{1}{n}\sum \frac{|y - y^*|}{y} \qquad (2)$$

$$RMSE = \sqrt{\frac{\sum (y - y^*)^2}{n}} \qquad (3)$$

$$\log_{10} error = \frac{1}{n}\sum |\log_{10}(y) - \log_{10}(y^*)| \qquad (4)$$

Where n refers to the total number of image pixels. The accuracy under threshold δ is calculated as the percentage of pixels below a threshold with the threshold being $th = \{1.25, 1.25^2, 1.25^3\}$.
The average results for all testing images of experiments 1 and 2 are shown in Table 1. Naturally, in experiment 1 where both training and tests sets come from the same dataset, the results are the best; however, the fine-tuned model generalizes quite well on the challenging real-world scenario. Our intuition is that the high RMSE values are due to the presence of some outliers.
In Figures 3 and 4 various examples of the predicted depth maps are shown. In Experiment 1 the predicted depth maps with our model are close to the ground truth ones; depth transitions are smooth and ordinal values seem to be consistent. In Experiment 2 the results behave similarly; however, some outliers are more likely to be present.

4.1.2 3D metrics. The 2D metrics tend to disregard the overall predicted 3D structure of the scene and cannot thus be reliable regarding the quality of the resulting 3D model. Although demonstrating high scores, most state-of-the-art methods fail to reliably reconstruct the 3D structure of the scene (Yin et al., 2019; 2020; 2021). To investigate this deficiency, in this study, apart from the standard 2D scores, we also consider the commonly used metrics for 3D point cloud quality, completeness (recall), accuracy (precision), and their harmonic mean F1-score (Knapitsch et al., 2017). Figures 5 and 7 present some indicative results of the 3D reconstructed point clouds, and Figures 6 and 8 their respective scores.
Although for some test images the 3D point cloud was reconstructed successfully with marginal distortions, in some other cases, particularly under strong perspective angles, the resulting 3D model is not reliably recovered.

[3] https://github.com/tensorflow/tensorflow

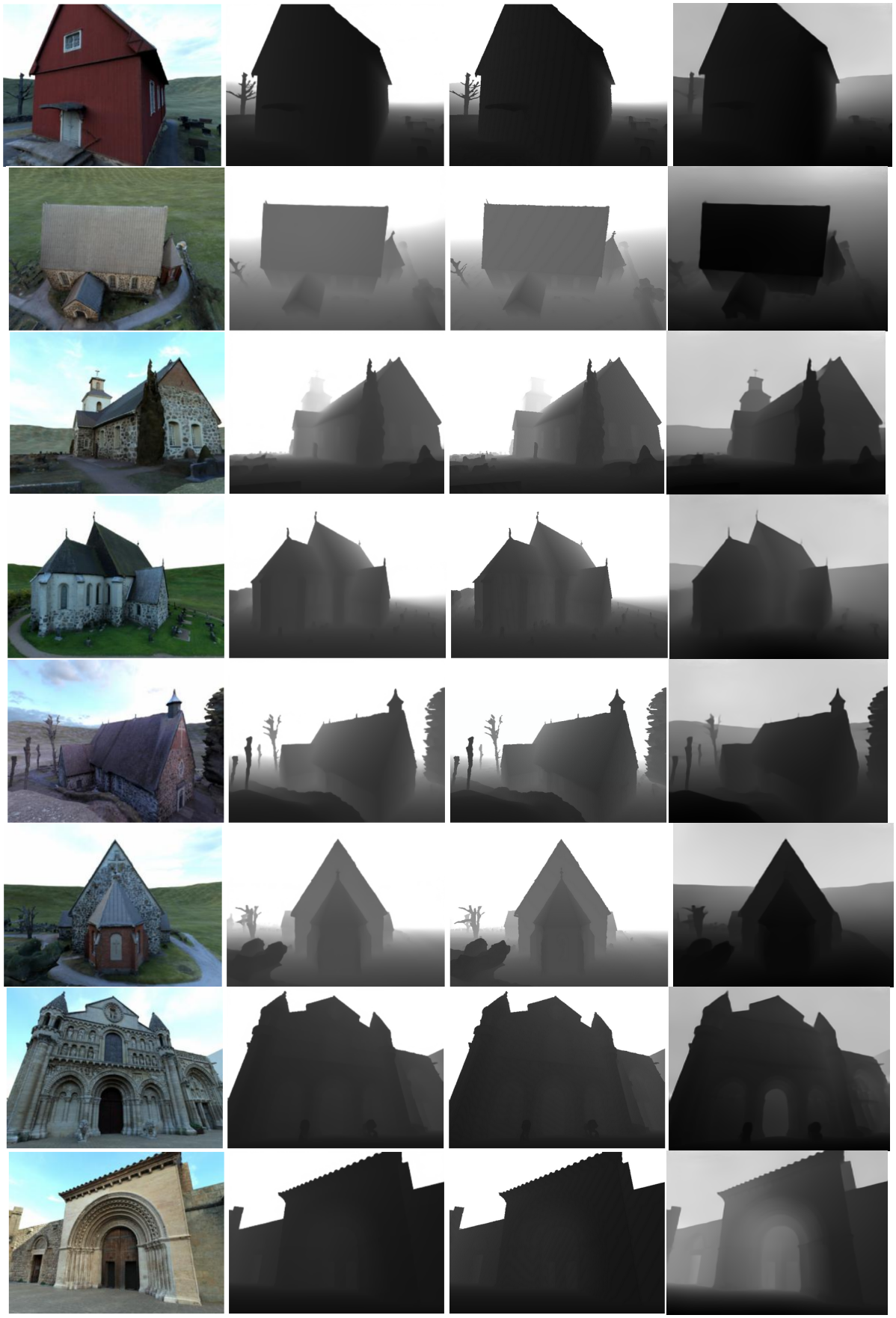

Figure 3. Experiment 1. (a) RGB input image (b) predicted depth map (c) GT depth map (d) LeRes depth map.

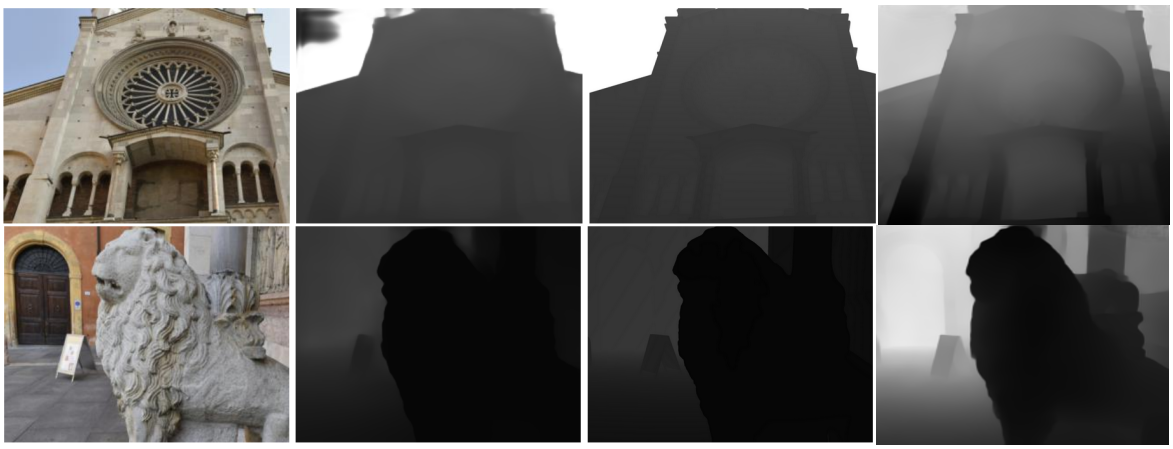

Figure 4. Experiment 2. (a) RGB input image (b) predicted depth map (c) GT depth map (d) LeRes depth map.

The low scores of Figure 5c are due to a present shift, probably because of incorrect estimation of absolute minimum and maximum depth values. However, in Experiment 1 these failure cases are rare. This fact is more present in Experiment 2.

4.2 Inference of LeRes

We predict on our images using the released pre-trained model of LeReS (Yin et al., 2021), a current state-of-the-art method for recovering the 3D structure of the scene from a single image. Although the method has demonstrated satisfying results in depth estimation and 3D reconstruction on various benchmarks in the original work, we observe that it does not generalize particularly well on our data. Depth maps seem to have kept the ordinal relations; however, the absolute and minimum depth values are not consistent with the GT (Figures 3,4), a fact also demonstrated in the low 2D scores in Table 1. The prediction on the 3D reconstruction behaves similarly, with evident distortions and scale inconsistencies (Figures 5 and 7).

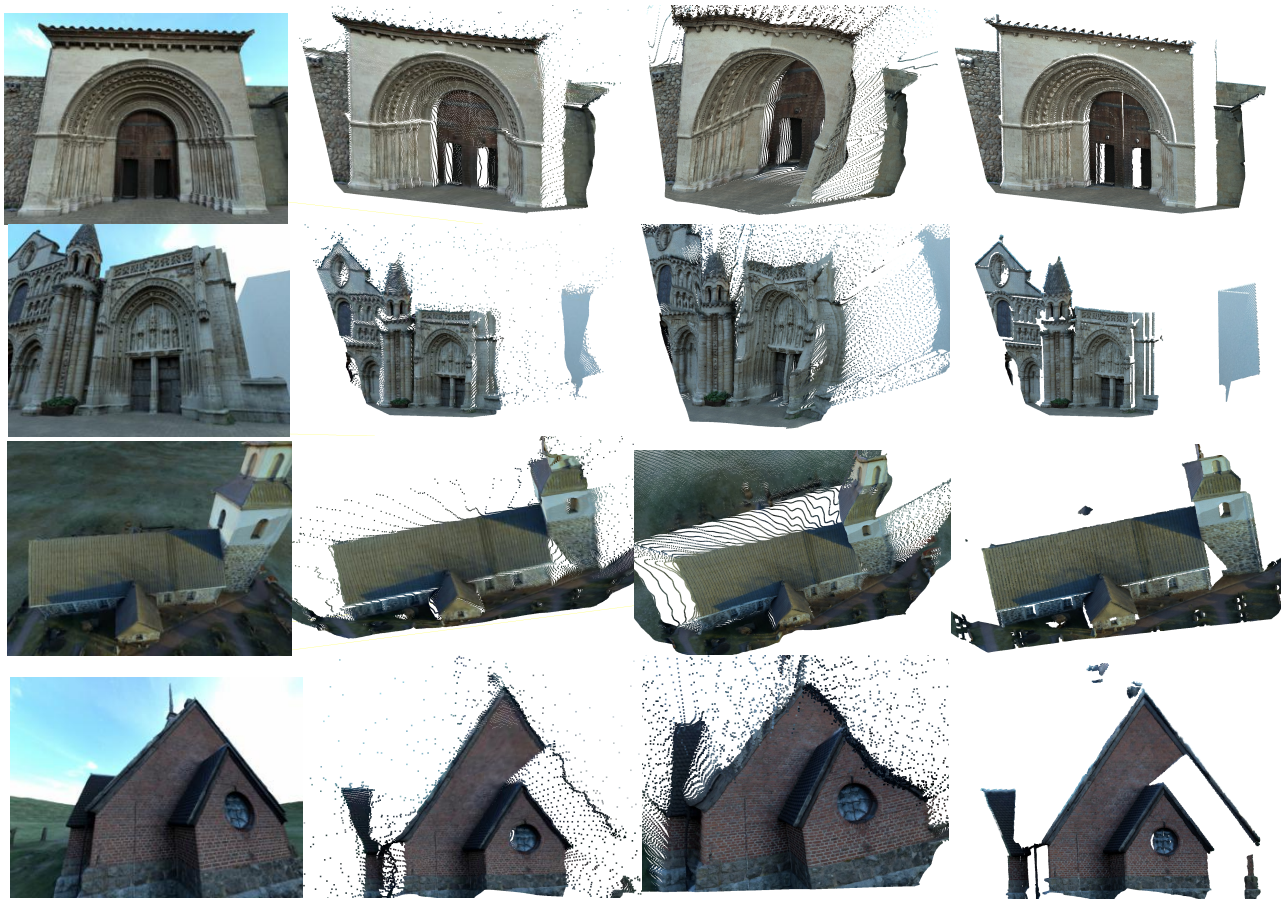

Figure 5. Experiment 1. (a) RGB input image (b) reconstructed point cloud (c) LeRes point cloud (d) GT point cloud.

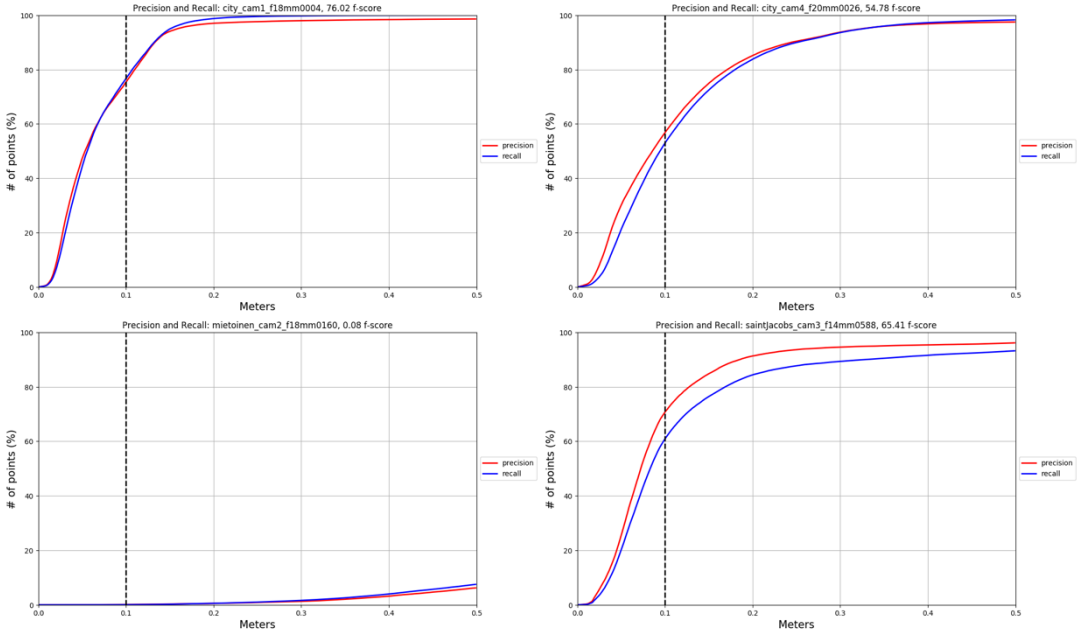

Figure 6. Experiment 1. Precision and recall curves for the four test images in Figure 5.

Figure 7. Experiment 2. (a) RGB input image (b) reconstructed point cloud (c) LeRes point cloud (d) GT point cloud.

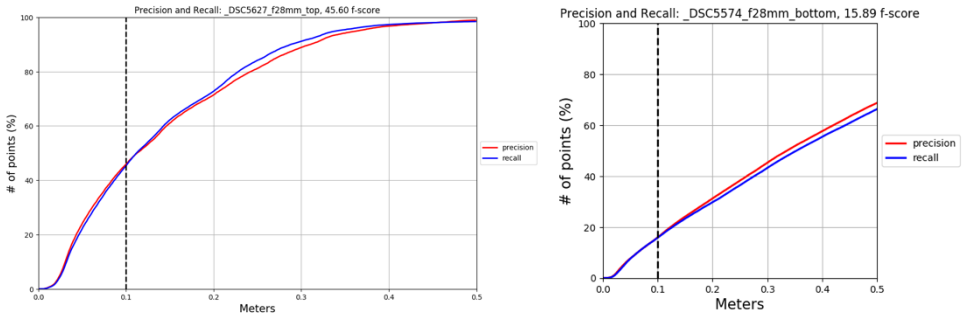

Figure 8. Experiment 2. Precision and recall curves for the two test images in Figure 6.

5. CONCLUSIONS

This paper proposes a new dataset for monocular depth prediction, composed of 24K images of outdoor scenes with great architectural details. Our dataset aims to provide high-quality metric depth benchmark data for training. We show its potential by training a straightforward encoder-decoder network and testing its robustness in predicting unseen views. The trained model was also fine-tuned using real-world images, typical for photogrammetric applications. Moreover, we employ a 3D reconstruction module to recover the shape of the scene using our predictions. Given that monocular depth estimation is by definition an ill-posed problem, such a reconstruction is not trivial without additional cues. Thus, despite the satisfying results on depth map prediction, improving the accuracy of the predictions in the 3D space is an open challenge. There is indeed

a need to shift the attention to 3D structure recovery and investigate more in this direction.

REFERENCES

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

Alhashim, I. and Wonka, P., 2018. High-quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*.

Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. PAMI*, *39*(12), pp. 2481-2495.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L., 2017. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. *IEEE Trans. PAMI*, *40*(4), pp. 834-848.

Chen, T., An, S., Zhang, Y., Ma, C., Wang, H., Guo, X. and Zheng, W., 2020. Improving monocular depth estimation by leveraging structural awareness and complementary datasets. Proc. ECCV, pp. 90-108.

Chen, R., Mahmood, F., Yuille, A. and Durr, N.J., 2018. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*.

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T. and Nießner, M., 2017. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. Proc. CVPR, pp. 5828-5839.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. CVPR, pp. 248-255.

Eigen, D., Puhrsch, C. and Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. *NIPS*, *27*.

Eigen, D. and Fergus, R., 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *Proc. ICCV*, pp. 2650-2658.

Fu, H., Gong, M., Wang, C., Batmanghelich, K. and Tao, D., 2018. Deep ordinal regression network for monocular depth estimation. Proc. *CVPR*, pp. 2002-2011.

Garg, R., Bg, V.K., Carneiro, G. and Reid, I., 2016. Unsupervised CNN for single view depth estimation: Geometry to the rescue. Proc. *ECCV*, pp. 740-756. Springer, Cham.

Geiger, A., Lenz, P. and Urtasun, R., 2012. Are we ready for autonomous driving? the KITTY vision benchmark suite. Proc. *CVPR*, pp. 3354-3361.

Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. *CVPR*, pp. 580-587.

Godard, C., Mac Aodha, O. and Brostow, G.J., 2017. Unsupervised monocular depth estimation with left-right consistency. Proc. *CVPR*, pp. 270-279.

Ioffe, S. and Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448-456. PMLR.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. Proc. *CVPR*, pp. 770-778.

He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask R-CNN. Proc. *ICCV*, pp. 2961-2969.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. Proc. *CVPR* pp. 807-814.

Hu, J., Ozay, M., Zhang, Y. and Okatani, T., 2019. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. Proc. *WACV*, pp. 1043-1051.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. Proc. CVPR, pp. 4700-4708.

Huang, B., Yi, H., Huang, C., He, Y., Liu, J. and Liu, X., 2021. M3VSNet: Unsupervised multi-metric multi-view stereo network. *Proc. ICIP*, pp. 3163-3167.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *NISP*, *25*, pp. 1097-1105.

Knapitsch, A., Park, J., Zhou, Q.Y. and Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM *Transactions on Graphics* (ToG), 36(4), pp.1-13.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F. and Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. Proc. *4th IEEE 3DV*, pp. 239-248.

Lee, J.H., Han, M.K., Ko, D.W. and Suh, I.H., 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*.

Li, Z. and Snavely, N., 2018. MegaDepth: Learning single-view depth prediction from internet photos. Proc. *CVPR*, pp. 2041-2050.

Liu, F., Shen, C., Lin, G. and Reid, I., 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. PAMI*, *38*(10), pp.2024-2039.

Liu, C., Kim, K., Gu, J., Furukawa, Y. and Kautz, J., 2019. PlaneRCNN: 3d plane detection and reconstruction from a single image. *Proc. CVPR*, pp. 4450-4459.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proc. *CVPR*, pp. 3431-3440.

Madhuanand, L., Nex, F. and Yang, M.Y., 2021. Self-supervised monocular depth estimation from oblique UAV videos. *ISPRS Journal of Photogrammetry and Remote Sensing*, *176*, pp.1-14.

McCormac, J., Handa, A., Leutenegger, S. and Davison, A.J., 2017. Scenenet RGB-D: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? Proc. *ICCV*, pp. 2678-2687.

Ranftl, R., Bochkovskiy, A. and Koltun, V., 2021. Vision transformers for dense prediction. Proc. *CVPR*, pp. 12179-12188.

Silberman, N., Hoiem, D., Kohli, P. and Fergus, R., 2012. Indoor segmentation and support inference from RGB-D images. Proc. *ECCV*, pp. 746-760. Springer, Berlin, Heidelberg.

Steenbeek, A., Nex. F., 2022. CNN-Based Dense Monocular Visual SLAM for Real-Time UAV Exploration in Emergency Conditions. *Drones,* Vol. 6(3): 79.

Song, S., Lichtenberg, S.P. and Xiao, J., 2015. Sun RGB-D: A RGB-D scene understanding benchmark suite. Proc. *CVPR*, pp. 567-576.

Tosi, F., Aleotti, F., Poggi, M. and Mattoccia, S., 2019. Learning monocular depth estimation infusing traditional stereo knowledge. Proc. *CVPR*, pp. 9799-9809.

Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), pp.600-612.

Xu, D., Ricci, E., Ouyang, W., Wang, X. and Sebe, N., 2017. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. Proc. *CVPR*, pp. 5354-5362.

Xu, Q. and Tao, W., 2020. Planar prior assisted PatchMatch multi-view stereo. Proc. *AAAI Conference on Artificial Intelligence*, *34*(7), pp. 12516-12523.

Yao, Y., Luo, Z., Li, S., Fang, T. and Quan, L., 2018. MVSNet: Depth inference for unstructured multi-view stereo. Proc. *ECCV*, pp. 767-783.

Yang, G., Tang, H., Ding, M., Sebe, N. and Ricci, E., 2021. Transformer-based attention networks for continuous pixel-wise prediction. Proc. ICCV, pp. 16269-16279.

Yin, W., Liu, Y., Shen, C. and Yan, Y., 2019. Enforcing geometric constraints of virtual normal for depth prediction. Proc. *ICCV*, pp. 5684-5693.

Yin, W., Wang, X., Shen, C., Liu, Y., Tian, Z., Xu, S., Sun, C. and Renyin, D., 2020. DiverseDepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*.

Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S. and Shen, C., 2021. Learning to recover 3d scene shape from a single image. *Proc. CVPR,* pp. 204-213.

Zbontar, J. and LeCun, Y., 2015. Computing the stereo matching cost with a convolutional neural network. Proc. *CVPR*, pp. 1592-1599.

Zhou, Q.Y., Park, J. and Koltun, V., 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*.