THE UNIVERSITY of ADELAIDE

3D Scene Reconstruction from A Monocular Image

Wei Yin

A thesis submitted for the degree of
DOCTOR OF PHILOSOPHY
The University of Adelaide

January 23, 2022

Contents

List of Figures

List of Tables

University of Adelaide

Abstract

3D Scene Reconstruction from A Monocular Image

by Wei Yin

3D scene reconstruction is a fundamental task in computer vision. The established approaches to address this task are based on multi-view geometry, which create correspondence of feature points with consecutive frames or multiple views. Finally, 3D information of these feature points can be recovered. In contrast, we aim to achieve dense 3D scene shape reconstruction from a single in-the-wild image. Without multiple views available, we rely on deep learning techniques. Recently, deep neural networks have been the dominant solution for various computer vision problems. Thus, we propose a two stage method based on learning-based methods. Firstly, we employ fully-convolutional neural networks to learn accurate depth from a monocular image. To recover high-quality depth, we lift the depth to 3D space and propose a global geometric constraint, termed virtual normal loss. To improve the generalization ability of the monocular depth estimation module, we construct a large-scale and diverse dataset and propose to learn the affine-invariant depth on that. Experiments demonstrate that our monocular depth estimation methods can robustly work in the wild and recover high-quality 3D geometry information. Furthermore, we propose a novel second stage to predict the focal length with a point cloud network. Instead of directly predicting it, the point cloud module leverages point cloud encoder networks that predict focal length adjustment factors from an initial guess of the scene point cloud reconstruction. The domain gap is significantly less of an issue for point clouds than that for images. Combing two stage modules together, 3D shape can be recovered from a single image input. Note that such reconstruction is up to a scale. To recover metric 3D shape, we propose to input the sparse points as guidance. Our proposed training method can significantly improve the robustness of the system, including robustness to various sparsity patterns and diverse scenes.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Wei Yin

Oct. 2021

Acknowledgements

Foremost, I want to express my deep and sincere gratitude to my supervisor Prof. Chunhua Shen for the continuous support, constant encouragement, and instructive advice to my Ph.D. study and research. His guidance helped me in all the time of research and finishing this thesis. He has taught me the methodology to carry out the research and to present the research outcomes as clearly as possible.

I would also like to thank Dr. Jianming Zhang, Dr. Oliver Wang, Dr. Simon Nicklaus, Dr. Simon Chen, Dr. Mai Long for their kind help on the difficulties during my internship.

I would also thank my friends and colleagues. They are Dr. Changming Sun, Dr. Zhi Tian, Dr. Hao Chen, Dr. Tong He, Dr. Yifan Liu, Dr. Xinyu Zhang, Dr. Bohan Zhuang, Jiawang Bian, Dr. Dong Gong, Libo Sun, Yunzhi Zhuge, Dr. Hu wang, Weian Mao, Dr. Lingqiao Liu, Dr. Zhipeng Cai. I am grateful for their support in my hard time.

Last but not the least, I would like to thank my families. They support me to focus on my research. Without their encouragements and love, I would not have gone through three such challenging but rewarding years.

Publications

This thesis contains the following works that have been published or prepared for publication:

- Enforcing Geometric Constraints of Virtual Normal for Depth Prediction.
  Wei Yin, Yifan Liu, Chunhua Shen, Youliang Yan.
  International Conference on Computer Vision (ICCV), 2019.

- Virtual Normal: Enforcing Geometric Constraints for Accurate and Robust Depth Prediction.
  Wei Yin, Yifan Liu, Chunhua Shen.
  IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021.

- Learning to Recover 3D Scene Shape from a Single Image.
  Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, Chunhua Shen.
  IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

- Diversedepth: Affine-invariant Depth Prediction Using Diverse Data.
  Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, Dou Renyin.
  ArXiv Preprint, 2020.

- Towards Domain-agnostic Depth Completion.
  Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Chunhua Shen.
  IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, Under review.

- Towards Accurate Reconstruction of 3D Scene Shape from Single Monocular Images.
  Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Long Mai, Chunhua Shen.
  IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Major Revision, 2021.

- The Devil Is in the Labels: Semantic Segmentation from Sentences.
  Wei Yin, Yifan Liu, Chunhua Shen, Anton van den Hengel, Baichuan Sun.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, Under review.

In addition, I have the following papers not included in this thesis:

- Generic Perceptual Loss for Modeling Structured Output Dependencies.
  Yifan Liu, Hao Chen, Yu Chen, **Wei Yin**, Chunhua Shen
  Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

- Task-aware Monocular Depth Estimation for 3D Object Detection.
  Xinlong Wang, **Wei Yin**, Tao Kong, Yuning Jiang, Lei Li, Chunhua Shen
  AAAI Conference on Artificial Intelligence (AAAI), 2020.

- Improving Monocular Visual Odometry Using Robustly Learned Depth,
  Libo Sun*, **Wei Yin***, Enze Xie, Zhengrong Li, Changming Sun, and Chunhua Shen.
  Transactions on Robotics (TRO, Minor Revision) 2021.

- Pseudo-LiDAR Based Road Detection,
  Libo Sun, Haokui Zhang, **Wei Yin**
  IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2022.

Chapter 1

Introduction

Human beings have strong abilities to interact with the world based on understanding of the 3D information of environments, such as grasping objects and avoiding obstacles. Even with a single eye, they can still understand the 3D layout of an environment. This implies that human beings are able to create an implicit reconstruction of a 3D environment from a single monocular viewpoint. The question of how to empower machines with this ability to understand 3D environments from a monocular viewpoint has been a long lasting problem. There have been many different 3D reconstruction methods proposed, which can be mainly categorized to active methods and passive methods. The former ones use the assistant optical information for reconstruction, such as coded patterns, time of flight and so on. By contrast, the passive methods focus on leveraging the geometry information and image features to recover the 3D information. In this thesis, we mainly discuss the passive methods.

Most traditional passive methods [117] for reconstructing the 3D scene are based on the multi-view geometry theory. They utilize hand-crafted features, such as SIFT [88] and ORB [108], to establish correspondences on consecutive frames or multiple views, recovers the camera intrinsic and extrinsic parameters, and then uses the triangulation to recover 3D information. With multi-view inputs, high-quality 3D reconstruction has been achieved.

However, there are many scenarios where only a monocular image is available, such as image editing, autonomous driving, and so on. In this thesis, we move forward to study the 3D reconstruction from a single image input on in-the-wild scenes. Previous geometry-based methods are not suited or applicable for this problem due to the limited amount of geometric information they are able to extract from monocular images. In recent years, the computer vision community has witnessed the continuously improved state-of-the-art performance on various vision problems with the help of deep-learning methods. They have surpassed traditional methods by large margins across a range of historically difficult tasks. Therefore, we propose to employ a supervised learning method capable of predicting depth, and other camera parameters from monocular images to create accurate 3D scene reconstructions from real world images. However, in order to adopt learning based methods, there are several problems that need to be solved: 1) how to enforce the constraint for the model to recover high-quality depth? 2) how can the system work robustly in the wild? 3) how to

recover the camera intrinsic parameters for the 3D reconstruction? 4) and can this system recover accurate metric information?

1.1 Motivation

In recent years, many methods [149, 35, 82, 150, 160, 75] have been proposed to solve the monocular depth estimation problem. They typically formulate the optimization problem as either point-wise regression or classification. The overall loss is summed over all pixels. However, we found that the reconstructed point cloud from the predicted depth is far from the ground truth 3D structure, despite their deceptively low amounts of error. To improve the geometry accuracy, some methods propose to incorporate the local geometry constraint, such as the surface normal. The problem is that many RGB-D datasets are captured by consumer-level sensors, such as Kinect and RealSense, which introduce much noise in local regions. Such noisy data would adversely affect the supervision. To solve these problems, we propose to explore the global geometric constraint enforced on the 3D point cloud, which is more stable and takes the long-range relations into account.

Most of the current methods mainly train and validate their methodologies on the dataset with a specific scene, such as indoor or outdoor scenarios. As a result, these specialised models are unable to generalise to diverse scenes. To solve this problem, some methods propose to construct a large-scale and diverse dataset, and explore the pair-wise ordinal relations for learning. However, only the relative depth can be predicted and geometric information is lost. To ensure both good generalization and high-quality 3D depth information, we seek to address these problems from 3 aspects. We firstly propose to construct a diverse dataset, which has diverse scenes and is constructed with various cameras and various camera poses. Compared with learning relative depth or metric depth, we propose to learn the affine-invariant depth.

With the proposed learning objective and datasets, the affine-invariant depth can be accurately and robustly predicted on diverse scenes. However, recovering a dense 3D scene reconstruction from an image still requires the challenges of depth shift and focal length to be solved. Some works propose to formulate these as regression problems and predict them from the 2D image input. By contrast, we discover that estimating these parameters from the point cloud is more robust, and results in improved generalisation with unobserved scenes. Importantly, owing to the smaller domain gap on point cloud than 2D images, we can train models on synthetic data. Therefore, we employ a point cloud network to recover focal length and depth shift in the second stage.

1.1.1 Contribution

The main contributions of this thesis include a set of algorithms for robustly reconstructing accurate 3D scene shape from a single image input. They are listed as follows.

- We propose a high-order geometric constraint enforced in the 3D space for the monocular depth estimation. Such global geometry information is instantiated with a simple yet effective concept termed virtual normal(VN). By enforcing a loss defined on VNs, we demonstrate the importance of 3D geometry information in depth estimation, and design a simple loss to exploit it. Most importantly, we demonstrate that this method can reconstruct high-quality 3D scene point clouds, from which other 3D geometry features can be directly recovered, such as the surface normal.

- To solve the generalization issue of monocular depth estimation, We propose a large-scale and high-diversity RGB-D dataset, DiverseDepth. It contains diverse scenes, and is captured with various cameras in various camera poses.

- To solve the training issue on large scale and diverse datasets, we propose to learn affine-invariant depth, which ensures both high generalization and high-quality geometric shapes of scenes. Furthermore, we propose a multi-curriculum learning method to boost the performance. Experiments on 8 zero-shot datasets show our method outperforms previous methods noticeably.

- We propose a novel framework for in-the-wild monocular 3D scene shape estimation. To our knowledge, this is the first method for this task, and the first method to leverage 3D point cloud neural networks for improving estimation of the structure of point clouds derived from depth maps.

- Furthermore, to improve the performance of the depth estimation, we propose an image-wise normalized regression loss and a pair-wise normal regression loss.

1.2 Thesis Outline

The structure of this thesis is organized as follows.

In Chapter 2, we firstly review existing state-of-the-art monocular depth estimation methods, current RGB-D datasets, some 3D reconstruction methods, and depth completion methods. Also we review the point cloud networks in detail.

In Chapter 3, we propose a global geometric constraint in 3D space, termed virtual normal loss. With the virtual normal loss, the monocular depth estimation performance is boosted a lot. Importantly, the depth can reconstruct high-quality 3D scene point clouds

In Chapter 4, to solve the generalization issue, we propose a large-scale and diverse dataset, DiverseDepth, and propose to learn affine-invariant depth on it.

In Chapter 5, to solve the 3D scene reconstruction from a single image input, we propose a staged method, which recovers the affine-invariant depth first and then predicts the camera focal length and depth shift to do the reconstruction.

In Chapter 6, we propose to combine the single image and a sparse depth map to recover the metric depth and do the metric reconstruction.

In Chapter 7, the conclusion and the potential research directions are discussed.

Chapter 2

Literature Review

Our method aims to solve the 3D reconstruction from a monocular image. In this section, we reviewed most-related methods, including monocular depth estimation, 3D reconstruction, depth completion, RGBD datasets, and 3D point cloud networks.

2.1 Monocular Depth Prediction

Monocular depth prediction aims to predict pixel-wise depth from a single image, which is important for many robotic and vision applications. It is an ill-posed problem because multiple 3D scenes can be projected to the same 2D image. To solve this problem, most of methods are data-driven. Based on their supervision methods, they are categorised to supervised monocular depth estimation [149, 15, 159, 161] and unsupervised/self-supervised depth estimation [39, 5, 73].

2.1.1 Supervised Monocular Depth Estimation.

Saxena *et al.* [113] are among the first ones proposing to predict depth from a single image. They construct a Markov Random Field (MRF) model that incorporates multi-scale local and global image features. Later, a few methods [114, 79] based on the probabilistic model are proposed. When the powerful deep convolutional neural network emerges and benefits various computer vision tasks, many CNN-based methods are also proposed. Eigen *et al.* [30, 29] propose the first multi-scale network for dense prediction, including monocular depth prediction, surface normal estimation, and semantic estimation. Liu *et al.* [80] proposes to combines the CNN and CRF for depth estimation. Besides the study on the network architecture, many endeavours [36, 161, 8, 75, 101, 27] have been done on leveraging supervisions to improve the performance. Some works [161, 36, 75, 27] model the depth prediction as a classification problem. Fu *et al.* [36] transfer the depth to multiple bins and propose a multiple 2-class classification loss, i.e. ordinal loss, to supervise the network. By contrast, some works [75, 161, 9] model the depth estimation as a multi-class classification problem. Furthermore, to boost the depth quality, some works [101] propose to combine the depth estimation with other geometry features estimation together, such as surface normal. Qi *et al.* [101] propose to jointly predict the surface normal and depth, which can refine the depth map based on the constraints from the surface normal. Fei [32]

proposed a semantically informed geometric loss while Yin et al. [146] introduced a virtual normal loss to exploit the structure information.

All previous methods propose to learn the metric depth. The problem is that such methods are difficult to generalize to diverse test scenes, mainly due to the lack of sufficiently large-scale and high-quality training datasets. To improve generalization, methods that learn the relative depth [77, 149, 150, 17, 15] are proposed, as relative depth is much easier to obtain than metric depth. Chen *et al.* [17] construct the first large-scale and highly diverse dataset for learning the relative depth. As they use the ordinal relations and there is a large-scale dataset for training, their method can produce a model with good generalization. To construct better quality training data, Xian *et al.* [149, 150] propose to collect stereo images or videos and use optical flow methods to obtain the inverse depth. Although learning relative depth can obtain a robust model, the relative depth can only represent depth ordinal relations and lost the geometry information, i.e. one point is farther or closer than another one. To solve this problem, some works [104, 162, 159, 160] propose to learn affine-invariant depth. Ranftl *et al.* [104] propose the scale-shift invariant loss to leverage the training on multi-source data, which can achieve promising generalization on diverse scenes. Yin *et al.* [162] propose a heterogeneous loss training strategy, which can achieve state-of-the-art performance on multiple zero-shot testing datasets.

2.1.2 Unsupervised/Self-supervised Monocular Depth Estimation.

Apart from these supervised learning methods, some works [5, 169, 42, 39, 123] propose to solve the monocular depth estimation problem without sensor captured ground truth depth but leveraging the training signal from consecutive temporal frames or stereo videos. Zhou *et al.* [169] propose the first monocular self-supervised approach, which trains a monocular depth estimation network along with a separate camera pose estimation network from monoculr videos. They use an image alignment loss, which is obtained by warping the source image to the neighboring frames with the predicted depth and ego-motion, to supervise the network. To remove the non-rigid scene motion that violates the rigid warping process, they propose to use an additional motion explanation mask to ignore specific regions. Yin *et al.* [163] propose to decompose motion into rigid and non-rigid components, using depth and optical flow to explain object motion. This can improve the flow estimation, but jointly training both flow and depth cannot see more improvement. To close the gap with fully-supervised methods, Godard *et al.* [43] propose to leverage the consistency signal from consecutive frames and stereo views. Yang *et al.* [157] constrain the predicted depth to be consistent with predicted surface normals, and [156] enforced edge consistency. To improve the scale consistency between consecutive frames, Bian *et al.* [5] propose the geometry consistency loss. Ranjan *et al.* [105] propose to solve multiple low-level vision problems simultaneously, including depth, camera motion, optical

flow, and moving objects segmentation, because such fundamental problems are coupled together through geometric constraints. Furthermore, several works [21, 60, 170] propose to leverage the geometric relations between consecutive frames.

2.2 RGB-D Datasets

Datasets [114, 41, 124, 25, 136] are significant for the advancement of data-driven depth prediction methods. According to the quality of the ground truth depth, these datasets can be summarized into two categories. Depth sensors are used to directly collect high-quality RGB-D pairs, which can construct accurate metric depth dataset. Make3D [114] is the first outdoor RGB-D dataset constructed for monocular depth prediction study. KITTI [41] and NYUD [124] are captured by LIDAR on outdoor streets and Kinect in indoor rooms. Larger-scale RGB-D datasets are also constructed, such as ScanNet [25], Taskonomy [165], DIML [23], DIODE [136]. These datasets usually only contain very limited scenes.

To improve the generalization of depth estimation methods on diverse scenes, several large-scale and diverse datasets are constructed, but the depth is not of high quality. Chen *et al.* [17] construct the largest RGB-D dataset, where the ground-truth depth maps are manually annotated with only one pair of ordinal relations. Similarly, Youtube3D [14] is also constructed to learn the relative depth but with more pairs of ordinal relations. MegaDepth [77] employs structure from motion to construct the depth supervision on the still and rigid scenes. To include more non-rigid and diverse scenes, Xian *et al.* [149] and Wang *et al.* [137] employ optical flow methods to construct datasets of relative depth. Chen *et al.* [15] propose the diverse OASIS dataset, which includes both depth ordinal annotations and camera intrinsic parameters. Yin *et al.* [160, 159] propose another large-scale and diverse RGB-D datasets, DiverseDepth.

2.3 Sparse Depth Completion

Depth completion aims to densify a sparse depth input. As the sparse depths captured by different solutions have varying sparsity types, several methods are proposed to solve these problems. Depth maps captured with low-cost LiDAR only have a few hundreds or thousands of valid measurements per image. Several methods [22, 96, 21, 14, 155] propose to leverage the texture information to complete these types of sparsity patterns. Besides such very sparse depth types, commodity-level RGB-D cameras such as Kinect, RealSense, and Tango produce depth images that are semi-dense but missing certain regions. This often happens due to objects with low reflective properties and objects beyond the maximum supported distance. Several methods treat this as a depth inpainting task and leverage smoothness priors [49], background surface extrapolation [90], and surface normals [168]. These methods have shown promising results, but they focus on only a single sparse depth type. In contrast, we design a unified solution for all these depth sparsity patterns. Additionally, we

propose to use a pretrained scale-shift-invariant depth prediction model as a scene prior to improve the depth completion quality.

2.4 3D Reconstruction

A number of works have addressed reconstructing different types of objects from a single image [3, 138, 148], such as humans [110, 111], cars, planes, tables, etc. The main challenge is how to best recover objects details, and how to represent them with limited memory. Pixel2Mesh [138] proposes to reconstruct the 3D shape from a single image and express it in a triangular mesh. PIFu [110, 111] proposes an memory-efficient implicit function to recover high-resolution surfaces, including unseen/occluded regions, of humans. However, all these methods rely on learning priors specific to a certain object class or instance, typically from 3D supervision, and can therefore not work for full scene reconstruction.

On the other hand, several works have proposed reconstructing 3D scene structure from a single image. Saxena *et al.* [114] assume that the whole scene can be segmented into several pieces, of which each one can be regarded as a small plane. They predict the orientation and the location of the planes and stitch them together to represent the scene. Other works propose to use image cues, such as shading [97] and contour edges [61] for scene reconstruction. However, these approaches use hand-designed priors and restrictive assumptions about the scene geometry. Our method is fully data driven, and can be applied to a wide range of scenes.

2.5 3D Deep Learning Models

3D vision attracts increased attention recently because of wide applications in AR/VR, autonomous driving, and so on. 3D data are usually in the format of 3D point cloud, which is a set of (x, y, z) coordinates. How to represent the point cloud and extract features from them are the main challenge. Many works [91, 99, 145, 24, 86] employ volumes to represent and process 3D data. Maturana *et al.* [91] proposed the first voxelnet. Qi *et al.* [99] systematically analyze CNNs based upon volumetric representations and CNNs based upon multi-view representations. They propose two volumetric CNN network architectures that significantly improve volumetric CNNs on 3D shape classification. To reduce the memory consumption of processing point cloud, Wang *et al.* [140] incorporate the octree into volumetric CNNs. In 3D scene segmentation problems, volumetric representation based methods [145, 24, 86] are also widely applied.

Apart from volumetric representation, some works [76, 144, 100, 98] propose to directly process unordered point cloud. Qi *et al.* [98] propose the first point cloud network which takes advantage of the symmetric function to process point cloud. Their following work, PointNet++ [100], increases the model capacity by stacking PointNets hierarchically to leverage the nearest neighborhood information. Besides,

some works [154, 76] propose to use the dynamically created convolution kernels to extract neighborhood features instead of the symmetric function.

2.6 Curriculum learning.

For many applications, introducing concepts in ascending difficulty to the learner is a common practice. Several works have demonstrated that curriculum learning [147, 46, 4] can boost the performance of deep learning methods. Weinshall *et al.* [147] combine the transfer learning and curriculum learning methods to construct a better curriculum, which can improve both the speed of convergence and the final accuracy. Hacohen and Weinshall [46] propose a bootstrapping method to train the network by self-tutoring. To train a robust depth model on noisy datasets, Yin *et al.* [160, 159] construct an easy-to-hard curriculum to load data samples.

Statement of Authorship

Title of Paper	Enforcing geometric constraints of virtual normal for depth prediction.
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Yin, Wei, Yifan Liu, Chunhua Shen, and Youliang Yan. "Enforcing geometric constraints of virtual normal for depth prediction." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5684-5693. 2019.

Principal Author

Name of Principal Author (Candidate)	Wei Yin
Contribution to the Paper	Design new methods and conduct the experiments.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/13/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chunhua Shen
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Name of Co-Author	Yifan Liu
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Please cut and paste additional co-author pane

Name of Co-Author	Youliang Yan
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/12/2024

Chapter 3

Geometric Constraints for Accurate Monocular Depth Prediction

3.1 Introduction

Monocular metric depth estimation aims to predict the metric distance between scene objects and the camera from a single monocular image, which is a critical task for understanding the 3D scene, such as recognizing a 3D object and parsing a 3D scene. Most of current methods mainly focus on enforcing pixel-wise regression or classification loss to supervise the network. By contrast, we propose to leverage the geometry information.

In this chapter, we first investigate the local geometry constraint, such as surface normal. To improve the geometry constraint's robustness to noise, we propose a global geometry constraint, i.e. virtual normal, to improve the performance of the monocular depth estimation.

3.2 Background

Although the monocular depth prediction is an ill-posed problem because many 3D scenes can be projected to the same 2D image, many deep convolutional neural networks (DCNN) based methods [29, 30, 36, 44, 68, 75, 115] have achieved impressive results by using a large amount of labelled data, thus taking advantage of prior knowledge in labelled data to solve the ambiguity.

These methods typically formulate the optimization problem as either point-wise regression or classification. That is, with the i.i.d. assumption, the overall loss is summing over all pixels. To improve the performance, some endeavours have been made to employ other constraints besides the pixel-wise term. For example, a continuous conditional random field (CRF) [83] is used for depth prediction, which takes pair-wise information into account. Other high-order geometric relations [32, 101] are also exploited, such as designing a gravity constraint for local regions [32] or incorporating the depth-to-surface-normal mutual transformation inside the optimization

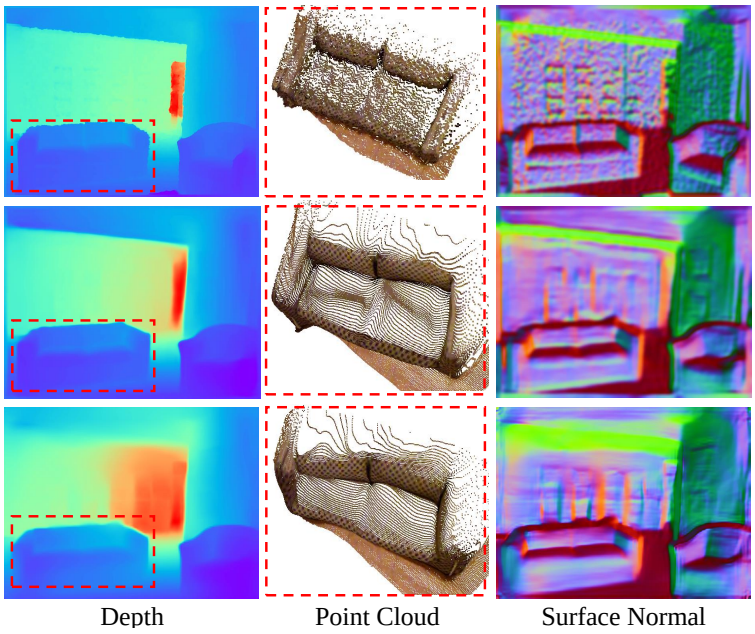


FIGURE 3.1. Example results of ground truth (the first row), our method (the second row) and Hu *et al.* [53] (the third row). By enforcing the geometric constraints of virtual normals, our reconstructed 3D point cloud can represent better shape of sofa (see the left part) and the recovered surface normal has much less errors (see green parts) even though the absolute relative error (rel) of our predicted depth is only slightly better than Hu *et al.* (0.108 *vs.* 0.115).

pipeline [101]. Note that, for the above methods, almost all the geometric constraints are 'local' in the sense that they are extracted from a small neighborhood in either 2D or 3D. Surface normal is 'local' by nature as it is defined by the local tangent plane. As the ground truth depth maps of most datasets are captured by consumer-level sensors, such as the Kinect, depth values can fluctuate considerably. Such noisy measurement would adversely affect the precision and subsequently the effectiveness of those local constraints inevitably. Moreover, local constraints calculated over a small neighborhood have not fully exploited the structure information of the scene geometry that may be possibly used to boost the performance.

To address these limitations, here we propose a more stable geometric constraint from a global perspective to take long-range relations into account for predicting depth, termed *virtual normal*. A few previous methods already made use of 3D geometric information in depth estimation, almost all of which focus on using surface normal. *We instead reconstruct the 3D point cloud from the estimated depth map explicitly.* In other words, we generate the 3D scene by lifting each RGB pixel in the 2D image to its corresponding 3D coordinate with the estimated depth map. This 3D point cloud serves as an intermediate representation. With the reconstructed point cloud, we can exploit many kinds of 3D geometry information, not limited to the surface normal. Here we consider the long-range dependency in 3D space by randomly sampling three non-colinear points with the large distance to form a *virtual plane*, of

which the normal vector is the proposed *virtual normal* (VN). The direction divergence between ground-truth and predicted VN can serve as a high-order 3D geometry loss. Owing to the long-range sampling of points, the adverse impact caused by noises in depth measurement is much alleviated compared to the computation of the surface normal, making VN significantly more accurate. Moreover, with randomly sampling we can obtain a large number of such constraints, encoding the global 3D geometric. *By converting estimated depth maps from images to 3D point cloud representations it opens many possibilities of incorporating algorithms for 3D point cloud processing to 2D images and 2.5D depth processing.* Here we show one instance of such possibilities.

By combining the high-order geometric supervision and the pixel-wise depth supervision, our network can predict not only an accurate depth map but also the high-quality 3D point cloud, subsequently other geometry information such as the surface normal. It is worth noting that we do not use a new model or introduce network branches for estimating the surface normal. Instead it is computed directly from the reconstructed point cloud. The second row of Fig. 3.1 demonstrates an example of our results. By contrast, although the previously state-of-the-art method [53] predicts the depth with low errors, the reconstructed point cloud is far away from the original shape (see, e.g., left part of 'sofa'). The surface normal also contains many errors. We are probably the first to achieve high-quality monocular depth and surface normal prediction with a single network.

Experimental results on NYUD-v2 [124] and KITTI [41] datasets demonstrate state-of-the-art performance of our method. Besides, when training with the lightweight backbone, MobileNetV2 [112], our framework provides a better trade-off between network parameters and accuracy. Our method outperforms other state-of-the-art real-time systems by up to 29% with a comparable number of network parameters. Furthermore, from the reconstructed point cloud, we directly calculate the surface normal, with a precision being on par with that of specific DCNN based surface normal estimation methods.

In summary, our main contributions of this work are as follow.

- We demonstrate the effectiveness of enforcing a high-order geometric constraint in the 3D space for the depth prediction task. Such global geometry information is instantiated with a simple yet effective concept termed *virtual normal* (VN). By enforcing a loss defined on VNs, we demonstrate the importance of 3D geometry information in depth estimation, and design a simple loss to exploit it.
- Our method can reconstruct high-quality 3D scene point clouds, from which other 3D geometry features can be calculated, such as the surface normal. In essence, we show that for depth estimation, one should not consider the information represented by depth only. Instead, converting depth into 3D point clouds and exploiting 3D geometry are likely to improve many tasks including depth estimation.

- Experimental results on NYUD-V2 and KITTI illustrate that our method achieves state-of-the-art performance.

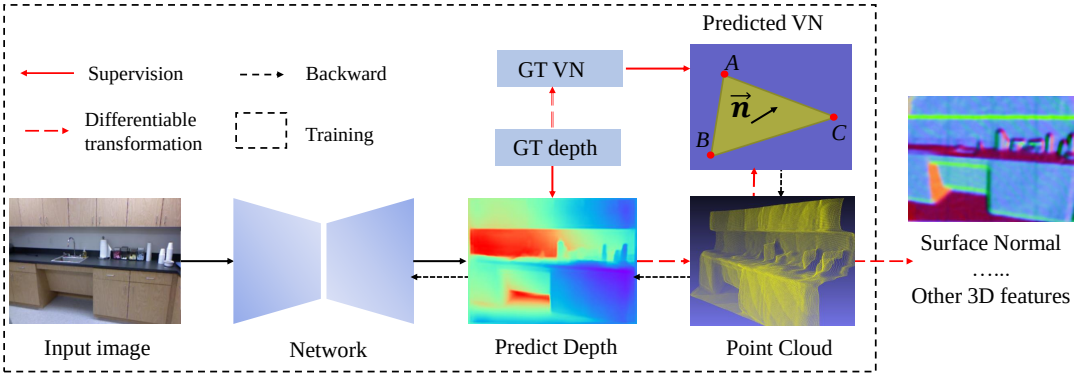


FIGURE 3.2. Illustration of the pipeline of our method. An encoder-decoder network is employed to predict the depth, from which the point cloud can be reconstructed. A pixel-wise depth supervision is firstly enforced on the predicted depth, while a geometric supervision, virtual normal constraint, is enforced in 3D space. With the well trained model, other 3D features, such as the surface normal, can be directly recovered from the reconstructed 3D point cloud in the inference.

3.3 Method

Our approach resolves the monocular depth prediction and reconstructs the high-quality scene 3D point cloud from the predicted depth at the same time. The pipeline is illustrated in Fig. 3.2.

We take an RGB image I_{in} as the input of an encoder-decoder network and predict the depth map D_{pred}. From the D_{pred}, the 3D scene point cloud P_{pred} can be reconstructed. The ground truth point cloud P_{gt} is reconstructed from D_{gt}.

We enforce two types of supervision for training the network. We firstly follow standard monocular depth prediction methods to enforce pixel-wise depth supervision over D_{pred} with D_{gt}. With the reconstructed point clouds, we then align the spatial relationship between the P_{pred} and the P_{gt} using the proposed *virtual normal*.

When the network is well trained, we not only obtain accurate depth map but also high-quality point clouds. From the reconstructed point clouds, other 3D features can be directly calculated, such as the surface normal.

3.3.1 High-order Geometric Constraints

Surface Normal. The surface normal is an important 'local' feature for many point-cloud based applications such as registration [109] and object detection [50, 45]. It appears to be a promising 3D cue for improving depth prediction. One can apply the angular difference between ground-truth and calculated surface normal to be a geometric constraint. One major issue of this approach is, when computing surface

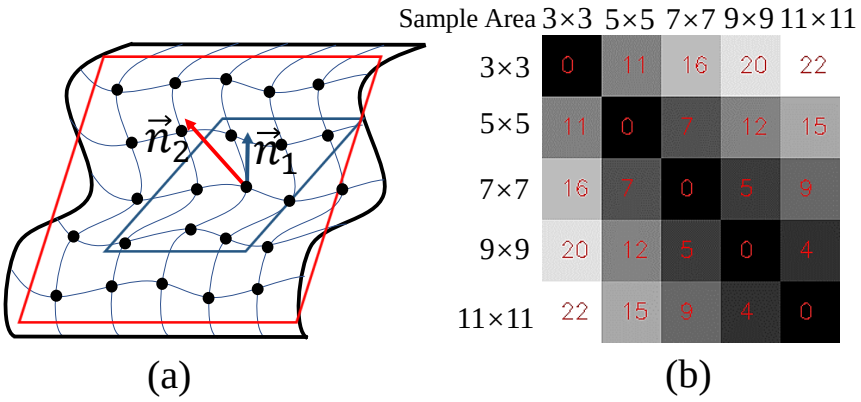


FIGURE 3.3. Illustration of fitting point clouds to obtain the local surface normal. The directions of the surface normals is fitted with different sampling sizes on a real point cloud (a). Because of noise, the surface normals vary significantly. (b) compares the angular difference between surface normals computed with different sample sizes in Mean Difference Error. The error can vary significantly.

normal from either a depth map or 3D point cloud, it is sensitive to noise. Moreover, surface normal only considers short-range local information.

We follow [64] to calculate the surface normal. It assumes that local 3D points locate in the same plane, of which the normal vector is the surface normal. In practice ground-truth depth maps are usually captured by a consumer-level sensor with limited precision, so depth maps are contaminated by noise. The reconstructed point clouds in the local region can vary considerably due to noises as well as the size of local patch for sampling (Fig. 3.3(a)).

We experiment on the NYUD-V2 dataset to test the robustness of the surface normal computation. Five different sampling sizes around the target pixel are employed to sample points, which are used to calculate its surface normal. The sample area is $a = (2i+1) \cdot (2i+1), i = 1, ..., 5$. The Mean Difference Error (Mean) [29] between calculated surface normals is evaluated. Errors are shown in Fig. 3.3(b). We can learn that the surface normal varies significantly with different sampling sizes. For example, the Mean between 3×3 and 11×11 is 22°. Such unstable surface normal negatively affects its effectiveness for learning. Likewise, other 3D geometric constraints demonstrating the 'local' relative relations also encounter this problem.

Virtual Normal. In order to enforce robust high-order geometric supervision in the 3D space, we propose the virtual normal (VN) to establish 3D geometric connections between regions in a much larger range. The point cloud can be reconstructed from the depth based on the pinhole camera model. For each pixel $p_i(u_i, v_i)$, the 3D location $P_i(x_i, y_i, z_i)$ in the world coordinate can be obtained by the prospective projection. We set the camera coordinate as the world coordinate. Then the 3D coordinate P_i is denoted as follows:

$$z_i = d_i, x_i = \frac{d_i \cdot (u_i - u_0)}{f_x}, y_i = \frac{d_i(v_i - v_0)}{f_y} \tag{3.1}$$

where d_i is the depth. f_x and f_y are the focal length along the x and y coordinate axis respectively. u_0 and v_0 are the 2D coordinate of the optical center.

We randomly sample N groups points from the depth map, with three points in each group. The corresponding 3D points are $\mathcal{S} = \{(P_A, P_B, P_C)_i | i = 0...N\}$. Three points in a group are restricted to be non-colinear based on the restriction $\mathcal{R}_1$. $\angle(\cdot)$ is the angle between two vectors.

$$\begin{aligned} \mathcal{R}_1 = \{&\alpha \geq \angle(\overrightarrow{P_A P_B}, \overrightarrow{P_A P_C}) \geq \beta, \\ &\alpha \geq \angle(\overrightarrow{P_B P_C}, \overrightarrow{P_B P_A}) \geq \beta | P \in \mathcal{S}\} \end{aligned} \tag{3.2}$$

where α, β are hyper-parameters. In all experiments, we set $\alpha = 120°$, $\beta = 30°$

In order to sample more long-range points, which have ambiguous relative locations in 3D space, we perform long-range restriction $\mathcal{R}_2$ for each group in $\mathcal{S}$.

$$\mathcal{R}_2 = \{\|\overrightarrow{P_k P_m}\| > \theta | k, m \in [A, B, C], P \in \mathcal{S}\} \tag{3.3}$$

where $\theta = 0.6m$ in our experiments.

Therefore, three 3D points in each group can establish a plane. We compute the normal vector of the plane to encode geometric relations, which can be written as

$$\begin{aligned} \mathcal{N} = \{\boldsymbol{n_i} = &\frac{\overrightarrow{P_{Ai} P_{Bi}} \times \overrightarrow{P_{Ai} P_{Ci}}}{\left\|\overrightarrow{P_{Ai} P_{Bi}} \times \overrightarrow{P_{Ai} P_{Ci}}\right\|} | \\ &(P_A, P_B, P_C)_i \in \mathcal{S}, i = 0...N\} \end{aligned} \tag{3.4}$$

where $\boldsymbol{n_i}$ is the normal vector of the virtual plane i.

Robustness to Depth Noise. Compared with local surface normal, our virtual normal is more robust to noise. In Fig. 3.4, we sample three 3D points with large distance. P_A and P_B are assumed to locate on the XY plane, P_C is on the Z axis. When P_C varies to $P_C{}'$, the direction of the virtual normal changes from $\boldsymbol{n}$ to $\boldsymbol{n}'$. $P_C{}''$ is the intersection point between plane $P_A P_B P_C{}'$ and Z axis. Because of restrictions $\mathcal{R}_1$ and $\mathcal{R}_2$, the difference between $\boldsymbol{n}$ and $\boldsymbol{n}'$ is usually very small, which is simple to show:

$$\begin{aligned} \angle(\boldsymbol{n}, \boldsymbol{n}') = &\angle(\overrightarrow{OP_C}, \overrightarrow{OP_C{}''}) = \arctan \frac{\|\overrightarrow{P_C P_C{}''}\|}{\|\overrightarrow{OP_C}\|} \approx 0, \\ &\|\overrightarrow{P_C P_C{}''}\| \ll \|\overrightarrow{OP_C}\| \end{aligned} \tag{3.5}$$

.

Furthermore, we conduct a simple experiment to verify the robustness of our proposed virtual normal against data noise. We create an unit sphere and then add Gaussian noise to simulate the ideal noise-free data and the real noisy data (see Fig. 3.5a). We then sample 100K groups of points from the noisy surface and the ideal one to compute the virtual normal respectively, while 100K points are sampled to compute the surface normal as well. For the gaussian noise, we use different deviations

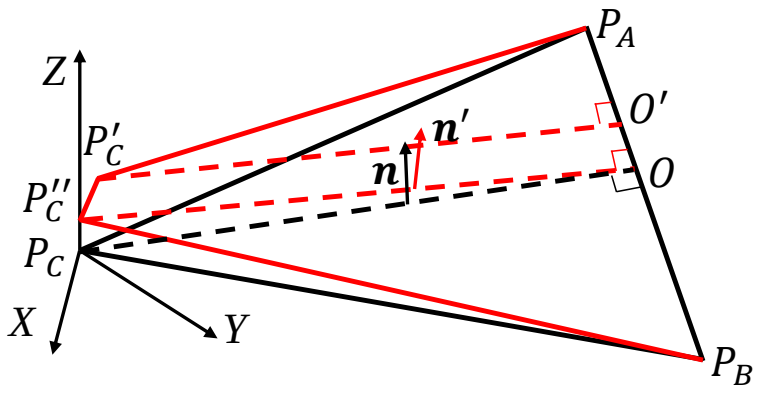


FIGURE 3.4. Robustness of VN to depth noise.

to simulate different noise levels by varying deviation $\sigma = [0.0002, ..., 0.01]$, and the mean being $\mu = 0$. The experimental results are illustrated in Fig. 3.5b. We can learn that our proposed virtual normal is much more robust to the data noise than the surface normal. Other local constraints are also sensitive to data noise.

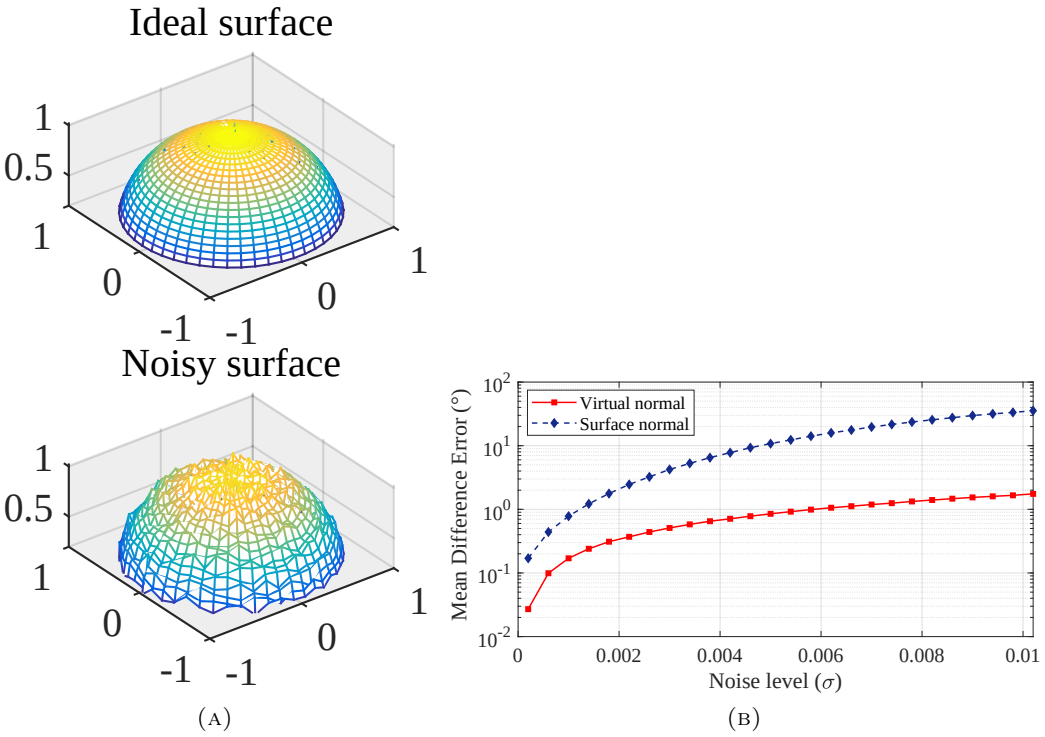


FIGURE 3.5. Robustness of virtual normal and surface normal against data noise. (a) The ideal surface and noisy surface. (b) The Mean Difference Error (Mean) is applied to evaluate the robustness of virtual normal and surface normal against different noise level. Our proposed virtual normal is more robust.

Most 'local' geometric constraints, such as the surface normal, actually enforcing the first-order smoothness of the surface but are less useful for helping the depth map prediction. In contrast, the proposed VN establishes long-range relations in the 3D space. Compared with pairwise CRFs, VN encodes triplet based relations, thus being

of high order.

Virtual Normal Loss. We can sample a large number of triplets and compute corresponding VNs. With the sampled VNs, we compute the divergence as the Virtual Normal Loss (VNL):

$$\ell_{VN} = \frac{1}{N}(\sum_{i=0}^{N} \|\boldsymbol{n}_i^{pred} - \boldsymbol{n}_i^{gt}\|_1) \tag{3.6}$$

where N is the number of valid sampling groups satisfying $\mathcal{R}_1, \mathcal{R}_2$. In our experiments, we have employed online hard example mining.

Pixel-wise Depth Supervision. We also use a standard pixel-wise depth map loss. We quantize the real-valued depth and formulate the depth prediction as a classification problem instead of regression, and employ the cross-entropy loss. In particular we follow [9] to use the weighted cross-entropy loss (WCEL), with the weight being the information gain. See [9] for details.

To obtain the accurate depth map and recover high-quality 3D information, we combine WCEL and VNL together to supervise the network output. The overall loss is:

$$\ell = \ell_{WCE} + \lambda \ell_{VN}, \tag{3.7}$$

where λ is a trade-off parameter and is set to 5 in all experiments to make the two terms roughly of the same scale.

Note that the above overall loss function is differentiable. The gradient of the ℓ_{VN} loss can be easily computed as Eq. (3.4) and Eq. (3.6) are both differentiable.

3.3.2 Network Architecture

An architecture overview of our model is illustrated in Fig.3.6. The network is mainly composed of two parts, an encoder to establish features in different levels from I_{in}, and a decoder to reconstruct the depth map. Inspired by [75], the decoder is composed of several adaptive merging blocks (AMBs) to fuse features from different levels and dilated residual blocks (DRB) to transform features. In order to improve the receptive field of the decoder, we set the dilation rates of all 3×3 convolutions in DRB to 2 and insert an Astrous Spatial Pyramid Pooling (ASPP) module (dilation rate: 2, 4, 8) [12] between the encoder and the decoder. Furthermore, we establish 4 flip connections from different levels of encoder blocks to the decoder to merge more low-level features. The AMB will learn a merging parameter for adaptive merging. Apart from the features at the highest level with 512 channels, the feature dimension of other flips is 256. At last, a prediction module, a 3×3 convolution and a softmax, is utilized to transfer the features dimensions from 256 channels to 150 depth bins.

In the lightweight backbone network experiment, the backbone is replaced with MobileNetV2. In order to further reduce parameters, the dimensions of four flip connections are reduced to $(128, 64, 64, 64)$. In the prediction module, the features are transferred from 64 channels to 60 depth bins.

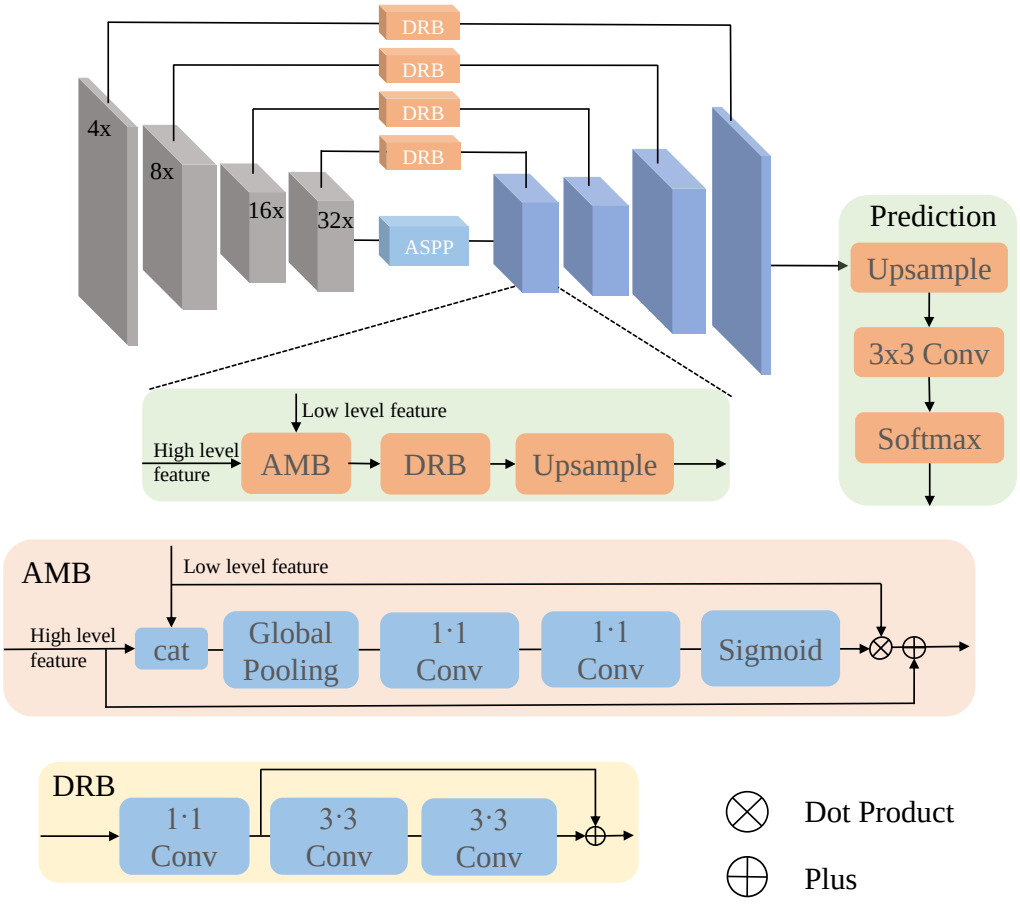


FIGURE 3.6. Model architecture. The encoder-decoder network has four flip connections to merge low-level features.

3.4 Experiments

In this section, we conduct several experiments to compare ours against state-of-the-art methods. We evaluate our methods on two datasets, NYUD-V2 and KITTI.

3.4.1 Datasets

NYUD-V2. The NYUD-V2 dataset consists of 464 different indoor scenes, which are further divided into 249 scenes for training and 215 for testing. We randomly sample 29K images from the training set to form NYUD-Large. Note that DORN uses the whole training set, which is significantly larger than that what we use. Apart from the whole dataset, there are officially annotated 1449 images (NYUD-Small), in which 795 images are split for training and others are for testing. In the ablation study, we use the NYUD-Small data.

KITTI. The KITTI dataset contains over 93K outdoor images and depth maps with the resolution around 1240×374. All images are captured on driving cars by stereo cameras and a Lidar. We test on 697 images from 29 scenes split by Eigen *et al.* [30], validate on 888 images, and train on about 23488 images from the remaining 32 scenes.

3.4.2 Implementation Details

The ResNeXt-101 [152] ($32 \times 4d$) model pre-trained on ImageNet [26] is used as our backbone model. A polynomial decaying method with the base learning rate 0.0001 and the power of 0.9 are applied in SGD. The weight decay and the momentum are set to 0.0005 and 0.9 respectively. The batch size is 8 in our experiments. The model is trained for 10 epochs on NYUD-Large and KITTI, and is trained for 40 epochs on NYUD-Small in the ablation study. We perform data augmentation on the training samples by the following methods. For NYUD-V2, the RGB image and the depth map are randomly resized with ratio $[1, 0.92, 0.86, 0.8, 0.75, 0.7, 0.67]$, randomly flipped in the horizon, and finally randomly cropped with the size 384×384 for NYUD-V2. The similar process is applied to KITTI but resizing with the ratio $[1, 1.1, 1.2, 1.3, 1.4, 1.5]$ and cropping with 384×512. Note that the depth map should be scaled with the corresponding resizing ratio.

3.4.3 Evaluation Metrics

We follow previous methods [68] to evaluate the performance of monocular depth prediction quantitatively based on following metrics: mean absolute relative error (rel), mean $\log_{10}$ error ($\log_{10}$), root mean squared error (rms) , root mean squared log error (rms (log)) and the accuracy under threshold ($\delta_i < 1.25^i, i = 1, 2, 3$).

3.4.4 Comparison with State-of-the-art

In this section, we detail the comparison of our methods with state-of-the-art methods.
NYUD-V2. In this experiment, we compare with other state-of-the-art methods on the NYUD-V2 dataset. Table 3.1 demonstrates that our proposed method outperforms other state-of-the-art methods across all evaluation metrics significantly. Compare to DORN, we have improved the accuracy from 0.2% to 18% over all evaluation metrics that they report.

In addition to the quantitative comparison, we demonstrate some visual results between our method and the state-of-the-art DORN in Fig. 3.7. Clearly, the predicted depth by the proposed method is much more accurate. The plane of ours is much smoother and has fewer errors (see the wall regions colored with red in the 1st, 2nd, and 3rd row). Furthermore, the last row in Fig. 3.7 manifests that our predicted depth is more accurate in the complicated scene. We have fewer errors in shelf and desk regions.
KITTI. In order to demonstrate that our proposed method can still reach the state-of-the-art performance on outdoor scenes, we test our method on the KITTI dataset. Results in Table 3.2 show that our method has outperformed all other methods on all evaluation metrics except root mean square (rms) error. The rms error is only slightly behind that of DORN. Note that for outdoor scenes, the rms (log) error, instead of rms, is usually the metric of interest, in which ours is better.

TABLE 3.1. Results on NYUD-V2. Our method outperforms other state-of-the-art methods over all evaluation metrics.

Method	**rel**	**log10**	**rms**	$\boldsymbol{\delta_1}$	$\boldsymbol{\delta_2}$	$\boldsymbol{\delta_3}$
	Lower is better			Higher is better		
Saxena *et al.*. [115]	0.349	-	1.214	0.447	0.745	0.897
Karsch *et al.*. [62]	0.349	0.131	1.21	-	-	-
Liu *et al.*. [84]	0.335	0.127	1.06	-	-	-
Ladicky *et al.*. [67]	-	-	-	0.542	0.829	0.941
Li *et al.*. [72]	0.232	0.094	0.821	0.621	0.886	0.968
Roy *et al.*. [107]	0.187	0.078	0.744	-	-	-
Liu *et al.*. [83]	0.213	0.087	0.759	0.650	0.906	0.974
Wang *et al.*. [139]	0.220	0.094	0.745	0.605	0.890	0.970
Eigen *et al.*. [29]	0.158	-	0.641	0.769	0.950	0.988
Chakrabarti [10]	0.149	-	0.620	0.806	0.958	0.987
Li *et al.*. [74]	0.143	0.063	0.635	0.788	0.958	0.991
Laina *et al.*. [68]	0.127	0.055	0.573	0.811	0.953	0.988
DORN [36]	0.115	0.051	0.509	0.828	0.965	0.992
DenseDepth [1]	0.123	0.053	0.465	0.846	0.974	0.994
DSN [103]	0.132	0.056	0.429	0.834	0.959	0.987
Chen *et al.* [18]	0.111	0.048	0.514	0.878	0.977	0.994
Huynh *et al.* [56]	0.108	-	**0.412**	**0.882**	**0.980**	**0.996**
Ours (ResNet101)	0.112	0.051	0.465	0.859	0.970	0.993
Ours (ResNeXt101)	**0.108**	**0.048**	**0.416**	**0.875**	**0.976**	**0.994**

TABLE 3.2. Results on KITTI. Our method outperforms other methods over all evaluation metrics except rms.

Method	$\boldsymbol{\delta_1}$	$\boldsymbol{\delta_2}$	$\boldsymbol{\delta_3}$	**rel**	**rms**	**rms (log)**
	Higher is better			Lower is better		
Make3D [115]	0.601	0.820	0.926	0.280	8.734	0.361
Eigen *et al.*. [30]	0.692	0.899	0.967	0.190	7.156	0.270
Liu *et al.*. [83]	0.647	0.882	0.961	0.114	4.935	0.206
Semi. [66]	0.862	0.960	0.986	0.113	4.621	0.189
Guo *et al.*. [44]	0.902	0.969	0.986	0.090	3.258	0.168
DORN [36]	0.932	0.984	0.994	0.072	**2.727**	0.120
DenseDepth [1]	0.886	0.965	0.986	0.093	4.170	0.171
DSN [103]	0.934	0.986	0.996	0.075	3.253	0.119
Ours	**0.938**	**0.990**	**0.998**	**0.072**	3.258	**0.117**

3.4.5 Ablation Studies

In this section, we conduct several ablation studies to analyze the details of our approach.

Effectiveness of VNL. In this study, in order to prove the effectiveness of the proposed VNL we compare it with two types of pixel-wise depth map supervision, a pair-wise geometric supervision, and a high-order geometric supervision: 1) the ordinary cross-entropy loss (CEL); 2) the L_1 loss (L_1); 3) the surface normal loss (SNL); 4) the pair-wise geometric loss (PL). We reconstruct the point cloud from the

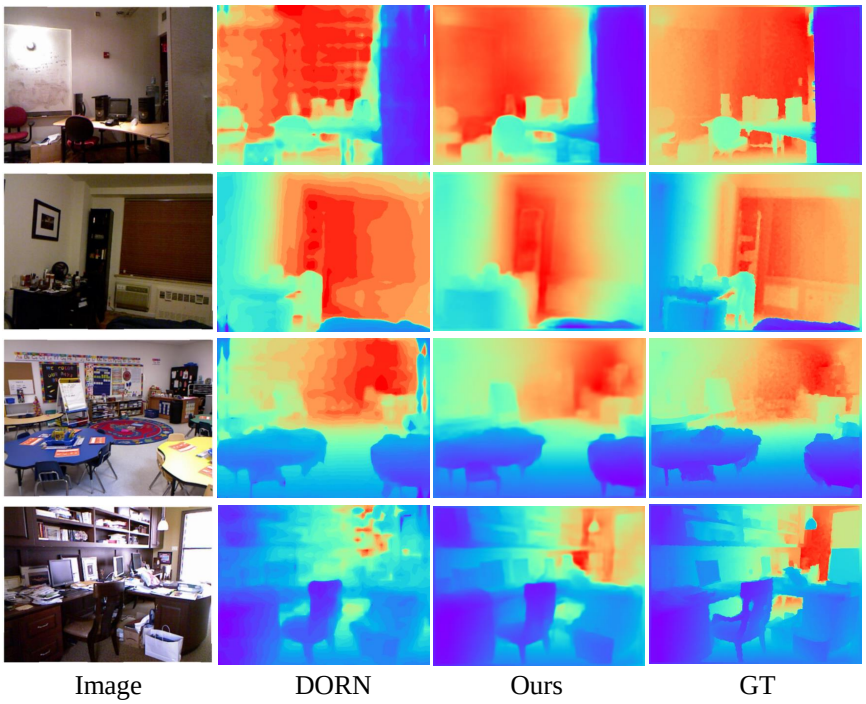


FIGURE 3.7. Examples of predicted depth maps by our method and the state-of-the-art DORN on NYUD-V2. Color indicates the depth (red is far, purple is close). Our predicted depth maps have fewer errors in planes (see walls) and have high-quality details in complicated scenes (see the desk and shelf in the last row)

.

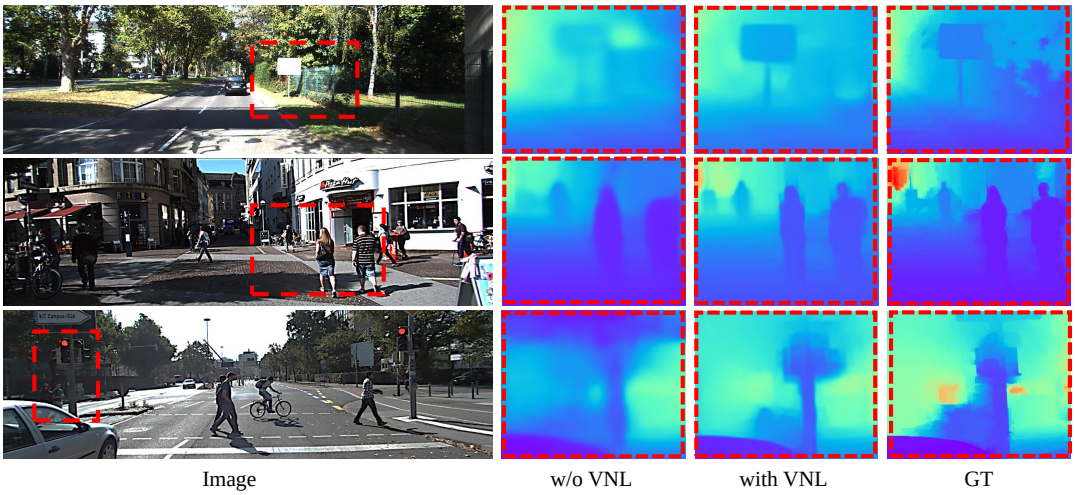


FIGURE 3.8. Examples of predicted depth on KITTI. Depth maps in the red dashed boxes with sign, pedestrian and traffic lights are zoomed in. One can see that with the help of virtual normal, predicted depth maps in these ambiguous regions are considerably more accurate.

depth map and further recover the surface normal from the point cloud. The angular discrepancy between the ground truth and recovered surface normal is defined as the surface normal loss, which is a high-order geometric supervision in 3D space. The pair-wise loss is the direction difference of two vectors in 3D, which are established by randomly sampling paired points in ground-truth and predicted point cloud. The

TABLE 3.3. Illustration of the effectiveness of VNL.

Metrics	**rel**	**log10**	**rms**	$\boldsymbol{\delta_1}$	$\boldsymbol{\delta_2}$	$\boldsymbol{\delta_3}$
Pixel-wise Depth Supervision						
CEL	0.1456	0.061	0.617	0.8087	0.9559	0.9862
WCEL	0.1427	0.060	0.511	0.8117	0.9611	0.9895
WCEL+L1	0.1429	0.061	0.626	0.8098	0.9539	0.9858
Pixel-wise Depth Supervision + Geometric Supervision						
WCEL+PL‡	0.1380	0.059	0.504	0.8212	0.9643	0.9913
WCEL+PL+VNL	0.1341	0.056	0.485	**0.8336**	**0.9671**	0.9913
WCEL+SNL†	0.1406	0.059	0.599	0.8209	0.9602	0.9886
WCEL+VNL‡ (Ours)	**0.1337**	**0.056**	**0.480**	0.8323	0.9669	**0.9920**

† 'Local' geometric supervision in 3D.
‡ 'Global' geometric supervision in 3D.

loss function of PL is as follows:

$$\ell_{PL} = \frac{1}{N}\sum_{i=0}^{N}(1 - \frac{\overrightarrow{P^*_{Ai}P^*_{Bi}} \cdot \overrightarrow{P_{Ai}P_{Bi}}}{\left\|\overrightarrow{P^*_{Ai}P^*_{Bi}}\right\| \cdot \left\|\overrightarrow{P_{Ai}P_{Bi}}\right\|}) \tag{3.8}$$

where $(P^*_A, P^*_B)_i$ and $(P_A, P_B)_i$ are paired points sampled from the ground truth and the predicted point cloud respectively and N is the total number of pairs.

We also employ the long-range restriction $\mathcal{R}_2$ for the paired points. Therefore, similar to VNL, PL can also be seen as a global geometric supervision in 3D space. The experimental results are reported in Table. 3.3. WCEL is the baseline for all following experiments.

Firstly, we analyze the effect of pixel-wise depth supervision for prediction performance. As WCE employs a weight in the CE loss, its performance is slightly better than that of CEL. However, when we enforce two pixel-wise supervision (WCEL+L1) on the depth map, the performance cannot improve any more. Thus using two pixel-wise loss terms does not help.

Secondly, we analyze the effectiveness of the supplementary 3D geometric constraint (PL, SNL, VNL). Compared with the baseline (WCEL), three supplementary 3D geometric constraints can promote the network performance with varying degrees. Our proposed VNL combining with WCEL has the best performance, which has improved the baseline performance by up to 8%.

Thirdly, we analyze the difference of three geometric constraints. As SNL can only exploit geometric relations of homogeneous local regions, its performance is the lowest among the three constraints over all evaluation metrics. Compared with SNL, since PL constrains the global geometric relations, its performance is clearly better. However, the performance of WCEL+PL is not as good as our proposed WCEL+VNL. When we further add our VNL on top of WCEL+PL, the precision can further be slightly improved and is comparable to WCEL+VNL. Therefore, although PL is a global geometric constraint in 3D, the pair-wise constraint cannot encode as strong geometry information as our proposed VNL.

At last, in order to further demonstrate the effectiveness of VNL, we analyze the

TABLE 3.4. Performance on NYUD-V2 with MobileNetV2 backbone. †Trained without VN. ‡Trained with VN.

Metrics	CReaM [127]	RF-LW[93]	Ours-B†	Ours-VN‡
$\boldsymbol{\delta_1}$	0.704	0.790	0.814	**0.829**
$\boldsymbol{\delta_2}$	0.917	0.955	0.947	**0.956**
$\boldsymbol{\delta_3}$	0.977	**0.990**	0.972	0.980
rel	0.190	0.149	0.144	**0.134**
rms	0.687	0.565	0.502	**0.485**
rms (log)	0.251	0.205	0.201	**0.185**
params	**1.5M**	3.0M	2.7M	2.7M

results of network trained with and without VNL supervision on the KITTI dataset. The visual comparison is shown in Fig. 3.8. One can see that VNL can improve the performance of the network in ambiguous regions. For example, the sign (1st row), the distant pedestrian (2nd row), and traffic light in the last row of the figure can demonstrate the effectiveness of the proposed VNL.

In conclusion, the geometric constraints in the 3D space can significantly boost the network performance. Moreover, the global and high-order constraints can enforce stronger supervision than the 'local' and pair-wise ones in 3D space.

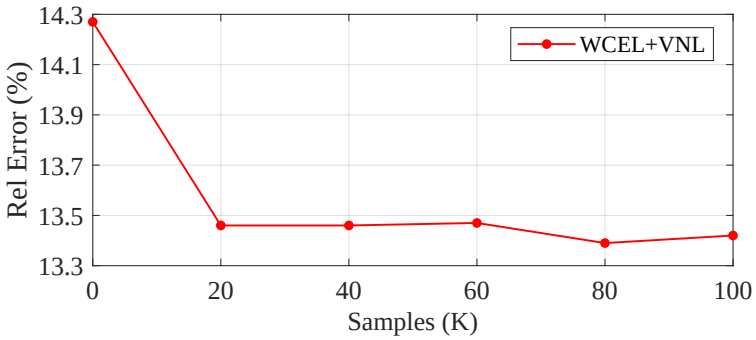


FIGURE 3.9. Illustration of the impact of the samples size. The more samples will promote the performance.

Impact of the Amount of Samples. Previously, we have proved the effectiveness of VNL. Here the impact of the size of samples for VNL is discussed. We sample six different sizes of point groups, 0K, 20K, 40K, 60K, and 80K and 100K, to establish VNL. '0K' means that the model is trained without VNL supervision. The rel error is reported for evaluation. Fig. 3.9 demonstrates that 'rel' slumps by 5.6% with 20K point groups to establish VNL. However, it only drops slightly when the samples for VNL increase from 20K to 100K. Therefore, the performance saturates with more samples, when samples reach a certain number in that the diversity of samples is enough to construct the global geometric constraint.

Lightweight Backbone Network.

We train the network with the MobileNetV2 backbone to evaluate the effectiveness of the proposed geometric constraint on the light network. We train it on the NYUD-Large for 10 epochs. Results in Table 3.4 show that the proposed VNL can improve

FIGURE 3.10. Comparison of reconstructed point clouds from estimated depth maps between DORN [36] and ours. We can see that our point cloud results contain less noise and are closer to groud-truth than that of DORN.

the performance by 1% - 8%. Comparing with previous state-of-the-art methods, we have improved the accuracy by around 29% over all evaluation metrics and achieved a better trade-off between parameters and the accuracy.

3.4.6 Recovering 3D Features from Estimated Depth

We have argued that, with geometric constraints in the 3D space, the network can achieve more accurate depth and also obtain higher-quality 3D information. Here we show the recovered 3D point cloud and the surface normal to support this.

3D Point Cloud. Firstly, we compare the reconstructed 3D point cloud from our predicted depth and that of DORN. Fig. 3.10 demonstrate that the overall quality of ours outperforms theirs significantly. Although our predicted depth is only slightly better than theirs on evaluation metrics, the reconstructed wall (see the 2nd row in 3.10) of ours is much flatter and has fewer errors. The shape of the bed is more similar to the ground truth. From the bird view, it is hard to recognize the bed shape of their results. The point cloud in Fig. 3.1 also leads to a similar conclusion.

Surface Normal.

The surface normal is another important information to describe the geometry of a scene. Normal prediction from the monocular image is also a long-lasting problem. Previous methods mainly design a separate network or decoder to estimate the surface normal. Actually, the surface normal can be directly calculated from the 3D point cloud. Therefore, the surface normal can also demonstrate the quality of the predicted depth. We compare the calculated surface normal with previous state-of-the-art methods and demonstrate the quantitative results in Table 3.5. The ground

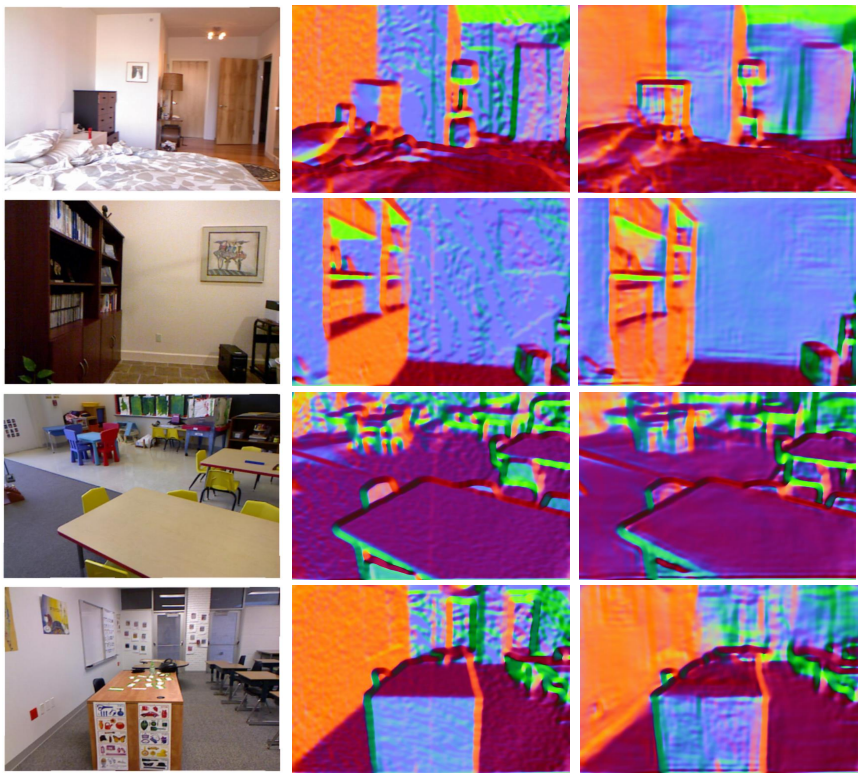


FIGURE 3.11. Recovered surface normal from 3D point cloud. According to the visual effect, the surface normal is in high-quality in planes (1st row) and the complicated curved surface (2nd and last row).

truth is obtained as described in [29]. We first compare our geometrically calculated results with DCNN-based optimization methods. Although we do not optimize a sub-model to achieve the surface normal, our results can outperform most of such methods and even are the best on **30°** metric.

TABLE 3.5. Evaluation of the surface normal on NYUD-V2.

Method	**Mean**	**Median**	**11.2°**	**22.5°**	**30°**
	Lower is better		Higher is better		
Predicted Surface Normal from the Network					
3DP [33]	33.0	28.3	18.8	40.7	52.4
Ladicky *et al.*. [166]	35.5	25.5	24.0	45.6	55.9
Fouhey *et al.*. [34]	35.2	17.9	40.5	54.1	58.9
Wang *et al.*. [142]	28.8	17.9	35.2	57.1	65.5
Eigen *et al.*. [29]	**23.7**	**15.5**	**39.2**	**62.0**	71.1
Calculated Surface Normal from the Point cloud					
GT-GeoNet† [101]	36.8	32.1	15.0	34.5	46.7
DORN‡ [36]	36.6	31.1	15.7	36.5	49.4
Ours	24.6	17.9	34.1	60.7	**71.7**

† Cited from the original paper.
‡ Using authors' released models.

Furthermore, we compare the surface normals directly computed from the reconstructed point cloud with that of DORN [36] and GeoNet [101]. Note that we run the released code and model of DORN to obtain depth maps and then calculate surface normals from the depth, while the evaluation of GeoNet is cited from the original paper. In Table 3.5, we can see that, with high-order geometric supervision, our method outperforms DORN and GeoNet by a large margin, and even is close to Eigen method which trains to output normals. It suggests that our method can lead the model to learn the shape from images.

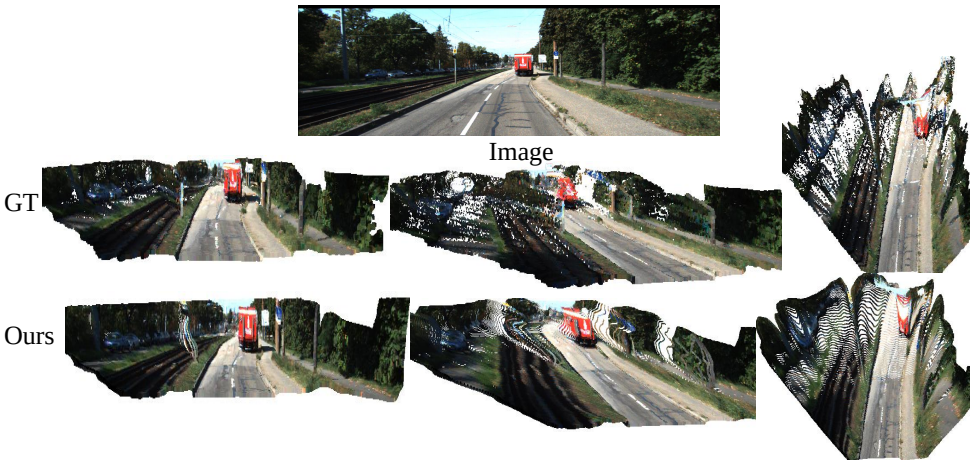
Apart from the quantitative comparison, the visual effect is shown in Fig. 3.11, demonstrating that our directly calculated surface normals are not only accurate in planes (the 1st row), but also are of higher quality in regions with sophisticated curved surface (the 2nd and last row).

3.4.7 3D point cloud

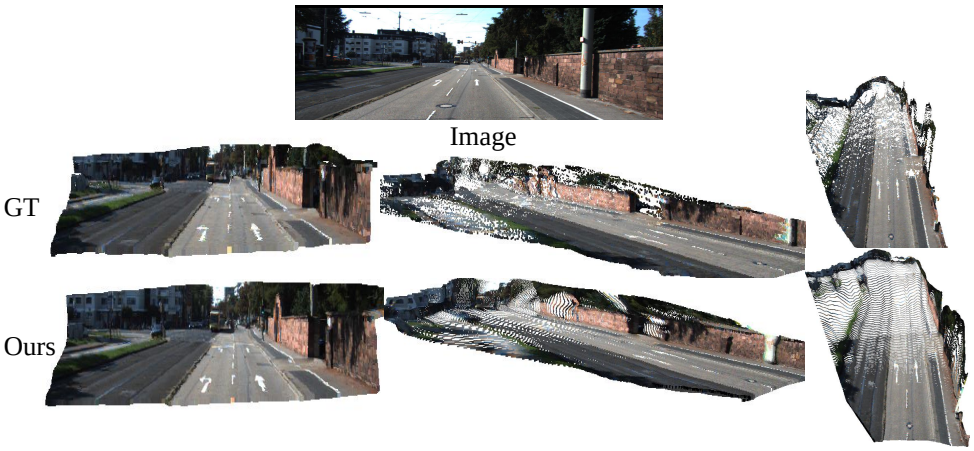
In order to further show the quality of reconstructed point cloud from the predicted depth, we randomly select 3 scenes from the testing part of NYUD-V2 and KITTI. 3 views are randomly selected to display the reconstructed point cloud. The results are shown in Fig. 3.12.

3.5 Conclusion

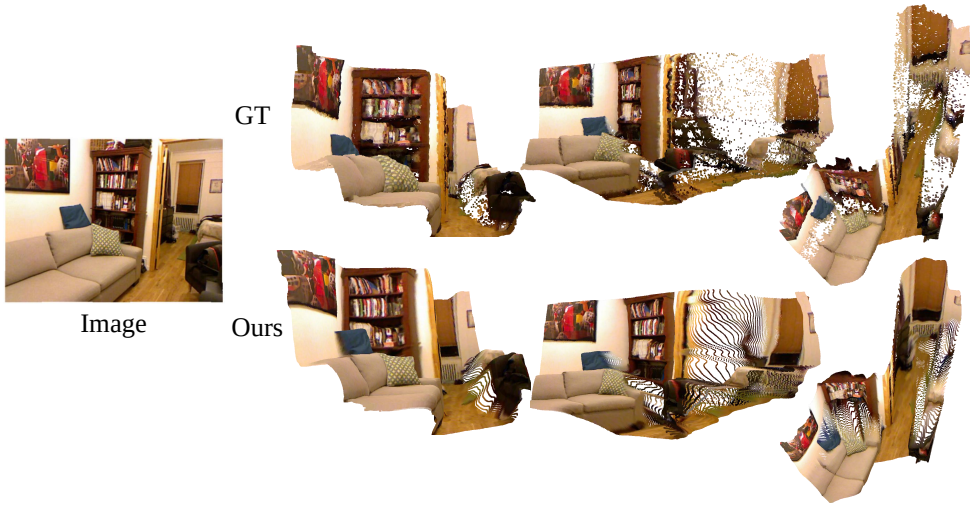
In this paper, we have proposed to construct a high-order global geometric constraint (VNL) in the 3D space for monocular depth prediction. In contrast to previous methods with only pixel-wise depth supervision in 2D space, our method cannot only obtain



(A)



(B)



(C)

FIGURE 3.12. Reconstructed point clouds. Three scenes are randomly selected from NYUD-V2 and KITTI. For the reconstructed point cloud of each scene, 3 views are selected to demonstrate the point cloud. The first column is the RGB image. The last 3 columns of are different views of the reconstructed point for each scene. (a) Scene 1; (b) Scene 2; (c) Scene 3.

the accurate depth maps but also recover high-quality 3D features, such as the point cloud and the surface normal, eliminating necessities to optimize a new sub-model. Compared with other 3D constrains, our proposed VNL is more robust to noise and can encode strong global constraints. Experimental results on NYUD-V2 and KITTI have proved the effectiveness of our method and the state-of-the-art performance.

Statement of Authorship

Title of Paper	Diversedepth: Affine-invariant depth prediction using diverse data
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Yin Wei, Wang Xinlong, Yifan Liu, Chunhua Shen, Tian Zhi, Sun Changming, Xu Songcun, Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. arXiv preprint arXiv:2002.00569. 2020

Principal Author

Name of Principal Author (Candidate)	Wei Yin
Contribution to the Paper	Design new methods and conduct the experiments.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/13/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chunhua Shen
Contribution to the Paper	Discussion and review the paper.
Signature	10/13/2021

Name of Co-Author	Yifan Liu
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Zhi Tian		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/11/2021

Name of Co-Author	Songcun Xu		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/11/2021

Name of Co-Author	Renyin Dou		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/09/2021

Please cut and paste additional co-author pane

Name of Co-Author	Xinlong Wang		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/12/2021

Please cut and paste additional co-author pa

Name of Co-Author	Changming Sun		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	13/10/2021

Statement of Authorship

Title of Paper	Virtual Normal: Enforcing Geometric Constraints for Accurate and Robust Depth Prediction
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Yin, Wei, Yifan Liu, Chunhua Shen, Virtual Normal: Enforcing Geometric Constraints for Accurate and Robust Depth Prediction, IEEE transactions on pattern analysis and machine intelligence, 2021,

Principal Author

Name of Principal Author (Candidate)	Wei Yin
Contribution to the Paper	Design new methods and conduct the experiments.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/11/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chunhua Shen
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Name of Co-Author	Yifan Liu
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Please cut and paste additional co-author panels here as required.

Chapter 4

Affine Invariant Depth Estimation

4.1 Introduction

In the previous chapter, we have demonstrated that the proposed global geometry constraint, i.e. virtual normal loss, can significantly boost the monocular metric depth estimation and improve the geometry quality of the predicted depth. However, learning metric depth on a small dataset cannot generate a robust depth model to generalize to in-the-wild scenes.

In this chapter, we aims to solve the generalization issue of monocular depth estimation. We analyze existing learning metric depth and learning relative depth methods in detail. Although current learning relative depth methods can produce a robust model, the predicted relative depth can only represent relative depth relations, i.e. one point is farther or closer than another one, but at the loss of geometry information. To solve these problems, we propose a method to construct large-scale and diverse RGBD dataset, and propose to learn affine-invariant depth on the dataset.

4.2 Background

Monocular depth estimation is a challenging problem. As there exists no easy way to enforce geometric constraints to recover the depth from a still image, various data-driven approaches are proposed to exploit comprehensive cues [36, 161, 25, 8].

Previous methods of depth estimation based on deep convolutional neural networks (DCNN) have achieved outstanding performance on popular benchmarks [161, 36, 1]. They can be mainly summarized into two categories. (1) The first group enforces the pixel-wise metric supervision to produce the accurate metric depth map typically on some specific scenes, such as indoor environments, but in general does not work well on diverse scenes. For example, the recent virtual normal method of [161] can achieve state-of-the-art performances on various benchmarks with the training and testing done on each benchmark separately. (2) The second group aims to address the issue of generalization to multiple scene data by learning with relative depth such that large-scale datasets of diverse scenes can be collected much easier. A typical example is the depth-in-the-wild (DIW) dataset [17]. Such methods often explore the pair-wise ordinal relations for learning and only the relative depth can be predicted.

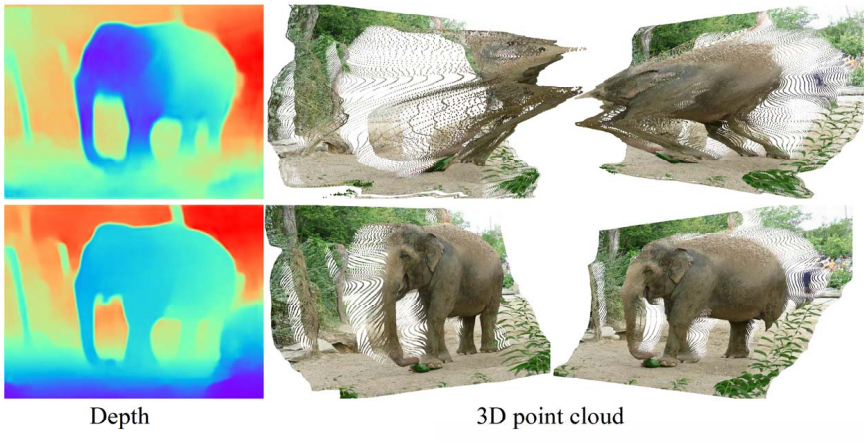


FIGURE 4.1. Qualitative comparison of depth and reconstructed 3D point cloud between our method and that of the recent learning relative depth method of Xian *et al.* [149]. The first row is the predicted depth and reconstructed 3D point cloud from the depth of theirs, while the second row is ours. The relative depth model fails to recover the 3D geometric shape of the scene (see the distorted elephant and ground area). Ours does much better. Note that this test image is sampled from the DIW dataset, which does not overlap with our training data.

A clear drawback is that these models fail to recover the high-quality geometric 3D shapes as only ordinal relations are used in learning. For example, the reconstructed 3D point cloud from the relative depth (first row in Figure 4.1) is completely distorted and cannot represent the shape of the elephant.

In order to ensure both good generalization and high-quality 3D depth information, there are two obstacles: (1) lacking diverse and high-quality training data; (2) an appropriate learning objective function that is easy to optimize, yet preserving as much geometric information as possible.

In this work, we seek to address these problems from three aspects: (1) constructing a large-scale dataset with diverse scenes, Diverse Scene Depth dataset (DiverseDepth), including both rigid and non-rigid contents in both indoor and outdoor environments. With our proposed dataset construction method, such a dataset can be relatively easy to expand. Existing datasets are either difficult to expand (metric depth), or only annotated with weak geometric information (relative depth). Our dataset strikes a balance between these. (2) enforcing the DCNN model to learn the affine-invariant depth instead of a specific depth value or relative depth on the diverse scales dataset; (3) proposing a multi-curriculum learning method for the effective training on this complex dataset. Current available RGB-D datasets can be summarized into two categories: (1) RGB-depth pairs captured by a depth sensor have high precision, typically accommodating only few scenes as it can be very costly to acquire a very large dataset of diverse scenes. For example, the KITTI dataset [41] is captured with LIDAR on road scenes only, while the NYU dataset [124] only contains several indoor rooms. (2) Images with much more diverse scenes are available online and can be annotated with coarse depth with reasonable effort. The large-scale DIW [17]

dataset is manually annotated with only one pair of ordinal depth relations for each image. To construct our large and diverse dataset, we harvest stereoscopic videos and images with diverse contents and use stereo matching methods to obtain depth maps. The dataset contains both rigid and non-rigid foregrounds, such as people, animals, and cars. Our DiverseDepth is more diverse than metric depth datasets, while it contains more geometric information than existing relative depth datasets because depth in our dataset is metric depth up to an affine transformation. We have sampled some images from Taskonomy [165] and DIML [23] and added them into our dataset.

The commonly used learning objectives can be summarized into two categories: (1) directly minimizing the pixel-wise divergence to the ground-truth metric depth [36, 30, 161]; (2) exploiting the uniformity of pair-wise ordinal relations [17, 149]. However, both two methods cannot well balance the high generalization and enriching the model with abundant geometric information. In contrast, we reduce the difficulty of depth prediction by explicitly disentangling depth scales during training. The model will ignore the depth scales and make the predicted depth invariant to the affine transformation, (i.e., translation, scale). Several loss functions can satisfy the requirement. For example, the surface normal and virtual normal loss [161] are affine-invariant because they are based on normals. Besides, the scale-and-shift-invariant loss (SSIL) [70] explicitly recovers the scaling and shifting gap between the predicted and ground-truth depth. We combine a geometric constraint and SSIL to supervise the model. The second row in Figure 4.1 shows the predicted affine-invariant depth of a DIW image and the reconstructed 3D point cloud, which can clearly represent the shape of the elephant and ground. Experiments on 8 zero-shot datasets show the effectiveness of learning affine-invariant depth.

Furthermore, training the model on the large-scale and diverse dataset effectively is also a problem. We propose a multi-curriculum learning method for training. Hacohen and Weinshall [46] have proved that an easy-to-hard curriculum will not change the global minimum of the optimization function but increase the learning speed and improve the final performance on test data. Here, we separate the diverse learning materials to different curriculums and introduce each curriculum with increasing difficulty to the network. Experiments show this method can significantly promote the performance on various scenes.

In conclusion, our contributions are outlined as follows.

- We construct a large scale and high-diversity RGB-D dataset, DiverseDepth;
- We are the first to propose to learn affine-invariant depth on the diverse dataset, which ensures both high generalization and high-quality geometric shapes of scenes;
- We propose a multi-curriculum learning method to effectively train the model on the large-scale and diverse dataset. Experiments on 8 zero-shot datasets show our method outperforms previous methods noticeably.

TABLE 4.1. Comparison with previous RGB-D datasets. Our dataset features both diverse scenes and high-quality ground-truth depth.

Dataset	Diversity	Dense	Accuracy	Images
Captured by RGB-D sensor				
NYU [124]	Low	✓	High	407K
KITTI [41]	Low	✓	High	93K
SUN-RGBD [126]	Low	✓	High	10K
ScanNet [25]	Low	✓	High	2.5M
Make3D [114]	Low	✓	High	534
Taskonomy [165]	Low	✓	High	4.5M
DIML [23]	Low	✓	High	2M
DIODE [135]	Low	✓	High	26K
Crawled online				
DIW [17]	High		Low	496K
Youtube3D [14]	High		Low	794K
RedWeb [149]	Medium	✓	Medium	3.6K
WSVD [137]	Medium	✓	low	1.5M
MegaDepth [77]	Medium	✓	Medium	130K
Ours	High	✓	Medium	320K

4.3 Method

4.3.1 Diverse Scene Depth Dataset Construction

Dataset statistics. Table 4.1 compares the released popular RGB-D datasets. RGB-D sensors can capture high-precision depth data, but they only contain limited scenes. By contrast, crawling large-scale online images can promote scene diversity. Previous datasets only have sparse ordinal depth annotations, such as DIW and Youtube3D. Although RedWeb and MegaDepth advance the ground-truth depth quality, RedWeb only has 3600 images and MegaDepth only contains static scenes.

Therefore, to feature diversity, quality, as well as data size, we construct a new dataset with multiple data collection sources. Firstly, we collect large-scale online stereo images and videos to construct foreground objects part, termed *Part-fore*, such as plants, people and animals. Then, we sample some images from Taskonomy and DIML to constitute the indoor and outdoor background part, termed *Part-in* and *Part-out*.

Foreground part. The processes of *Part-fore* data construction is outlined as follows.

(1) Crawling online stereoscopic images and videos. We summarize three websites for data collection: Flickr, 3DStreaming and YouTube. We firstly discard invalid frames and images that are not left/right stereo contents by comparing the similarity of left/right parts, i.e., removing low similarity frames. Then we manually inspect outliers.

(2) Retrieving disparities from stereo materials, then reversing and scaling them to obtain depths. As parameters of all stereo cameras are unknown, we cannot rectify

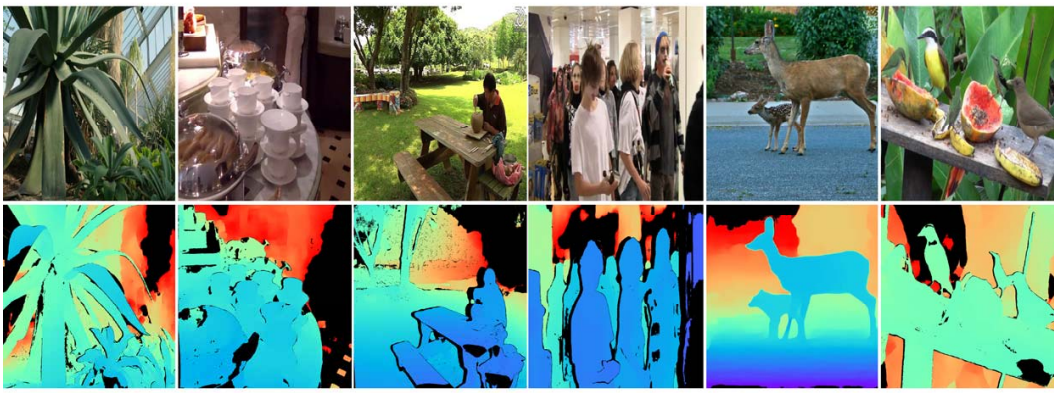


FIGURE 4.2. Examples of the DiverseDepth dataset. Purple parts are closer, while red regions are farther.

the stereo images, remove lens distortion, and align the epipolar line. Existing stereo matching methods [51] based on comparing the local or semi-global features along the epipolar line cannot obtain the disparity. Instead, we utilize the optical flow [57] method to match the paired pixels in stereo samples and take the horizontal matching as the disparity. The depth is obtained by reversing and scaling the disparity.

(3) Filtering depth maps. We find that many outliers and noises residing in depths are mainly caused by large distortions, small baselines, and poor images features. Here, we take 3 metrics to mask out such noises. Firstly, pixels with vertical disparities larger than 5 are removed. Secondly, pixels with the left-right disparity difference greater than 2 are removed. Furthermore, images with valid pixels less than 30% are discarded. After these filtering processes, We totally collect more than 90K RGB-D pairs for the *Part-fore*. Example images are shown in Figure 4.2.

Background part. In order to enrich the diverse background environments, we sample 100K images from an indoor and an outdoor dataset respectively, i.e., Taskonomy [165] and DIML [23]. The Taskonomy samples constitute our indoor background data, *Part-in*, while the DIML ones are the outdoor background part, *Part-out*.

Therefore, our DiverseDepth dataset has around 300K diverse RGB-D pairs, which is composed of three different parts, i.e., *Part-fore*, *Part-in* and *Part-out*. There are around 18K images for testing.

4.3.2 Affine-invariant Depth Prediction

The geometric model of the monocular depth estimation system is illustrated in Figure 4.3. The ground-truth object in the scene is A^*, and the real camera system is $O\text{-}XYZ$ (the black one in Figure 4.3). When learning the metric depth, the model $\mathcal{G}(\mathbf{I}, \theta)$ may predict the object at location A. $\mathbf{I}$ is the input image. The learning objective of such methods is to minimize the divergence between A and A^*, i.e., $\min_\theta |\mathcal{G}(\mathbf{I}, \theta) - d^*|$, where d^* is the ground-truth depth and θ is the network parameters. As such methods mainly train and test the model on the same benchmark, where the camera system and the scale remain almost the same, the model can implicitly

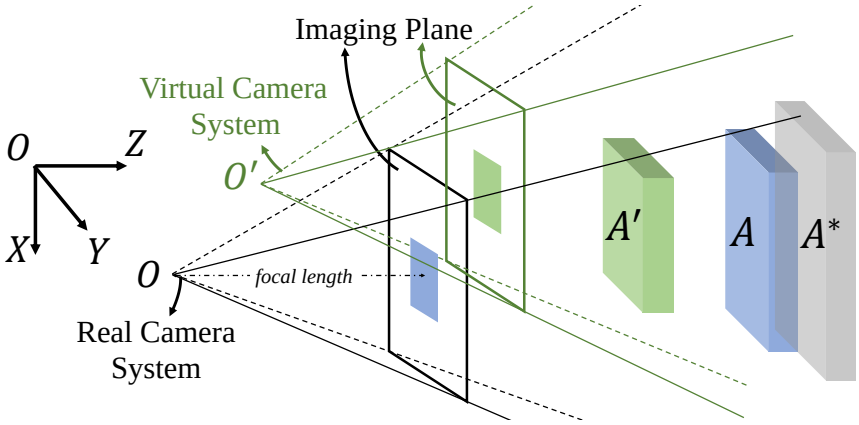


FIGURE 4.3. The geometric model of an imaging system. A^* is the ground-truth location for an object. A is the predicted location by learning metric depth method, while A' is the predicted location by our learning affine-invariant depth method.

learn the camera system and produce accurate depth on the testing data [28]. The typical loss functions for learning metric depth are illustrated in Table 4.2. However, when training and testing on the diverse dataset, where the camera system and scale vary, it is theoretically impossible for the model to accommodate multiple camera parameters. The tractable approach is to feed camera parameters of different camera systems to the network as part of the input in order to predict metric depth. This requires the access to camera parameters, which are often unavailable when harvesting online image data. Our experiments show failure cases of learning metric depth on the diverse dataset (see Table 4.3, Table 4.6, and Figure 4.4).

Learning the relative depth reduces the difficulty of depth prediction from predicting the accurate metric depth to the ordinal relations. With enough diverse training data, this method can predict relative depth on diverse scenes, but it loses geometric information of the scene, such as the geometric shape. For example, the reconstructed 3D point cloud from the relative depth in Figure 4.7 and Figure 4.1 cannot represent the shape of the sofa and elephant respectively.

In this paper, we propose to learn the affine-invariant depth from the diverse dataset. On the diverse dataset, we define a virtual camera system, O'-$X'Y'Z'$ (the green one in Figure 4.3), which has the same viewpoint as the real one but has the different optical center location and the focal length. Therefore, there is an affine transformation, i.e., translation T and scaling s, between the real camera system O-XYZ and the virtual one O'-$X'Y'Z'$. For the predicted depth under the virtual camera system, it has to take an affine transformation to recover the metric depth under the real camera system, i.e., $P_A = s \cdot (P_{A'} + T)$, where $P = (x, y, d)^T$. The learning objective is defined as follows.

$$L = \min_{\theta} |\mathcal{K}(\mathcal{G}(\mathbf{I}, \theta)) - d^*| \tag{4.1}$$

where $\mathcal{K}(\cdot)$ is the affine transformation to recover the scaling and translation. Through

TABLE 4.2. Illustration of different loss functions

Loss	Definition
Metric Depth Loss	
MSE	$L_{mse} = \frac{1}{N}\sum_{i=1}^{N}(d_i - d_i^*)^2$
Silog [30]	$L_{si} = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \frac{1}{N^2}(\sum_{i=1}^{N} y_i)^2$ $y_i = \log(\frac{d_i}{d_i^*})$
Relative Depth Loss	
Ranking [149]	$L_{rank} = \begin{cases} \log(1+\exp(d_i - d_j)l_{ij}), l_{ij} = \pm 1 \\ (d_i - d_j)^2, l_{ij} = 0 \end{cases}$
Affine-invariant Depth Loss	
Scale-shift-invariant [70]	$L_{ssi} = \frac{1}{2N}\sum_{i=1}^{N}(\overrightarrow{\mathbf{d}_i}^{\top}\mathbf{h} - d_i^*)^2$ $\mathbf{h} = (\sum_{i=1}^{N}\overrightarrow{\mathbf{d}_i}\overrightarrow{\mathbf{d}_i}^{\top})^{-1}(\sum_{i=1}^{N}\overrightarrow{\mathbf{d}_i}\mathbf{d}_i^*)$, $\overrightarrow{\mathbf{d}_i} = (d_i, 1)^{\top}$
Virtual normal [161]	$L_{vn} = \frac{1}{N}(\sum_{i=0}^{N}\|\mathbf{n}_i - \mathbf{n}_i^*\|_1)$ $\mathbf{n}$ is the virtual normal
Surface normal [101]	$L_{sn} = \frac{1}{N}(\sum_{i=0}^{N}\|\mathbf{n'}_i - \mathbf{n'}_i^*\|_1)$ $\mathbf{n'}$ is the surface normal

explicitly defining a virtual camera system and disentangling the affine transformation between the diverse real camera system and the virtual one, we simplify the objective of monocular depth prediction. The predicted depth will be invariant to various scales and translations. Therefore, it will be easier to generalize to diverse scenes by learning affine-invariant depth than metric depth. Besides, such learning objective can maintain more geometric information than that of learning relative depth.

In Table 4.2, several losses can disentangle the scaling and translation and enforce the model to learn the affine-invariant depth. The virtual normal loss (VNL) and surface normal loss [101, 161] are constructed based on the normals, which are essentially invariable to scaling and translation. Furthermore, the scale-and-shift-invariant loss (SSIL) [70] explicitly recovers the scaling and translation before minimizing the divergence to ground truth. Therefore, we take the high-order geometric loss and the SSIL to optimize the network. The overall loss function is illustrated as follows. λ is to balance the two terms, where λ is set to 1 in our experiments.

$$\ell = L_{vn}(d, d^*) + \lambda L_{ssi}(d, d^*) \tag{4.2}$$

4.3.3 Multi-curriculum Learning

Most existing methods uniformly sample a sequence of mini-batches $[\mathbb{B}_0, ..., \mathbb{B}_M]$ from the whole dataset for training. However, as our DiverseDepth has a wide range of scenes, experiments illustrate that such training paradigm cannot effectively optimize the network. We propose a multi-curriculum learning method to solve this problem.

Algorithm 1: multi-curriculum learning algorithm

Input : scoring function $\mathcal{F}$, pacing function $\mathcal{H}$, dataset $\mathbb{X}$
Output: mini-batches sequence $\{\mathbb{B}_i | i = 0 \dots M\}$.

train the model $\mathcal{G}_j$ on the data part $\mathbb{D}_j$ as the teacher
sort each data part $\mathbb{D}_j$ with ascending difficulty according to $\mathcal{F}$, the ranked data is $\mathbb{C}_j$
for $k = 0$ **to** K **do**
    for $i = 0$ **to** M **do**
        for $j = 0$ **to** P **do**
            subset size $s_{kj} = \mathcal{H}(k, j)$
            subset $\mathbb{S}_{kj} = \mathbb{C}_j[0, \dots, s_{kj}]$
            uniformly sample batch $\mathbb{B}_{ij}$ from $\mathbb{S}_{kj}$
        end
        concatenate P batches sampled from different data parts together $\mathbb{B}_i = \{\mathbb{B}_{ij}\}_{j=0}^{P}$
        append $\mathbb{B}_i$ to the mini-batches sequence
    end
end

We sort the training data by the increasing difficulty and sample a series of mini-batches that exhibit an increasing level of difficulty. Therefore, there are two problems that should be solved: 1) how to construct the curriculum; 2) how to yield a sequence of easy-to-hard mini-batches for the network. Pseudo-code for multi-curriculum algorithm is shown in Algorithm 1.

Constructing the curriculum. Three parts of DiverseDepth, i.e., *part-fore, part-in* and *part-out*, are termed as $\mathbb{X} = \{\mathbb{D}_j\}_{j=0}^{P}$. Let $\mathbb{D}_j = \{(x_{ij}, y_{ij}) | i = 0, ..., N\}$ represents the N data points of the part j, where x_{ij} denotes a single data, y_{ij} is the corresponding label. Previous monocular depth estimation methods show that training on limited scenes are easy to converge, so we train three models, $\mathcal{G}_j$, separately on 3 parts as teachers. The absolute relative error (Abs-Rel) is chosen as the *scoring function* $\mathcal{F}(\cdot)$ to evaluate the difficulty of each training sample. If $\mathcal{F}(\mathcal{G}_j(x_{ij}), y_{ij}) > \mathcal{F}(\mathcal{G}_j(x_{(i+1)j}), y_{(i+1)j})$, then we define the data (x_{ij}, y_{ij}) is more difficult to learn. Finally, we sort 3 parts according to the ascending Abs-Rel error and the ranked datasets are $\mathbb{C}_j = \{(x_{ij}, y_{ij}) | i = 0, ..., N\}$.

Mini-batch sampling. The *pacing function* $\mathcal{H}(\cdot)$ determines a sequence of subsets of the dataset so that the likelihood of the easier data will decrease in this sequence, i.e. $\{\mathbb{S}_{0j}, \dots, \mathbb{S}_{Kj}\} \subseteq \mathbb{C}_j$, where $\mathbb{S}_{kj}$ represents the first $\mathcal{H}(k, j)$ elements of $\mathbb{C}_j$. From each subset $\mathbb{S}_{kj}$, a sequence of mini-batches $\{\mathbb{B}_{0j}, ..., \mathbb{B}_{Mj} | j = 0, 1, 2\}$ are uniformly sampled. Here we utilize the stair-case function as the *pacing function*, which is determined by the starting sampling percentage p_j, the current *step* k, and the fixed *step length* I_o (the number of iterations in each step). In each *step* k, there are I_o iterations and the $\mathcal{H}(k, j)$ remains constant, thus the step $k = \left\lfloor \frac{iter}{I_o} \right\rfloor$, where *iter* is the iteration index. $\mathcal{H}(k, j)$ is defined as follows.

$$\mathcal{H}(k, j) = min(p_j \cdot k, 1) \cdot N_j \tag{4.3}$$

where N_j is the size of part $\mathbb{D}_j$.

4.4 Experiments

In order to demonstrate the generalization and effectiveness of our method, we test our method quantitatively and qualitatively on several zero-shot datasets and compare it with other state-of-the-art methods.

Experiment setup. We test on 8 zero-shot datasets to illustrate the performance and generalization of our method, i.e., NYU [124], KITTI [41], DIW [17], ETH3D [118], ScanNet [25], TUM-RGBD [129], DiverseDepth-H-Realsense, and DiverseDepth-H-SIMU. The last two testing datasets are constructed by us to test the performance on scenes with foreground people. We use two different RGB-D sensors, Realsense and SIMU, to capture people in several indoor and outdoor scenes. DiverseDepth-H-Realsense contains 2329 images, while DiverseDepth-H-SIMU has 8685 images. We use the model used in [161] with the pre-trained ResNeXt-50 [152] backbone. The SGD is utilized for optimization with the initial learning rate of 0.0005 for all layers. The learning rate is decayed every 5K iterations with the ratio 0.9. The batch size is set to 12. Note that we evenly sample images from three data parts to constitute a batch. During the training, images are flipped horizontally, resized with the ratio from 0.5 to 1.5, and cropped with the size of 385×385. In the testing, we will resize, pad, and crop the image to keep a similar aspect ratio.

Evaluation metrics. We mainly take the absolute relative error (Abs-Rel) for evaluation except DIW, which is evaluated with the Weighted Human Disagreement Rate (WHDR) [149]. Besides, when evaluating the depth of foreground people, we follow the approach in [78] to take scale-invariant root mean squared error (Si-RMS) and Abs-Rel for evaluation. As our model can only predict the affine-invariant depth of the scene, we explicitly scale and translate the depth to recover the metric depth when evaluating the metric depth. The scaling and translation factors are obtained by the least-squares method.

4.4.1 Comparison with State-of-the-art Methods

TABLE 4.3. The comparison with state-of-the-art methods on five zero-shot datasets. Our method outperforms previous learning the relative depth or metric depth methods significantly.

Method	Training dataset	Backbone	Testing on zero-shot datasets				
			DIW WHDR	NYU	KITTI	ETH3D Abs-Rel	ScanNet
Learning Metric Depth + Single-scene Dataset							
Yin *et al.* [161]	NYU	ResneXt-101	27.0	10.8	35.1	29.6	13.66
Alhashim *et al.* [1]	NYU	DenseNet-169	26.8	12.3	33.4	34.5	12.5
Yin *et al.* [161]	KITTI	ResneXt-101	30.8	26.7	7.2	31.8	23.5
Alhashim *et al.* [1]	KITTI	DenseNet-169	30.9	23.5	9.3	32.1	20.5
Learning Relative Depth + Diverse-scene Dataset							
Li and Snavely [77]	MegaDepth	ResNet-50	24.6	19.1	19.3	29.0	18.3
Lasinger *et al.* [70]	MV + MegaDepth + RedWeb	ResNet-50	14.7	19.1	29.0	23.3	15.8
Chen *et al.* [17]	DIW	ResNeXt-50	11.5	16.7	25.6	25.7	16.0
Xian *et al.* [149]	RedWeb	ResNeXt-50	21.0	26.6	44.4	39.0	18.2
Learning Affine-invariant Depth + Diverse-scene Dataset							
Ours	DiverseDepth	ResNeXt-50	**14.3**	**11.7**	**12.6**	**22.5**	**10.4**

'__' The method has trained the model on the corresponding dataset.

Quantitative comparison on popular benchmarks. The quantitative comparison is illustrated in Table 4.3. Apart from Chen *et al.* [17] and Xian *et al.* [149], whose performance is retrieved by re-implementing the ranking loss and training with our model, the performances of other methods are obtained by running their released codes and models. For all methods, we scale and translate the depth before evaluation. Those results whose models have been trained on the testing scene are marked with an underline.

Firstly, from Table 4.3, we can see that previous state-of-the-art methods, which enforce the model to learn accurate metric depth, cannot generalize well to other scenes. For example, the well-trained models of Yin *et al.* [161] and Alhashim and Wonka [1] cannot perform well on other zero-shot scenes.

Secondly, although learning the relative depth methods can predict high-quality ordinal relations on the diverse DIW dataset, i.e., one point being closer or further than another one, the discrepancy between the relative depth and the ground-truth metric depth is very large, see Abs-Rel on other datasets. Such high Abs-Rel results in these methods are not able to recover high-quality 3D shape of scenes, see Figure 4.1 and Figure 4.7.

By contrast, through enforcing the model to learn the affine-invariant depth and constructing a high-quality diverse dataset for training, our method can predict high-quality depths on various zero-shot scenes. Our method can outperform previous methods by up to 70%. Noticeably, on NYU, our performance is even on par with existing state-of-the-art methods which have trained on NYU (ours 11.7% vs. Alhashim 12.3%).

Qualitative comparison on zero-shot datasets. Figure 4.4 illustrates the qualitative comparison on five zero-shot datasets. The transparent white masks denote the method has trained the model on the corresponding dataset. We can see those learning metric depth methods, Yin *et al.* [161] and Alhashim and Wonka [1], cannot work well on unseen scenes, while learning relative depth methods, see Lasinger *et al.*, cannot recover high-quality depth map, especially for distant regions (see the marked regions on KITTI, NYU, and ScanNet) and regions with high texture difference (see marked head and colorful wall on DIW). On the DIW dataset, our method can predict more accurate depth on diverse DIW scenes, such as the forest and sign. Furthermore, on popular benchmarks, such as ScanNet, KITTI, and NYU, our method can also produce more accurate depth maps.

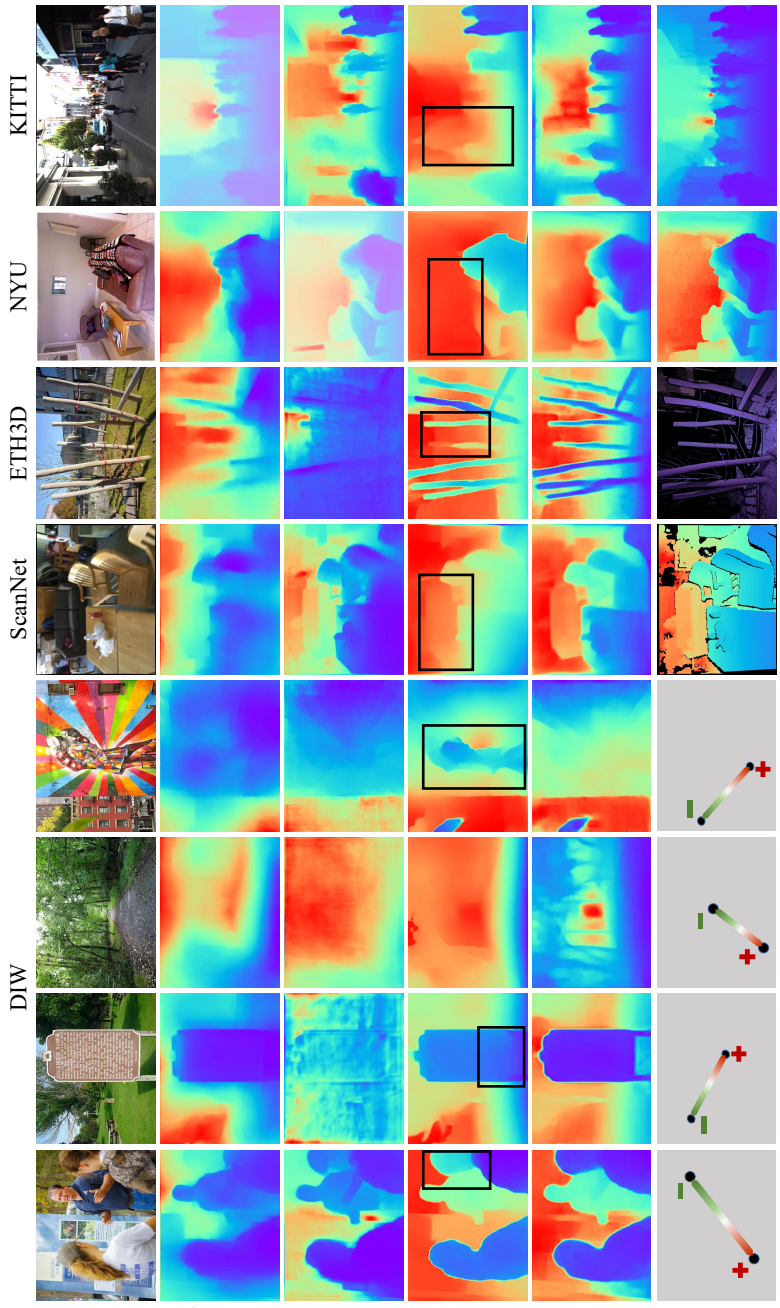


FIGURE 4.4. Qualitative comparison with state-of-the-art methods on zero-shot datasets. The transparent masks on images denote the method has been trained on the corresponding testing data. The black rectangles highlight the comparison regions. Our method not only predicts more accurate depth on diverse DIW, but also recovers better details on indoor and outdoor scenes, see marked regions on ScanNet, ETH3D, NYU, and KITTI. Note that ground truth of DIW only annotates the ordinal relation between two points.

TABLE 4.4. The performance comparison of the foreground people on three zero-shot datasets. Our method can predict more accurate depth on foreground people over three datasets.

Method	Training	Si-hum	Si-env	Si-RMS	Abs-Rel
Testing on TUM-RGBD					
Li-I	MC	0.294	0.334	0.318	0.204
Li-IFCM†	MC	0.302	0.330	0.316	0.206
Li-IDCM†	MC	0.293	**0.238**	0.272	**0.147**
Ours	DiverseDepth	**0.272**	0.270	**0.272**	0.192
Testing on DiverseDepth-H-Realsense					
Li-I	MC	0.343	0.305	0.319	0.264
Ours	DiverseDepth	**0.262**	**0.241**	**0.261**	**0.186**
Testing on DiverseDepth-H-SIMU					
Li-I	MC	0.373	0.466	0.419	0.391
Ours	DiverseDepth	**0.283**	**0.335**	**0.309**	**0.218**

† Input the mask of people and depth of background to the model.

Comparison of people. 'People' is a significant foreground content for various applications. To our best knowledge, Li *et al.* [78] are the first ones to focus on depth estimation for people. To promote the performance, they have to input the pre-computed depth of the background from two consecutive frames with structure from motion [116] and the mask of people regions to the network, see Li-IFCM and Li-IDCM in Table 4.4. Li-I denotes the method with a single image input for the network. By contrast, our method can also predict the high-quality depth for people with a still image. We make comparison on three datasets, i.e., TUM-RGBD, DiverseDepth-H-Realsense, and DiverseDepth-H-SIMU.

In Table 4.4, Si-env and Si-hum denote the Si-RMS errors of the background and people, respectively. On TUM-RGBD, our method outperforms three configurations of Li *et al.* [78] on foreground people up to 10%. Our overall performance, Si-RMS, is also much better. As Li-IDCM inputs the depth of the background, its Si-env error is lower than ours.

DiverseDepth-H-Realsense and DiverseDepth-H-SIMU have more scenes than TUM-RGBD. We compare our method with Li-I. It is clear that our method outperforms theirs significantly over all metrics with a still image.

Furthermore, we randomly select several images for qualitative comparison (see Figure 4.5). It is clear that Li-I cannot perform well on the bottom part of people, distant people, and regions with significant texture difference, while our method can predict much better depths on both people and the background.

4.4.2 Ablation Study

In this section, we carry out several experiments to analyze the effectiveness of the proposed multi-curriculum learning method, the effectiveness of different loss functions

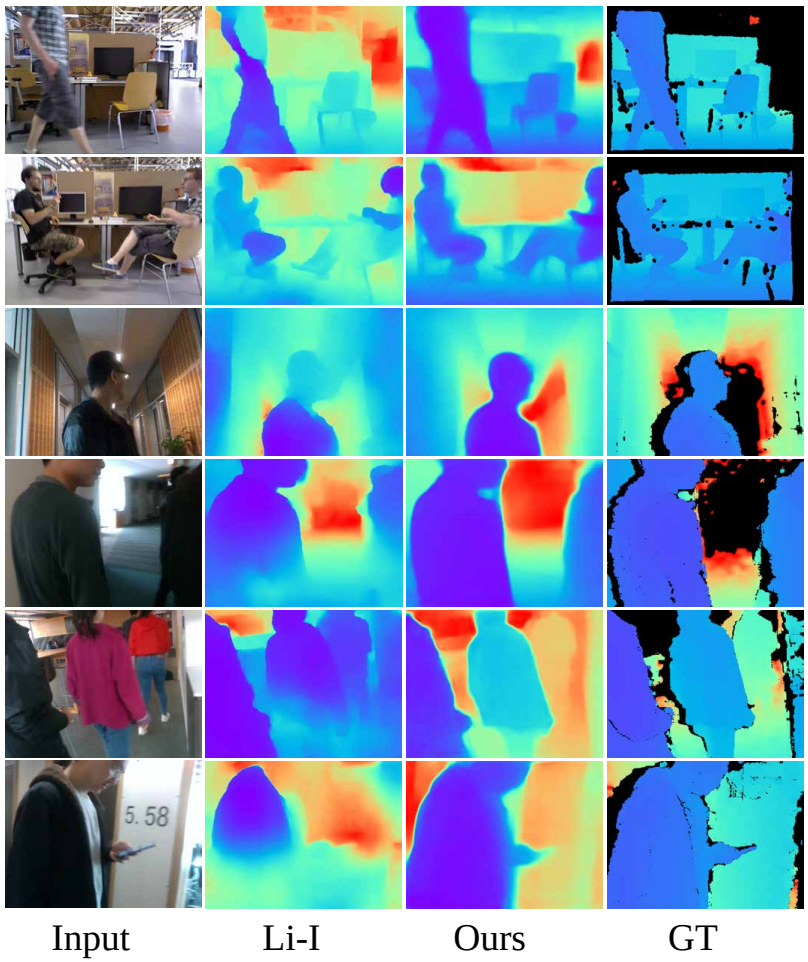


FIGURE 4.5. Qualitative comparison of the foreground people. Our method and Li *et al.* [78]-I have a single image input for the network. Our method can predict better depth on people and the background environments.

on the diverse data, the comparison of the reconstructed 3D point cloud among different methods, and the linear relations between the predicted affine-invariant depth and GT.

Effectiveness of multi-curriculum learning. To demonstrate the effectiveness of multi-curriculum learning method, we take three settings for the comparison: (1) sampling a sequence of mini-batches uniformly for training, termed Baseline; (2) using the reverse *scoring function*, i.e., $\mathcal{F}^{'} = -\mathcal{F}$, thus the training samples are sorted in the descending order on difficulty and the harder examples are sampled more than easier ones, termed MCL-R; (3) using the proposed multi-curriculum learning method for training, termed MCL. We make comparisons on 5 zero-shot datasets and our proposed DiverseDepth dataset. In Table 4.5, it is clear that MCL outperforms the baseline by a large margin over all testing datasets. Although MCL-R can also promote the performance, it cannot achieve the comparable performance as MCL. Furthermore, we demonstrate the validation error along the training in Figure 4.6. It is clear that the validation error of MCL is always lower than the baseline and MCL-R over the whole training process. Therefore, the MCL method with an easy-to-hard curriculum

TABLE 4.5. The comparison of different training methods on 5 zero-shot datasets and our DiverseDepth dataset. The proposed multi-curriculum learning method outperforms the baseline noticeably, while MCL-R can also promote the performance.

Method	DIW†	NYU†	KITTI†	ETH3D†	ScanNet†	DiverseDepth	
	WHDR	Abs-Rel				Abs-Rel	WHDR
Baseline	14.5	11.7	17.9	26.1	11.2	26.0	16.4
MCL-R	15.0	11.8	15.8	24.7	11.0	24.4	15.9
MCL	**14.3**	**11.7**	**12.6**	**22.5**	**10.4**	**20.6**	**15.0**

† Testing on zero-shot datasets.

can effectively train the model on diverse datasets.

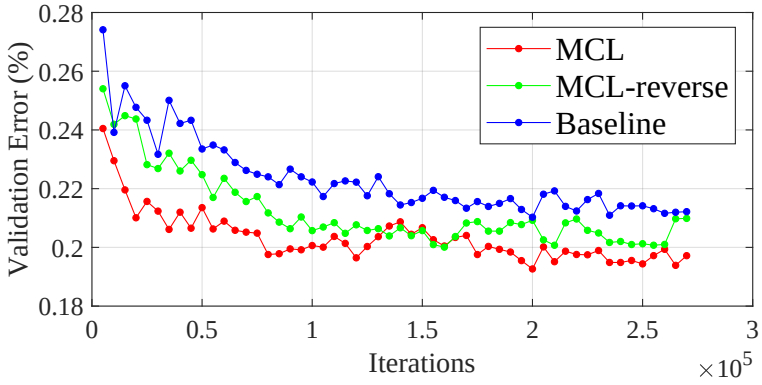


FIGURE 4.6. Validation error during the training process. The validation error of the proposed multi-curriculum learning method is always lower than that of the MCL-R and baseline.

Effects of different losses. In this section, we analyze the effectiveness of various loss functions for depth estimation on diverse datasets, including virtual normal loss (VNL), scale-shift-invariant loss (SSIL), Silog [30], Ranking, and MSE. We sample 10K images from each part of DiverseDepth separately for quick training then test the performances on 5 zero-shot datasets. All the experiments take a multi-curriculum learning method. In Table 4.6, the VNL and SSIL outperform others over five zero-shot datasets significantly, which demonstrates the effectiveness of learning the affine-invariant depth on diverse datasets. By contrast, as MSE loss enforces the network to learn the accurate metric depth, it fails to generalize to unseen scenes, thus cannot perform well on zero-shot datasets. Although Ranking can make the model predict good relative depth on diverse DIW, Abs-Rel errors are very high on other datasets because it cannot enrich model with any geometric information. By contrast, as Silog considers the varying scale in the dataset, it performs a little better than Ranking and MSE.

Comparison of the recovered 3D shape. In order to further demonstrate learning affine-invariant depth can maintain the geometric information, we reconstruct the 3D point cloud from the predicted depth of a random ScanNet image. We compare our methods with Lasinger *et al.* [70] and Yin-NYU [161]. We take four viewpoints for visual comparison, i.e., front, up, left, and right viewpoints. From Figure 4.7, it is clear that our reconstructed point cloud can clearly represent the shape of the sofa

TABLE 4.6. The effectiveness comparison of different losses on zero-shot datasets. VNL and SSIL outperform others noticeably. By contrast, the model supervised by MSE fails to generalize to diverse scenes, while Ranking can only enforce the model to learn the relative depth. Although Silog considers the varying scale in the dataset, its performance cannot equal VNL and SSIL.

Loss	Testing on zero-shot datasets				
	DIW	NYU	KITTI	ETH3D	ScanNet
VNL+SSIL (Ours)	**14.3**	**11.7**	**12.6**	**22.5**	**10.4**
VNL	15.2	12.2	21.0	28.9	11.5
SSIL	17.5	16.5	16.3	26.8	15.6
Silog	19.6	20.8	30.8	29.4	17.6
Ranking	24.3	23.4	47.9	39.5	18.1
MSE	35.3	33.2	36.0	30.2	21.6

and the wall from four views, while the sofa shapes of the other two methods are distorted noticeably and the wall is not flat.

Illustration of the affine transformation relation. To illustrate the affine transformation between the predicted affine-invariant depth and the ground-truth metric depth, we randomly select two images from KITTI and NYU respectively, and uniformly sample around 15K points from each image. The predicted depth has been scaled and translated for visualization. In Figure 4.8, the red line is the ideal linear relation, while the blue points are the sampled points. We can see the ground-truth depth and the predicted depth have a roughly linear relation. Note that as the precision of the sensor declines with the increase of depth, as expected.

Test on in-the-wild Scenes. To demonstrate the robustness of our methods on in-the-wild scenes, we test our predicted depths and reconstructed point cloud on several in-the-wild scenes. We capture some high-resolution images by a mobile phone, and the predicted affine-invariant depth results are shown in Fig. 4.9. We can see that the depth maps are of high quality. Edges and occlusion boundaries are clear. Furthermore, we collect several outdoor scenes to evaluate the quality of reconstructed 3D point clouds. Note that the focal length is provided for the reconstruction. Results are illustrated in Fig. 4.10.

4.5 Conclusion

We have proposed methods to solve the generalization issue of monocular depth estimation, at the same time maintaining as much geometric information as possible. Firstly, we construct a large-scale and highly diverse RGB-D dataset. Compared with previous diverse datasets, which only have sparse depth ordinal annotations, our dataset is annotated with dense and high-quality depth. Besides, we have proposed methods to learn the affine-invariant depth on our DiverseDepth dataset, which can ensure both good generalization and high-quality geometric shape reconstruction from the depth. In order to enable learning affine-invariant depth, we propose the high-order geometric loss, namely, virtual normal loss, which is more robust to noise

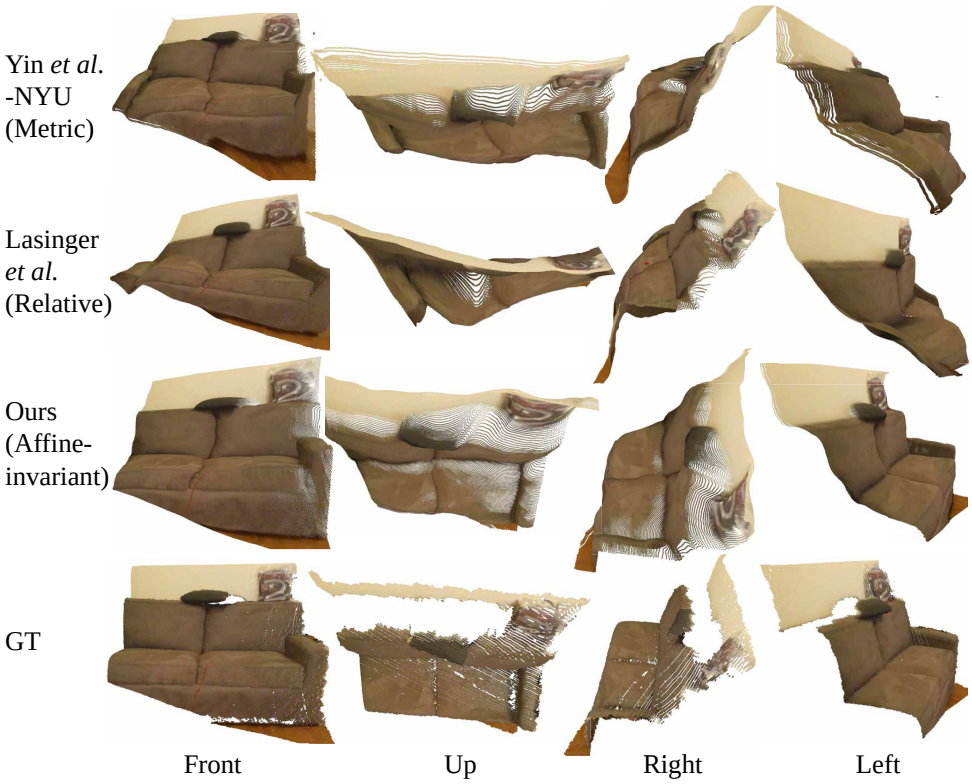
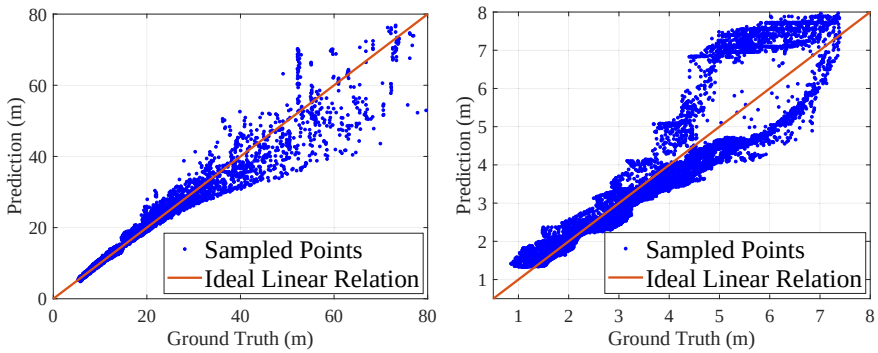


FIGURE 4.7. Qualitative comparison of the reconstructed 3D point cloud from the predicted depth of a ScanNet image. Our method can clearly recover the shapes of the sofa and wall, while the shape of other methods distort noticeably.

.



(A) (B)

FIGURE 4.8. Testing the linear relation between the ground-truth and predicted depth. (a) Testing on KITTI. (b) Testing on NYU. Predicted depth has been scaled and translated for visualization. Blue points are the sampled points, while the red line is the ideal linear relation. There is an approximately linear relation between the ground-truth and predicted depth.

and enables learning high-quality shapes from a single image. Furthermore, we propose a multi-curriculum learning method to train the model effectively on this diverse

FIGURE 4.9. Testing on high-resolution images captured by a phone.

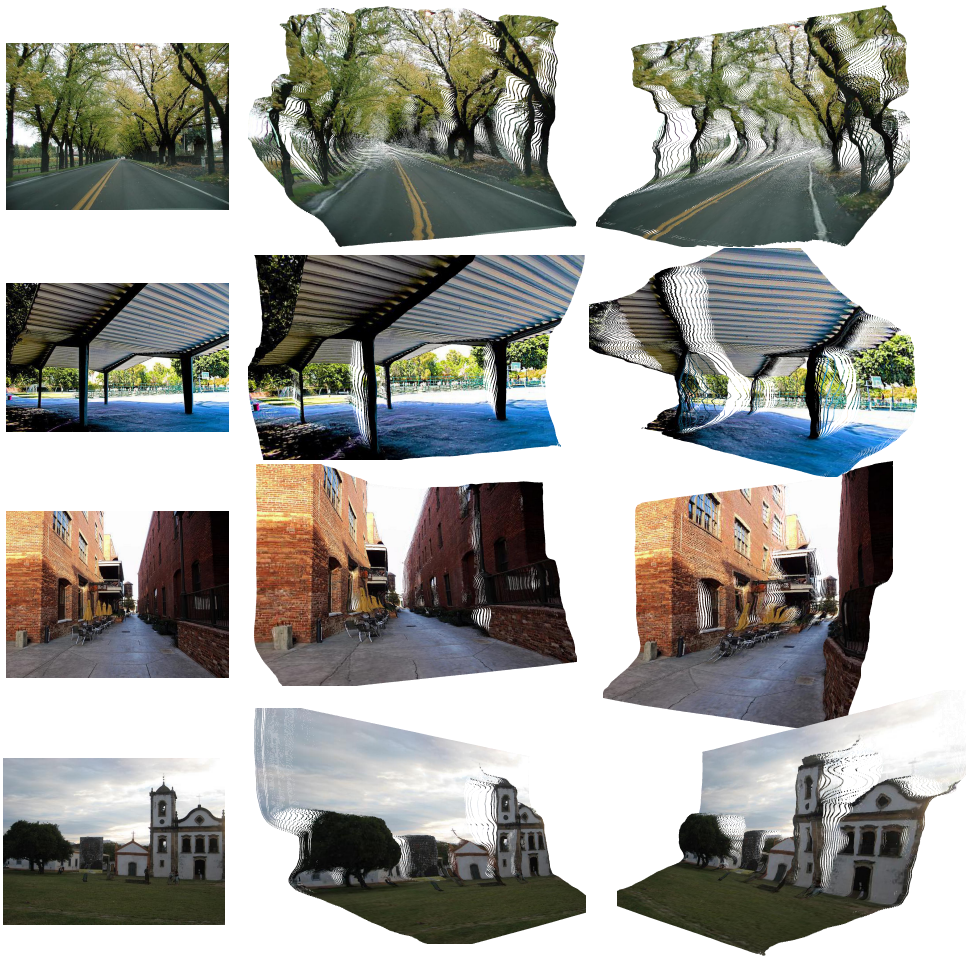


FIGURE 4.10. Reconstructing the 3D point cloud of some in-the-wild scenes.

dataset. Experiments on NYU and KITTI have demonstrated the effectiveness of virtual normal loss for monocular depth estimation. Besides, experimental results on 8 unseen datasets have shown the usefulness of our dataset for learning affine-invariant depth on diverse scenes.

Statement of Authorship

Title of Paper	Learning to recover 3d scene shape from a single image
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., & Shen, C. (2021). Learning to recover 3d scene shape from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 204-213).

Principal Author

Name of Principal Author (Candidate)	Wei Yin
Contribution to the Paper	Design new methods and conduct the experiments.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/13/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chunhua Shen
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Name of Co-Author	Jianming Zhang
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Oliver Wang		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/13/2021

Name of Co-Author	Long Mai		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10-13-2021

Name of Co-Author	Simon Niklaus		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	October 13, 2021

Name of Co-Author	Simon Chen		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/13/2021

Statement of Authorship

Title of Paper	Towards Accurate Reconstruction of 3D SceneShape from Single Monocular Images
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Long Mai, Chunhua Shen, IEEE transactions on pattern analysis and machine intelligence, (2021). .

Principal Author

Name of Principal Author (Candidate)	Wei Yin		
Contribution to the Paper	Design new methods and conduct the experiments.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	10/13/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chunhua Shen		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/13/2021

Name of Co-Author	Jianming Zhang		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/13/2021

Please cut and paste additional co-auth... panels here as required.

Name of Co-Author	Oliver Wang		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/13/2021

Name of Co-Author	Simon Niklaus		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	October 13, 2021

Name of Co-Author	Simon Chen		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/13/2021

Chapter 5

3D Scene Reconstruction from a Monocular Image

5.1 Introduction

In the last chapter, to solve the generalization problem of the monocular depth estimation, we propose to learn the affine-invariant depth on large-scale and diverse RGBD datasets. With the proposed curriculum learning training method, we can obtain a robust depth prediction model.

In this chapter, we go a step further to solve the scene reconstruction from a single image input. The camera intrinsic parameters estimation is indispensable for the reconstruction. As large-scale paired image to focal length datasets are unavailable, we propose to learn the focal length on synthetic 3D point cloud data. Furthermore, an unknown shift resides in the predicted affine-invariant depth, which will cause the reconstructed point cloud distortion. We propose to rectify such distortion and predict focal length from the point cloud.

5.2 Background

3D scene reconstruction is a fundamental task in computer vision. The established approach to address this task is multi-view geometry [47], which reconstructs 3D scenes based on feature-point correspondence with consecutive frames or multiple views. In contrast, we aim to achieve *dense 3D scene shape reconstruction from a single in-the-wild image.* Without multiple views available, we rely on monocular depth estimation. However, as shown in Fig. 5.1, existing monocular depth estimation methods [149, 81, 161] alone are unable to faithfully recover an accurate 3D point cloud. The key challenges are: 1) it is difficult to collect large-scale metric depth datasets with diverse scenes, which are needed to achieve good monocular depth estimation models; 2) alternatively, one can train models on large-scale *relative* depth datasets which are much easier to collect. We discover that learning depth on such datasets requires to estimate the depth shift and focal length in order to generate accurate 3D scene shapes. This problem was almost not studied in the literature, and we attempt to tackle this problem here.

Recent works have shown great progress by training deep neural networks on diverse in-the-wild data, *e.g.*, web stereo images and stereo videos [17, 16, 104, 137, 149, 150, 160, 159]. Chen *et al.* [17] propose the first large-scale and in-the-wild dataset, termed DIW. Each image only provides a pair of points and annotates their depth relations, *i.e.*, one is farther or closer than the other one. Xian *et al.* [149] propose to collect diverse web stereo images and use optical flow for finding pixel correspondence so as to create dense *relative* ground-truth depth because camera parameters are unknown and differ for each pair of stereo images.

However, web stereo images and videos can only provide depth supervision up to a scale and shift due to the unknown camera baselines and stereoscopic post-processing [69]. Moreover, the diversity of the training data also poses challenges for the model training, as training data captured by different cameras can exhibit significantly different image priors for depth estimation [31].

As a result, state-of-the-art in-the-wild monocular depth estimation models use various types of objective functions that are invariant to scale and shift to facilitate training. While an unknown scale in depth does not cause scene shape distortion, as it scales the 3D scene shape uniformly, an unknown depth shift does (see Sec. 5.3.1. As shown in Fig. 5.1, the walls are not flat because of the unknown shift). In addition, the camera focal length of a given image may not be accessible at test time, leading to more distortion of the 3D scene shape (see the angle between two walls of "Recovered shift" in Fig. 5.1). This scene shape distortion is a critical problem for downstream tasks such as 3D view synthesis and 3D photography.

To address these challenges, we propose a novel two-stage monocular scene shape estimation framework that consists of 1) a depth prediction module; and 2) a point cloud reconstruction module. The depth prediction module is a convolutional network trained on a mix of existing datasets that predicts depth maps up to a scale and shift. The point cloud reconstruction module leverages point cloud encoder networks that predict shift and focal length adjustment factors from an initial guess of the scene point cloud reconstruction. A key observation that we make here is that, *when operating on point clouds derived from depth maps, and not on images themselves, we can train models to learn 3D scene shapes using synthetic 3D data or data acquired by 3D laser scanning devices. The domain gap is significantly less of an issue for point clouds than that for images.* We empirically show that the point cloud network generalizes well to unseen datasets. Moreover, as two modules can be trained separately, we do not need the paired "RGB-Point Cloud" training data.

To obtain a robust model, we propose to mix multiple sources of data for training, including high-quality Lidar sensor data, medium-quality calibrated stereo data, and low-quality web stereo data. Considering the quality difference, we propose to distinguish them and use heterogeneous losses instead of an uniform form. For example, the low-quality data can only provide reliable depth ordinal relations, thus the ranking loss [149, 150] is applied. Other data have more accurate depths, but cameras are various. The training schedule on multiple heterogeneous data sources can have an impact on

the final performance. We propose a simple yet effective normalized regression loss for high-quality and medium-quality data. It transforms the depth data to a canonical scale-shift-invariant space for more robust training. Furthermore, to improve the geometry quality of the depth, we propose a pair-wise normal regression loss, which can account for both local and global geometry constraints. From high-quality data, reliable local normal information and global plane relations can be extracted, while other data can only provide co-plane information from semantics. Explicitly using these relations can significantly improve the depth quality.

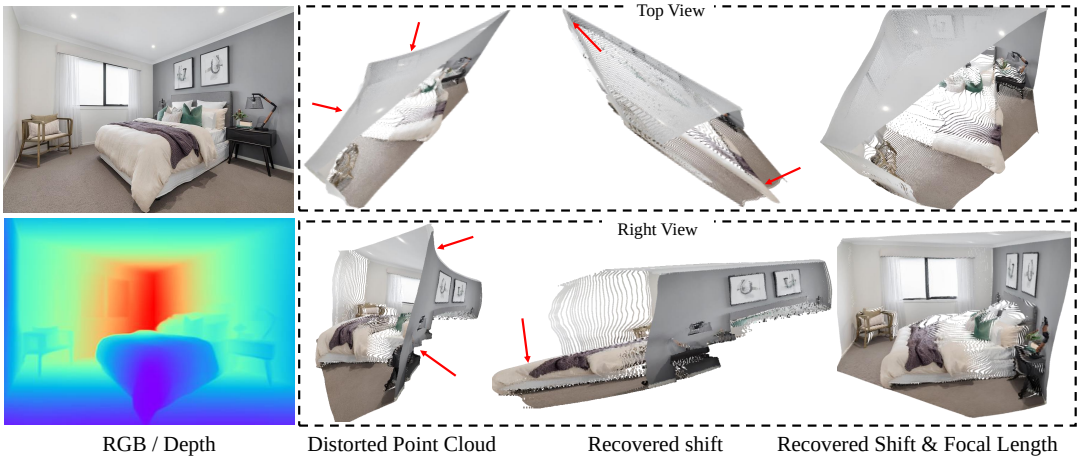


FIGURE 5.1. 3D scene structure distortion of projected point clouds. While the predicted depth map appears very good, the 3D scene shape of the point cloud suffers from noticeable distortions due to an unknown depth shift and focal length (2nd column). Our method recovers these parameters using 3D point cloud information. With the recovered depth shift, the wall and bed edges become straight. However, the overall scene is stretched (3rd column). Finally, with recovered focal length, an accurate 3D scene can be reconstructed (4th column).

To summarize, our main contributions are as follows.

- We propose a novel framework for in-the-wild monocular 3D scene shape estimation. To our knowledge, this is the first method for this task, and the first method to leverage 3D point cloud neural networks for improving estimation of the structure of point clouds derived from depth maps.
- We propose an image-wise normalized regression loss and a pair-wise normal regression loss for improving monocular depth estimation models trained on mixed multi-source datasets.

Experiments show that our point cloud reconstruction method can recover accurate 3D shapes from single monocular images (up to scale). Also, for depth prediction, our method achieves state-of-the-art results on zero-shot dataset transfer to 10 unseen datasets.

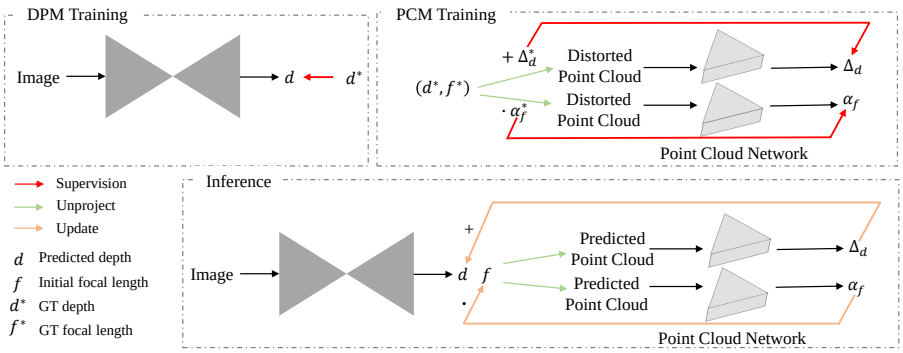


FIGURE 5.2. **The overall pipeline of our method.** During training, the depth prediction model (top left) and point cloud module (top right) are trained separately on different sources of data. During inference (bottom), the two networks are combined to predict depth d; and the depth shift Δ_d, the focal length $f \cdot \alpha_f$ using the predicted d, which together enable an accurate scene shape reconstruction. Note that we employ point cloud networks to predict shift and focal length scaling factors separately.

5.3 Our Methods

Our two-stage pipeline for 3D shape estimation from single images is shown in Fig. 5.2. It consists of a depth prediction module (DPM) and a point cloud module (PCM). The two modules are trained separately on different data sources, and are then combined together at inference time. The DPM takes an RGB image and outputs a depth map [160] with unknown scale and shift in relation to the true metric depth map. The PCM takes as input a distorted 3D point cloud that is computed using a predicted depth map d and an initial estimation of the focal length f,[1] and outputs shift adjustments to the depth map and focal length to improve the geometry of the reconstructed 3D scene shape. We describe the details of these two modules next.

5.3.1 Point Cloud Module

We assume a pinhole camera model for the 3D point cloud reconstruction, which means that the un-projection from 2D coordinates and depth to 3D points is:‘

$$\begin{cases} x = \frac{u-u_0}{f} d \\ y = \frac{v-v_0}{f} d \\ z = \quad d \end{cases} \tag{5.1}$$

where (u_0, v_0) is the camera optical center; f is the focal length, and d is the depth. The 3D point cloud is reconstructed based on the function $(x, y, z) = \mathcal{F}(u_0, v_0, f, d)$, see Eq. 5.1. The focal length affects the point cloud shape as it scales x and y coordinates, but not z. Similarly, a shift d affects the x, y, and z coordinates non-uniformly, which results in shape distortions.

For a human observer, these distortions are immediately recognizable when viewing the point cloud at an oblique angle, although they cannot be observed by looking at

[1] This initial value does not need to very accurate.

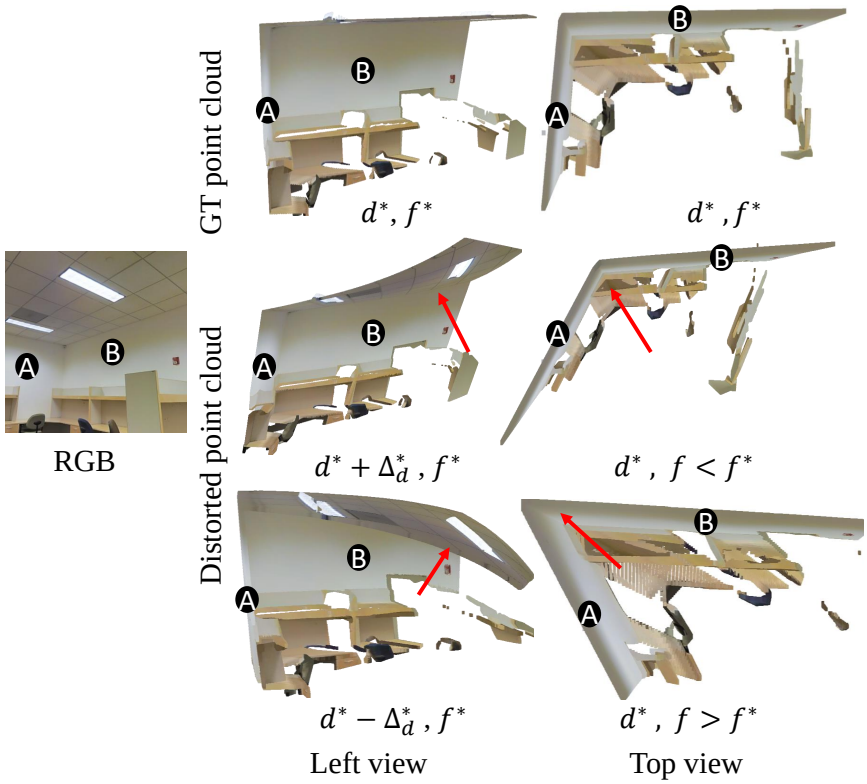


FIGURE 5.3. Illustration of the distorted 3D shape caused by incorrect shift and focal length. A ground-truth depth map is projected in 3D, which can create the ground truth point cloud (see the first row). `A` and `B` annotate the walls. When the focal length is incorrectly estimated ($f > f^*$ or $f < f^*$), we observe significant structural distortion, *e.g.*, see the angle between two walls `A` and `B` (see the third column). Second column: a shift ($d^* + \Delta_d$ or $d^* - \Delta_d$) also causes the shape distortion, see the roof. Note that different distortions are caused by the negative or positive shift.

a depth map alone. In Fig. 5.3, we can see that a shift for the depth will cause planes camber, while the focal length will change the angle between two planes (see the last column).

As a result, we propose to *directly analyze the point cloud to estimate the unknown shift and focal length, instead of working with 2D images*. We have tried a number of network architectures that take unstructured 3D point clouds as input, and found that the recent PVCNN [85] performs well for this task. Thus, we build our method on the PVCNN architecture here.

During training, a perturbed input point cloud with incorrect shift and focal length is synthesized by perturbing the known ground-truth depth shift and focal length. The ground-truth depth d^* is transformed by a shift Δ_d^* drawn from $\mathcal{U}(-0.25, 0.8)$, and the ground truth focal length f^* is transformed by a scale α_f^* drawn from $\mathcal{U}(0.6, 1.25)$ to keep the focal length positive and non-zero.

When recovering the depth shift, the perturbed 3D point cloud $\mathcal{F}(u_0, v_0, f^*, d^* + \Delta_d^*)$ is given as input to the shift point cloud network $\mathcal{N}_d(\cdot)$, trained with the objective:

$$L = \min_{\theta} |\mathcal{N}_d(\mathcal{F}(u_0, v_0, f^*, d^* + \Delta_d^*), \theta) - \Delta_d^*| \tag{5.2}$$

where θ are network weights and f^* is the true focal length.

Similarly, when recovering the focal length, the point cloud $\mathcal{F}(u_0, v_0, \alpha_f^* f^*, d^*)$ is fed to the focal length point cloud network $\mathcal{N}_f(\cdot)$, trained with the objective:

$$L = \min_{\theta} \left| \mathcal{N}_f(\mathcal{F}(u_0, v_0, \alpha_f^* f^*, d^*), \theta) - \alpha_f^* \right| \tag{5.3}$$

During inference, the ground-truth depth is replaced with the predicted affine-invariant depth d, which is normalized to $[0, 1]$ prior to the 3D reconstruction. We use an initial guess of focal length f, giving us the reconstructed point cloud $\mathcal{F}(u_0, v_0, f, d)$, which is fed to $\mathcal{N}_d(\cdot)$ and $\mathcal{N}_f(\cdot)$ to predict the shift Δ_d and focal length scaling factor α_f respectively. In our experiments, we simply use an initial focal length with a field of view (FOV) of 60°. We have also tried to employ a single network to predict both the shift and the scaling factor, but have empirically found that two separate networks can achieve a better performance.

5.3.2 Monocular Depth Prediction Network

Our monocular depth prediction network takes an RGB image I_{rgb} as input and produces an affine-invariant depth map d. We train our depth prediction on multiple data sources including high-quality LiDAR sensor data [165], and low-quality web stereo data [104, 137, 150] (see Sec. 5.4). As these datasets have varied depth ranges and web stereo datasets containing unknown depth scale and shift, we propose an image-level normalized regression (ILNR) loss to address this issue. Moreover, we propose a pair-wise normal regression (PWN) loss to exploit local geometry information.

Image-level normalized regression loss. Depth maps of different data sources can have varied depth ranges. Normalization is a critical step to transform data with variable ranges to a comparable range where large features no longer dominate smaller features [125]. Therefore, we propose to normalize the data to make the model training easier. Simple Min-Max normalization [38, 125] is sensitive to depth value outliers. For example, a large value at a single pixel will affect the rest of the depth map after the Min-Max normalization. We investigate more robust normalization methods and propose a simple but effective image-level normalized regression loss for mixed-data training.

Our image-level normalized regression loss transforms each ground-truth depth map to a similar numerical range based on its individual statistics. To reduce the effect of outliers and long-tail residuals, we combine tanh normalization [125] with a trimmed Z-score normalization, after which we can simply apply a pixel-wise mean average error (MAE) between the prediction and the normalized ground-truth depth maps. The ILNR loss is formally defined as follows.

$$L_{\text{ILNR}} = \frac{1}{N} \sum_{i}^{N} \left| d_i - \overline{d}_i^* \right| + \left| \tanh(d_i/100) - \tanh(\overline{d}_i^*/100) \right| \tag{5.4}$$

where $\overline{d}_i^* = {}^{(d_i^* - \mu_{\text{trim}})}/_{\sigma_{\text{trim}}}$ and μ_{trim} and σ_{trim} are the mean and the standard deviation of a trimmed depth map which has the nearest and farthest 10% of pixels removed. d is the predicted depth, and d^* is the ground-truth depth map.

We have tested a number of other normalization methods such as Min-Max normalization [125], Z-score normalization [37], and median absolute deviation normalization (MAD) [125]. In our experiments, we observe that our proposed ILNR loss achieves the best performance and generalization.

Pair-wise normal loss. Surface normals are an important geometric property, which have been shown to be a complementary modality to depth [124]. Many methods have been proposed to use normal constraints to improve the depth quality, such as the virtual normal loss [161]. However, as the virtual normal only leverages global structure, it may not help improve the local geometric quality, such as depth edges and planes. Recently, Xian *et al.* [150] proposed a structure-guided ranking loss, which can improve edge sharpness. Inspired by these methods, we follow their sampling method but enforce the supervision in the surface normal space. Moreover, our samples include not only edges but also planes. Our proposed pair-wise normal (PWN) loss can better constrain both the global and local geometric relations.

The detailed sampling method is described here. The first step is to locate image edges. We follow [161] to calculate the surface normal from the depth map with the local least squares fitting method. The Sobel edge detector is applied to find edges from the surface normal map and the input image. At each edge point, we then sample pairs of points on both sides of the edge. The ground-truth normals for these points are $\mathcal{N}^* = \{(\boldsymbol{n}_A^*, \boldsymbol{n}_B^*)_i | i = 0, ..., n\}$, while the predicted normals are $\mathcal{N} = \{(\boldsymbol{n}_A, \boldsymbol{n}_B)_i | i = 0, ..., n\}$. Before calculating the predicted surface normal, we align the predicted depth and the ground-truth depth with a scale and shift factor, which are retrieved by the least squares fitting [104]. To locate the object boundaries and planes folders, where the normals changes significantly, we set the angle difference of two normals greater than arccos(0.3). To balance the samples, we also get some negative samples, where the angle difference is smaller than arccos(0.95) and they are also detected as edges on the input image. The sample strategy is defined:

$$\mathcal{S}_1 = \{\boldsymbol{n}_A^* \cdot \boldsymbol{n}_B^* > 0.95, \boldsymbol{n}_A^* \cdot \boldsymbol{n}_B^* < 0.3 | (\boldsymbol{n}_A^*, \boldsymbol{n}_B^*)_i \in \mathcal{N}^*\} \tag{5.5}$$

To improve the quality of planes, we also sample points on the same plane and enforce the co-plane supervisions on predictions. We employ different methods to locate planes. For the high-quality Lidar data, Taskonomy [165], we locate planes by finding regions with the same surface normal. For noisy data, DIML, we use [13] to segment the roads, which we assume to be planar regions. Then we obtain the predicted surface normal for the samples. Note that we have also tried to get the virtual normal from samples and enforce the co-plane constraints, which has similar performance to the surface normal constraints. In doing so, we sample 100K paired points per training sample on average.

The sampled points are $\{(A_i, B_i), i = 0, ..., N\}$. The PWN loss is:

$$L_{\text{PWN}} = \frac{1}{N}\sum_{i}^{N} |n_{Ai} \cdot n_{B_i} - n^*_{Ai} \cdot n^*_{Bi}| \tag{5.6}$$

where n^* denotes ground truth surface normals. Note that if points are on the same plane, $n^*_{Ai} \cdot n^*_{Bi} = 1$. As this loss accounts for both local and global geometries, we find that it improves the overall reconstructed shape.

Finally, we also use a multi-scale gradient loss (MSG) [77] and structure-guided ranking loss (SR) [150]. The MSG loss is as follows.

$$L_{\text{MSG}} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N} \left|\bigtriangledown^k_x d_i - \bigtriangledown^k_x \bar{d}^*_i\right| + \left|\bigtriangledown^k_y d_i - \bigtriangledown^k_y \bar{d}^*_i\right| \tag{5.7}$$

where K is the number of scale, N is the number of valid samples.

The structure-guided ranking loss is as follows.

$$L_{\text{SR}} = \frac{1}{N}\sum_{i=0}^{N} \begin{cases} \log\left(1 + \exp\left[-l\left(d_{i0} - d_{i1}\right)\right]\right); & l \neq 0 \\ \left(d_{i0} - d_{i1}\right)^2, & l = 0, \end{cases} \tag{5.8}$$

where $l = \begin{cases} +1, & \text{if } d^*_{i0}/d^*_{i1} \geq 1 + \tau; \\ -1, & \text{if } d^*_{i1}/d^*_{i0} \geq 1 + \tau;. \\ 0, & \text{otherwise.} \end{cases}$ τ is a predefined threshold.

The training strategy for mixed datasets. We mix5 datasets to train the depth model. Based on their depth quality, they are categorized to high-quality data (Taskonomy [165] and 3d ken burns [95]), middle-quality data (DIML [63]), and low-quality web-stereo data (Holopix50K [54] and HRWSI [150]).

For the low-quality web-stereo data, as their inverse depths d_{inv} have unknown scale and shift, *i.e.*, $d_{inv} = s \cdot d^*_{inv} + \Delta_{d\_inv}$, the depth map ($d = 1/d_{inv} = 1/(s \cdot d^*_{inv} + \Delta_{d\_inv})$) can only demonstrate the relative depth relations. Therefore, we only enforce the structured-guided ranking loss on those data.

For the middle-quality data, such as DIML [63], we enforce the proposed image-level normalized regression loss, multi-scale gradient loss and ranking loss. As depths contain much noise in local regions (see Fig. 5.4), enforcing the pair-wise normal regression loss on noisy edges will cause many artifacts. Therefore, we enforce the pair-wise normal regression loss only on planar regions.

For the high-quality data, accurate edges and planes can be easily located. Therefore, we enforce all losses on these data.

The overall loss functions for different datasets are reported in Table 5.1.

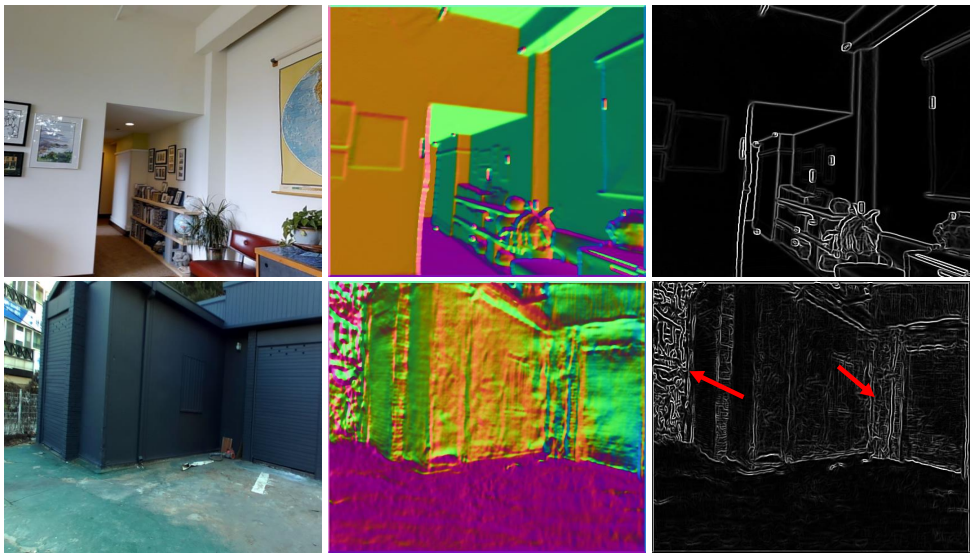


FIGURE 5.4. Comparison of edges of high-quality data and middle-quality data. The first row is taskonomy, while the second row is DIML. Red arrows highlight artifacts on edges.

TABLE 5.1. Losses enforced for different datasets based on their depth quality.

	L_{SR}	L_{ILNR}	L_{PWN} (Edges)	L_{PWN} (Planes)	L_{MSG}
High-quality data	✓	✓	✓	✓	✓
Middle-quality data	✓	✓		✓	✓
Low-quality data	✓				

5.3.3 Depth Completion

Depth completion is an important problem, which aims to produce accurate dense metric depth maps from sparse or incomplete depth maps. Existing completion methods can be classified into two categories according to the input sparsity pattern: depth inpainting methods that fill large holes [168, 120, 55], and sparse depth completion methods that fill sparsely distributed depth measurements [22, 96, 153, 102, 21]. When working on a specific sparsity pattern, e.g., on either NYU [124] or KITTI [132], recent approaches [96, 22, 20, 102] such as NLSPN [96] can obtain impressive performance. However, in real-world scenarios, the specific sparsity pattern may be unknown at training time, as it is a function of hardware, software, as well as the configuration of the scene itself.

Therefore, multiple models have to be trained to solve various sparse depth situations. Our proposed mix-data training strategy is also effective to improve the generalization of depth completion model. During training, we synthetically create the sparse depth input by sampling from the ground-truth depth. To improve the generalization to different sparsity types, we create several different sparse depth patterns, including uniform sampling [96], sampling points from feature points [19], and randomly masking multiple continuous regions from the ground-truth depth. We

TABLE 5.2. Overview of the test sets in our experiments.

Dataset	# Img	Scene Type	Evaluation Metric	Supervision Type
NYU	654	Indoor	AbsRel & δ_1	Kinect
ScanNet	700	Indoor	AbsRel & δ_1	Kinect
2D-3D-S	12256	Indoor	LSIV	LiDAR
iBims-1	100	Indoor	AbsRel & ε_{PE} &ε_{DBE}	LiDAR
KITTI	652	Outdoor	AbsRel & δ_1	LiDAR
Sintel	641	Outdoor	AbsRel & δ_1	Synthetic
ETH3D	431	Outdoor	AbsRel & δ_1	LiDAR
YouTube3D	58054	In the Wild	WHDR	SfM, Ordinal pairs
OASIS	10000	In the Wild	WHDR & LSIV	User clicks, Small patches with GT
DIODE	771	Indoor & Outdoor	AbsRel & δ_1	LiDAR
TUM-RGBD	1815	Indoor	AbsRel & SiLog	Kinect

use the high-quality data and middle-quality data to train the model. Note that on middle-quality data, we only enforce the pair-wise normal regression loss on planes. The loss function is as follows.

$$L = L_1 + L_{\text{PWN}} + L_{\text{SR}}, \tag{5.9}$$

where L_1 is the pixel-wise error.

5.4 Experiments

Datasets and implementation details. To train the PCM, we sample 100K Kinect-captured depth maps from ScanNet, 114K LiDAR-captured depth maps from Taskonomy, and 51K synthetic depth maps from the 3D Ken Burns paper [95]. We train the network using SGD with a batch size of 40, an initial learning rate of 0.24, and a learning rate decay of 0.1. For parameters specific to PVCNN, such as the voxel size, we follow the original work [85].

To train the DPM, we sample 114K RGBD pairs from LiDAR-captured Taskonomy [165], 51K synthetic RGBD pairs from the 3D Ken Burns paper [95], 121K RGBD pairs from calibrated stereo DIML [63], 48K RGBD pairs from web-stereo Holopix50K [54], and 20K web-stereo HRWSI [150] RGBD pairs. Note that for the ablation study about the effectiveness of PWN and ILNR, we sample a smaller dataset which is composed of 12K images from Taskonomy, 12K images from DIML, and 12K images from HRWSI. During training, 1000 images are withheld from all datasets as a validation set. We use the depth prediction architecture proposed in Xian *et al.* [150], which consists of a standard backbone for feature extraction (*e.g.*, ResNet50 [48] or ResNeXt101 [152]), followed by a decoder, and train it using SGD with a batch size of 40, an initial learning rate 0.02 for all layer, and a learning rate decay of 0.1. Images are resized to

TABLE 5.3. Effectiveness of recovering the shift from 3D point clouds with the PCM. Compared with the baseline, the AbsRel↓ is much lower after recovering the depth shift over all test sets.

Method	ETH3D	NYU	KITTI	Sintel	DIODE
	AbsRel↓				
Baseline	23.7	25.8	23.3	47.4	46.8
Recovered Shift	**15.9**	**15.1**	**17.5**	**40.3**	**36.9**

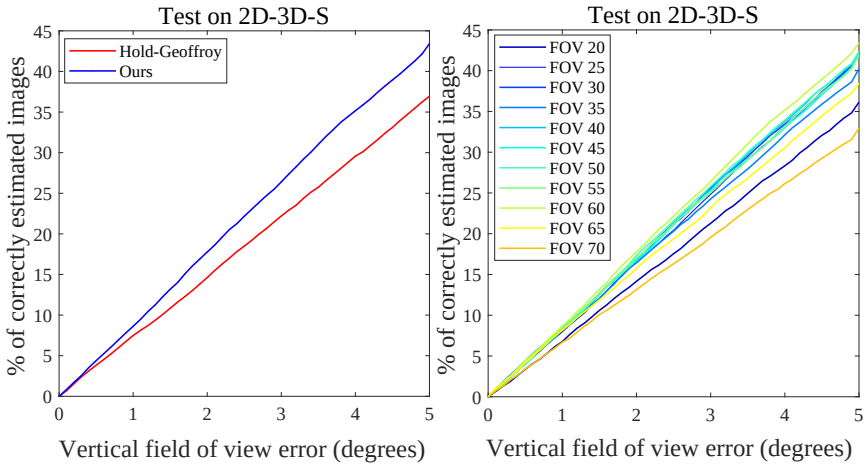


FIGURE 5.5. Comparison of the recovered focal length on the 2D-3D-S dataset. Left: our method outperforms Hold-Geoffroy *et al.* [52]. Right: we conduct an experiment on the effect of the initialization of field of view (FOV). Our method remains robust across different initial FOVs, with a slight degradation in quality beyond 25° and 65°.

448×448, and flipped horizontally with a 50% chance. Following [160], we load data from different datasets evenly for each batch.

Evaluation details. The accuracy of focal length prediction is evaluated on 2D-3D-S [2] following [52]. Furthermore, to evaluate the accuracy of the reconstructed 3D shape, we use the Locally Scale Invariant RMSE (LSIV) [16] metric on both OASIS [16] and 2D-3D-S [2]. It is consistent with the previous work [16]. The OASIS [16] dataset only has the ground-truth depth on some small regions, while 2D-3D-S has the ground truth for the whole scene.

To evaluate the generalization of our proposed depth prediction method, we test on 9 datasets which are unseen during training, including YouTube3D [14], OASIS [16], NYU [124], KITTI [40], ScanNet [25], DIODE [135], ETH3D [118], Sintel [7], and iBims-1 [65]. On OASIS and YouTube3D, we use the Weighted Human Disagreement Rate (WHDR) [149] for evaluation. On other datasets, except for iBims-1, we evaluate the absolute mean relative error (AbsRel↓) and the percentage of pixels with $\delta_1 = \max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) < 1.25$. We follow Ranftl *et al.* [104] and align the scale and shift before evaluation.

To evaluate the geometric quality of the depth, *i.e.*, the quality of edges and planes, we follow [95, 150] and evaluate the depth boundary error [65] $(\varepsilon_{\text{DBE}}^{\text{acc}}, \varepsilon_{\text{DBE}}^{\text{comp}})$ as well as the planarity error [65] $(\varepsilon_{\text{PE}}^{\text{plan}}, \varepsilon_{\text{PE}}^{\text{orie}})$ on iBims-1. $\varepsilon_{\text{PE}}^{\text{plan}}$ and $\varepsilon_{\text{PE}}^{\text{orie}}$ evaluate the flatness and orientation of reconstructed 3D planes compared to the ground truth 3D

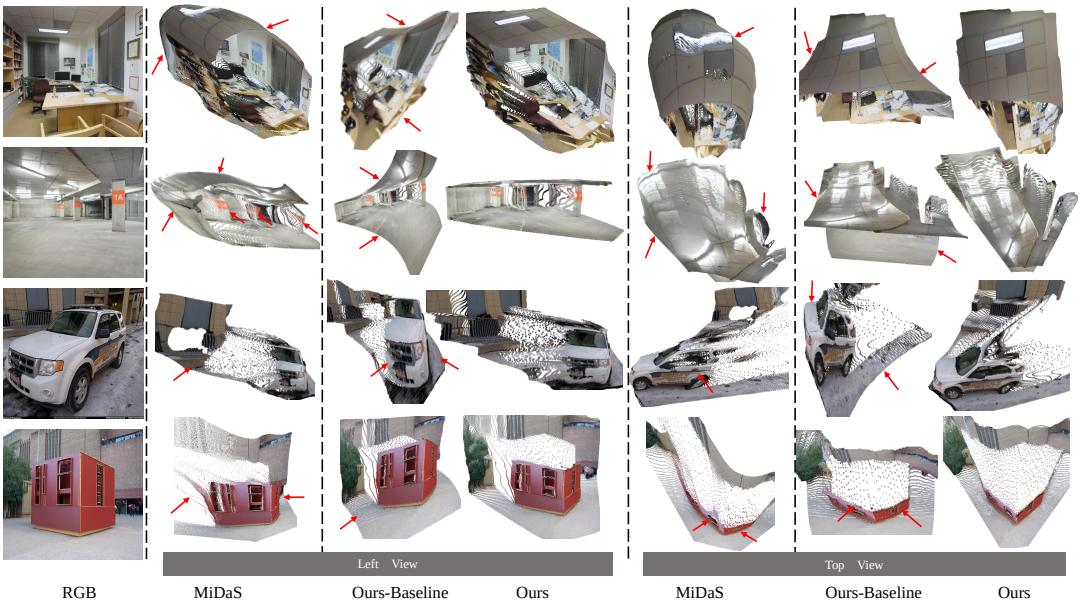


FIGURE 5.6. Qualitative comparison. We compare the reconstructed 3D shape of our method with several baselines. As MiDaS [104] does not estimate the focal length, we use the focal length recovered from [52] to convert the predicted depth to a point cloud. "Ours-Baseline" does not recover the depth shift or focal length and uses an orthographic camera, while "Ours" recovers the shift and focal length. We can see that our method better reconstructs the 3D shape, especially at edges and planar regions (see arrows).

planes respectively, while $\varepsilon_{\text{DBE}}^{\text{acc}}$ and $\varepsilon_{\text{DBE}}^{\text{comp}}$ measure the localization accuracy and the sharpness of edges respectively. More details as well as a comparison of these test datasets are summarized in Table 5.2

5.4.1 3D Shape Reconstruction

Shift recovery. To evaluate the effectiveness of our depth shift recovery, we perform zero-shot evaluation on 5 datasets unseen during training. We recover a 3D point cloud by unprojecting the predicted depth map, and then compute the depth shift using our PCM. We then align the unknown scale [6, 43] of the original depth and our shifted depth to the ground-truth, and evaluate both using the AbsRel↓ error. The results are shown in Table 5.3, where we see that, on all test sets, the AbsRel↓ error is lower after recovering the shift. We also trained a standard 2D convolutional neural network to predict the shift given an image composed of the un-projected point coordinates, but this approach did not generalize well to samples from unseen datasets.

Focal length recovery. To evaluate the accuracy of our recovered focal length, we follow Hold-Geoffroy *et al.* [52] and compare on the 2D-3D-S dataset, which is unseen during training for both methods. The model of [52] is trained on the in-the-wild SUN360 [151] dataset. Results are illustrated in Fig. 5.5, where we can see that our method demonstrates better generalization performance. Note that PVCNN is very lightweight and only has 5.5M parameters, but shows promising a generalization

TABLE 5.4. Quantitative evaluation of the reconstructed 3D shape quality on OASIS and 2D-3D-S. Our method can achieve better performance than previous methods. Compared with the orthographic projection, our method using the pinhole camera model can obtain better performance. DPM and PCM refers to our depth prediction module and point cloud module, respectively.

Method	OASIS LSIV ↓	2D-3D-S LSIV↓
Orthographic Camera Model		
MegaDepth [77]	0.64	2.68
MiDaS [104]	0.63	2.65
Ours-DPM	0.63	2.65
Pinhole Camera Model		
MegaDepth [77] + Hold-Geoffroy [52]	1.69	1.81
MiDaS [104] + Hold-Geoffroy [52]	1.60	0.94
MiDaS [104] + Ours-PCM	1.32	0.94
Ours	**0.52**	**0.80**

capability, which indicates that there is a much smaller domain gap between datasets in the 3D point cloud space than in the image space where appearance variation can be large.

Furthermore, we analyze the effect of different initial focal lengths during inference. We set the initial field of view (FOV) from 20° to 70° and evaluate the accuracy of the recovered focal length, Fig. 5.5 (right). The experimental results demonstrate that our method is not particularly sensitive to different initial focal lengths.

Evaluation of 3D shape quality. Following OASIS [16], we use LSIV for the quantitative comparison of recovered 3D shapes on the OASIS [16] dataset and the 2D-3D-S [2] dataset. OASIS only provides the ground truth point cloud on small regions, while 2D-3D-S covers the whole 3D scene. Following OASIS [16], we evaluate the reconstructed 3D shape with two different camera models, *i.e.*, the orthographic projection camera model [16] (infinite focal length) and the (more realistic) pinhole camera model. As MiDaS [104] and MegaDepth [77] do not estimate the focal length, we use the focal length recovered from Hold-Geoffroy [52] to convert the predicted depth to a point cloud. We also evaluate a baseline using MiDaS instead of our DPM with the focal length predicted by our PCM ("MiDaS + Ours-PCM"). From Table 5.4 we can see that with an orthographic projection, our method ("Ours-DPM") performs roughly as well as existing state-of-the-art methods. However, for the pinhole camera model our combined method significantly outperforms existing approaches. Furthermore, comparing "MiDaS + Ours-PCM" and "MiDaS + Hold-Geoffroy", we note that our PCM is able to generalize to different depth prediction methods.

A qualitative comparison of the reconstructed 3D shape for in-the-wild scenes is shown in Fig. 5.6. It demonstrates that our model can recover significantly more accurate 3D scene shapes. For example, planar structures such as walls, floors, and roads are much flatter in our reconstructed scenes, and the angles between surfaces (*e.g.*, walls) are also more realistic. Also, the shape of the car has less distortions.

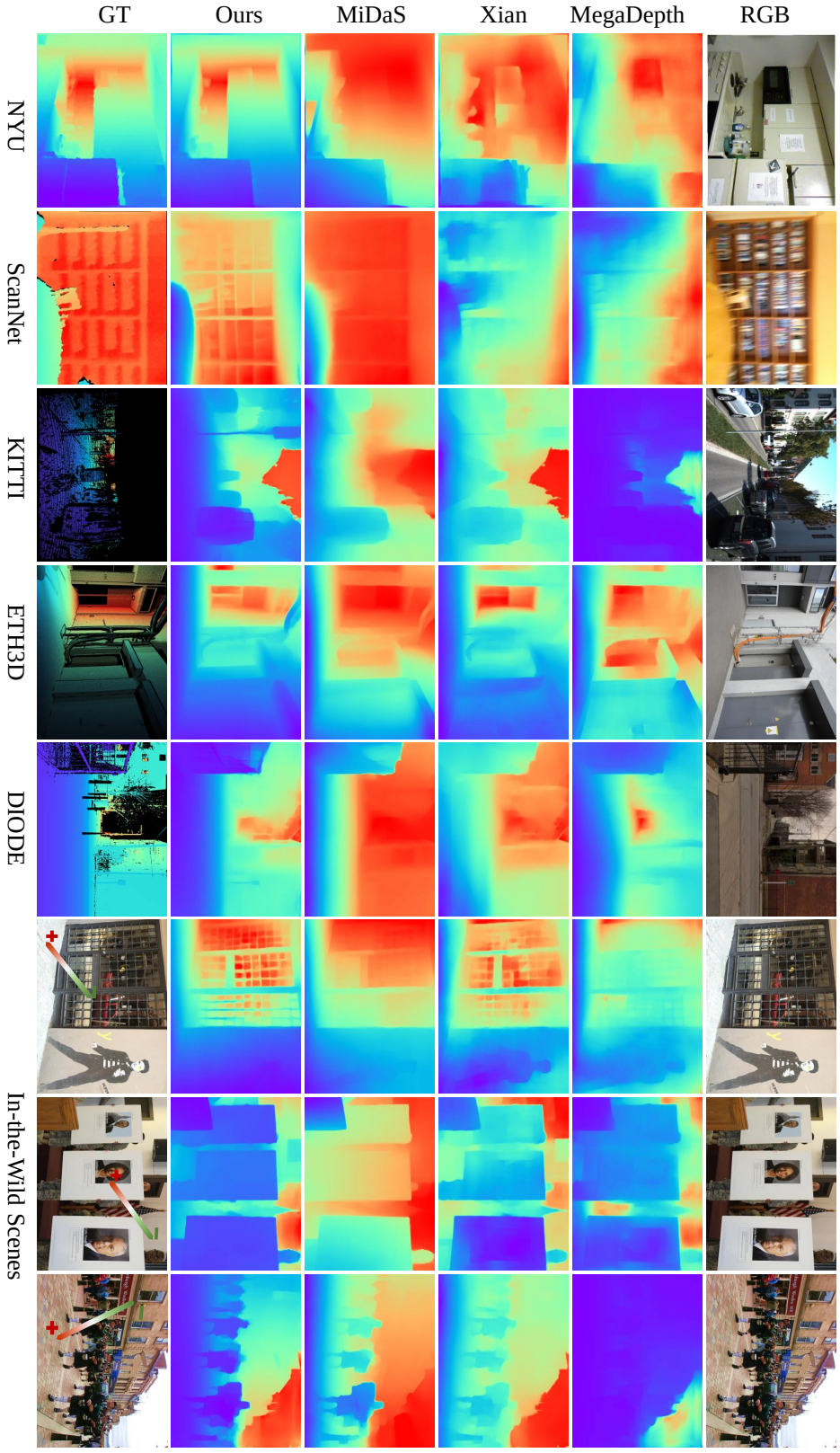


FIGURE 5.7. Qualitative comparisons with state-of-the-art methods, including MegaDepth [77], Xian *et al.* [150], and MiDaS [104]. It shows that our method can predict more accurate depths at far locations and regions with complex details. In addition, we see that our method generalizes better to in-the-wild scenes.

TABLE 5.5. Quantitative comparison of the quality of depth boundaries (DBE) and planes (PE) on the iBims-1 dataset. We use † to indicate when a method was trained on the small training subset.

Method	iBims-1				
	$\varepsilon_{\text{DBE}}^{\text{acc}} \downarrow$	$\varepsilon_{\text{DBE}}^{\text{comp}} \downarrow$	$\varepsilon_{\text{PE}}^{\text{plan}} \downarrow$	$\varepsilon_{\text{PE}}^{\text{orie}} \downarrow$	AbsRel↓↓
Xian [150]	7.72	9.68	5.00	44.77	0.301
MegaDepth [77]	4.09	8.28	7.04	33.03	0.20
MiDaS [104]	1.91	5.72	3.43	12.78	0.104
3D Ken Burns [95]	2.02	**5.44**	2.19	10.24	0.097
Ours† w/o PWN	2.05	6.10	3.91	13.47	0.106
Ours†	1.91	5.70	2.95	11.59	0.101
Ours Full	**1.90**	5.73	**2.0**	**7.41**	**0.079**

5.4.2 Monocular Depth Estimation

In this section, we conduct several experiments to demonstrate the effectiveness of our depth prediction method, including a comparison with state-of-the-art methods, a comparison of our proposed image-level normalized regression loss with other methods, and an analysis of the effectiveness of our pair-wise normal regression loss.

TABLE 5.6. Quantitative comparison of our depth prediction with state-of-the-art methods on eight zero-shot (unseen during training) datasets. Our method achieves better performance than existing state-of-the-art methods across all test datasets.

Method	Backbone	OASIS	YT3D	NYU		KITTI		DIODE		ScanNet		ETH3D		Sintel		Rank↓
		WHDR↓		AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	
OASIS [16]	ResNet50	32.7	27.0	21.9	66.8	31.7	43.7	48.4	53.4	19.8	69.7	29.2	59.5	60.2	42.9	6.7
MegaDepth [77]	Hourglass	33.5	26.7	19.4	71.4	20.1	66.3	39.1	61.5	19.0	71.2	26.0	64.3	39.8	52.7	6.7
Xian *et al.* [150]	ResNet50	31.6	23.0	16.6	77.2	27.0	52.9	42.5	61.8	17.4	75.9	27.3	63.0	52.6	50.9	6.7
WSVD [137]	ResNet50	34.8	24.8	22.6	65.0	24.4	60.2	35.8	63.8	18.9	71.4	26.1	61.9	35.9	54.5	6.6
Chen *et al.* [14]	ResNet50	33.6	20.9	16.6	77.3	32.7	51.2	37.9	66.0	16.5	76.7	23.7	67.2	38.4	57.4	5.6
DiverseDepth [160]	ResNeXt50	30.9	21.2	11.7	87.5	19.0	70.4	37.6	63.1	10.8	88.2	22.8	69.4	38.6	58.7	4.4
MiDaS [104]	ResNeXt101	29.5	19.9	11.1	88.5	23.6	63.0	33.2	71.5	11.1	88.6	18.4	75.2	40.5	60.6	3.5
Ours	ResNet50	30.2	19.5	9.1	91.4	**14.3**	**80.0**	28.7	75.1	9.6	90.8	18.4	75.8	34.4	62.4	1.9
Ours	ResNeXt101	**28.3**	**19.2**	**9.0**	**91.6**	14.9	78.4	**27.1**	**76.6**	**9.5**	**91.2**	**17.1**	**77.7**	**31.9**	**65.9**	**1.1**

TABLE 5.7. Quantitative comparison of different losses on zero shot generalization to 5 datasets unseen during training.

Method	RedWeb WHDR↓	NYU	KITTI	ScanNet AbsRel↓	DIODE
SMSG [137]	19.1	15.6	16.3	13.7	36.5
SSMAE [104]	19.2	14.4	18.2	13.3	34.4
MinMax	19.1	15.0	17.1	13.3	46.1
MAD	18.8	14.8	17.4	12.5	34.6
ILNR	**18.7**	**13.9**	**16.1**	**12.3**	**34.2**

Comparison with state-of-the-art methods. In this comparison, we test on datasets unseen during training. We compare with methods that have been shown to best generalize to in-the-wild scenes. Their results are obtained by running the publicly released codes. Each method is trained on its own proposed datasets. When comparing the AbsRel↓ error, we follow Ranftl [104] to align the scale and shift before the evaluation. The results are shown in the Table 5.6. Our method outperforms prior works, and using a larger ResNeXt101 backbone further improves the results. Some qualitative comparisons are shown in Fig. 5.7.

Pair-wise normal loss. To evaluate the quality of our full method and dataset on edges and planes, we compare our depth model with previous state-of-the-art methods on the iBims-1 dataset. In addition, we evaluate the effect of our proposed pair-wise normal (PWN) loss through an ablation study. As training on our full dataset is computationally demanding, we perform this ablation on the small training subset. To do the quick comparison, the ResNet50 backbone model supervised with the PWN (W PWN) or without the PWN (W/O PWN) is trained on the sampled small training set. The results are shown in Table 5.5. We can see that our full method outperforms prior work for this task. In addition, under the same settings, both edges and planes are improved by adding the PWN loss. We further show a qualitative comparison of depths and reconstructed point clouds in Fig. 5.8 and Fig. 5.9 respectively. We can see that the edges in depths are more accurate and sharper than those without PWN supervision, and the reconstructed point clouds have much less distortions.

Image-level normalized regression loss. To show the effectiveness of our proposed image-level normalized regression (ILNR) loss, we compare it with the scale-shift invariant loss (SSMAE) [104] and the scale-invariant multi-scale gradient loss [137]. Each of these methods is trained on the small training subset to limit the computational overhead, and comparisons are made to datasets that are unseen during training. All models have been trained for 50 epochs, and we have verified that all models fully converged by then. The quantitative comparison is shown in Table 5.7, where we can see an improvement of ILNR over other scale and shift invariant losses. Furthermore, we also analyze different options for normalization, including image-level Min-Max (MinMax) normalization and image-level median absolute deviation (MAD) normalization, and found that our proposed loss performs a bit better.

Comparison of depth prediction on people. Li *et al.* [78] propose the first work

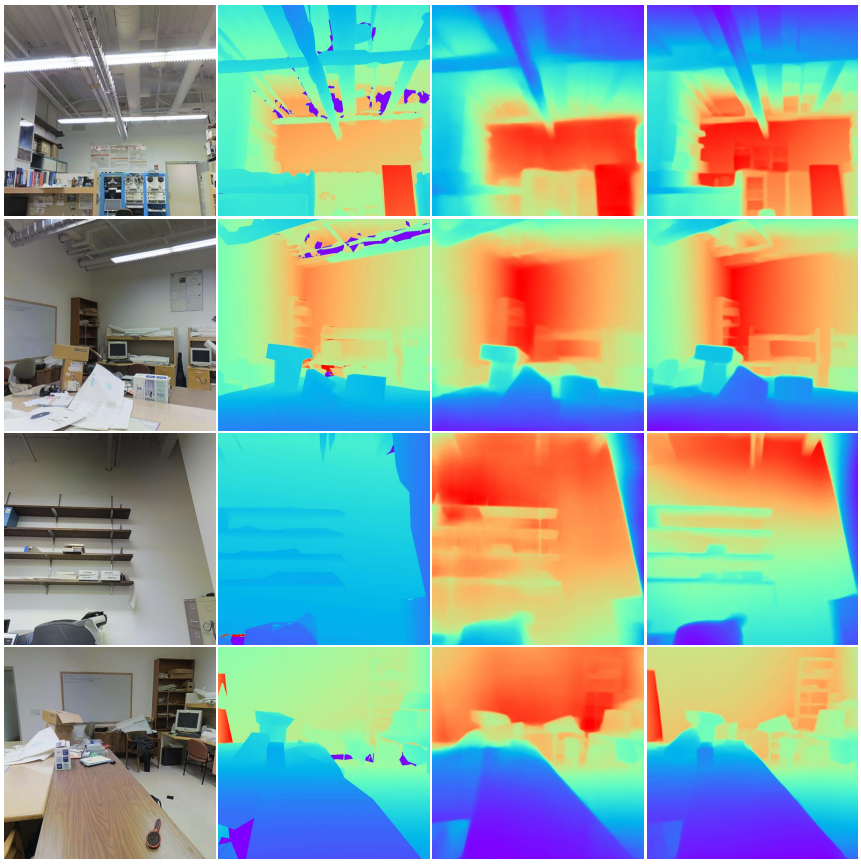


FIGURE 5.8. Qualitative comparison. Using the pair-wise normal loss (PWN), we can see that depths have finer details on edges.

to solve the depth prediction of moving people problem. Apart from RGB image, they propose to input the background depth which is obtained by the structure-from-motion method and the mask of humans as the guidance (see Li-IFCM and Li-IDCM in Table 5.8) to predict high quality depth of moving people. In comparison, our method only takes a single RGB image. Following [78], we conduct the comparison on the TUM-RGBD [128] dataset. The quantitative comparison illustrated in Table 5.8 shows that our method can achieve comparable performance with them. On humans, our depth is more accurate than other methods. Moreover, the visual results are illustrated in Fig. 5.10. We can see that our predicted depths have less artifacts and sharper edges than Li *et al.* [78] and DiverseDepth [160].

Additional qualitative results on in-the-wild scenes. Fig. 5.11 demonstrate more in-the-wild scenes examples. We can see that the predicted depths exhibit fine details on edges. Furthermore, we show reconstructed point clouds.

5.4.3 Depth Completion

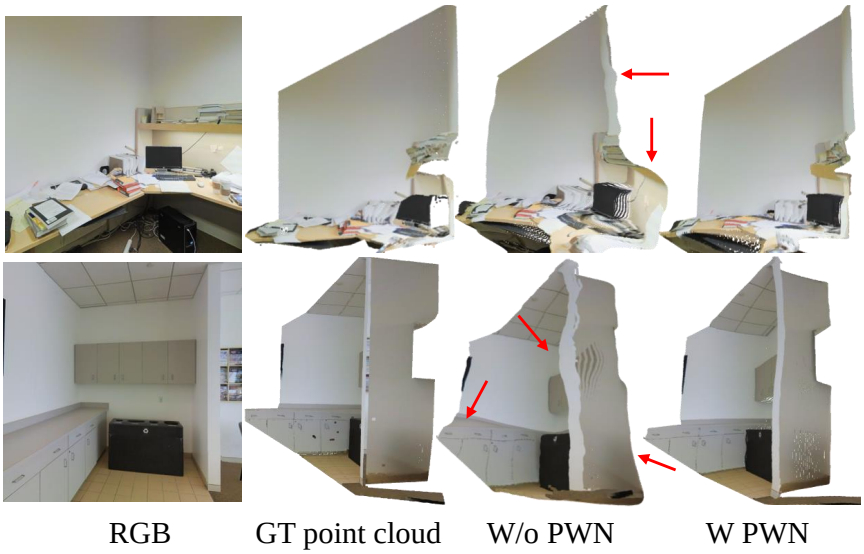


FIGURE 5.9. Qualitative comparison of reconstructed point clouds. Using the pair-wise normal loss (PWN), we can see that edges and planes are better reconstructed (see highlighted regions).

TABLE 5.8. Comparison of the foreground people on the TUM-RGBD datasets. Our overall performance is comparable with previous methods, while our depths are more accurate on foreground people. Note that [78] needs extra input such as the semantic human masks.

Method	Si-hum↓	Si-env↓	Si-RMS↓	AbsRel↓
DeMoN [133]	0.360	0.302	0.866	0.220
Li-I [78]	0.294	0.334	0.318	0.204
Li-IFCM [78]	0.302	0.330	0.316	0.206
Li-IDCM [78]	0.293	**0.238**	0.272	**0.147**
DiverseDepth [160]	<u>0.272</u>	0.270	<u>0.272</u>	0.192
Ours	**0.258**	<u>0.247</u>	**0.251**	<u>0.175</u>

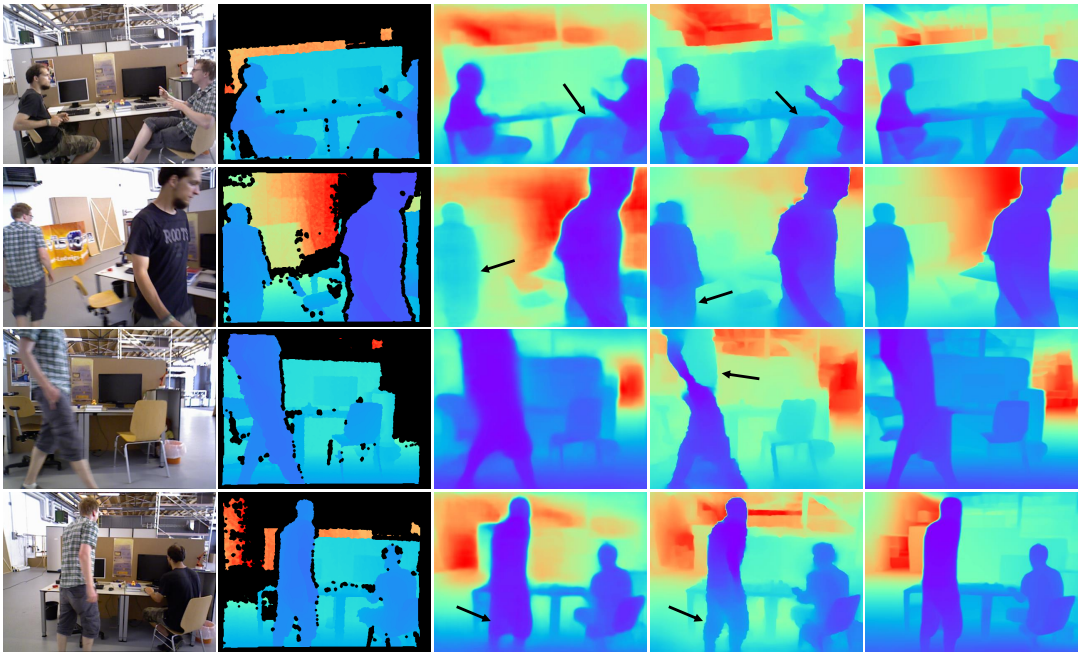


FIGURE 5.10. Qualitative comparison on the TUM-RGBD dataset. Following Li *et al.* [78], we compare the depth of moving people on the TUM-RGBD dataset.

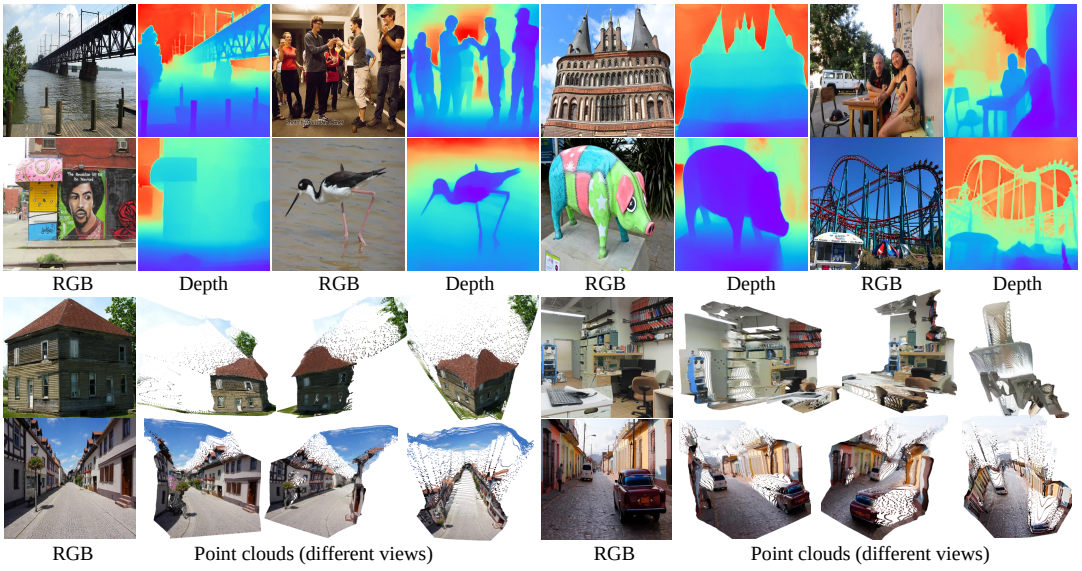


FIGURE 5.11. Qualitative results on some in-the-wild scenes. The reconstructed point clouds and depth maps of some in-the-wild scenes are illustrated.

TABLE 5.9. Comparison of our method with state-of-the-art methods on zero-shot test datasets. We create 4 different sparse depth types for evaluation. It is clear that our method has better generalization than previous methods on unseen data and different sparse depth patterns.

	NYU				ScanNet				DIODE			
	Features	Uni-10K	Lidar	Incomplete	Features	Uni-10K	Lidar	Incomplete	Features	Uni-10K	Lidar	Incomplete
	AbsRel↓				AbsRel↓				AbsRel↓			
NLSP [96]	0.096	0.204	0.158	0.574	0.653	0.632	1.327	0.620	19.005	14.940	38.157	15.259
Senushkin *et al.* [120]	5.199	5.299	5.187	4.846	1.523	0.726	0.591	0.070	3.660	3.104	2.583	0.693
Ours	**0.032**	**0.021**	**0.072**	**0.026**	**0.038**	**0.017**	**0.071**	**0.020**	**0.150**	**0.143**	**0.230**	**0.144**

TABLE 5.10. Quantitative comparison of our depth completion method with state-of-the-art methods on NYU dataset. Our method is *on par* with state-of-the-art methods, without training on the target dataset.

Methods	RMSE(m)↓	AbsRel↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
S2D [89]	0.230	0.044	97.1	99.4	99.8
S2D+SPN [87]	0.172	0.031	98.3	99.7	99.9
DepthCoeff [59]	0.118	0.013	99.4	99.9	-
CSPN [20]	0.117	0.016	99.2	99.9	100.0
CSPN++ [22]	0.116	-	-	-	-
DeepLiDAR [102]	0.115	0.022	99.3	99.9	100.0
DepthNormal [153]	0.112	0.018	99.5	99.9	100.0
NLSP [96]	0.092	0.012	99.6	99.9	100.0
Ours	0.19	0.036	98.4	99.6	100

In this section, we conduct several experiments to report the effectiveness of our training method for depth completion. To show the generalization of our method, we conduct the zero-shot testing on a few benchmark datasets. Note that we only train a single model to solve different sparse depth situations, while previous methods [96, 120] train different models for different sparse patterns.

Comparison with state-of-the-art depth completion methods. We test on standard benchmarks, NYU [124] and Matterport3D [11]. Note that our models have not been trained on these datasets. Two benchmarks have different types of sparse types. On NYU, the sparse depth only have 500 valid pixels, while Matterport3D provides the incomplete sensor-captured depth map.

Table 5.10 demonstrates results on NYU. Our method is on par with previous methods. Table 5.11 shows the comparison on the Matterport3D dataset. Ours can outperform previous methods on some metrics. Note that we do not fine tune our model on the target Matterport3D dataset.

Moreover, we demonstrate some visual comparisons in Fig. 5.12. Although the state-of-the-art method can achieve better quantitative performance than ours, our method, supervised by the geometric loss, can reconstruct more accurate scene structure. It is clear that our reconstructed walls are flatter than [120].

Generalization to different sparse depth types. To demonstrate the robustness of our methods to zero-shot test datasets and different sparse depth types, we create 4 sparse depth patterns and enclose 3 unseen datasets for evaluation. We compare our methods to the state-of-the-art methods on NYUD and Matterport3D benchmarks, i.e. NLSPN [96] and Senushkin *et al.* [120]. NLSPN method aims to complete the depth with only hundreds of valid points, while Senushkin *et al.* [120] method design to complete contiguous holes. Our created 'Uni-10K' are same to the sparsity pattern on NYU benchmark but have 10000 valid points. 'Features' employ the Fast corner detectors to sample points from the GT depth to create the sparse depth input. 'Lidar' aims to simulate the Lidar sensor, which captures depth in a linear scanning pattern. 'Incomplete' pattern is similar to the sparsity pattern of Matterport3D benchmark,

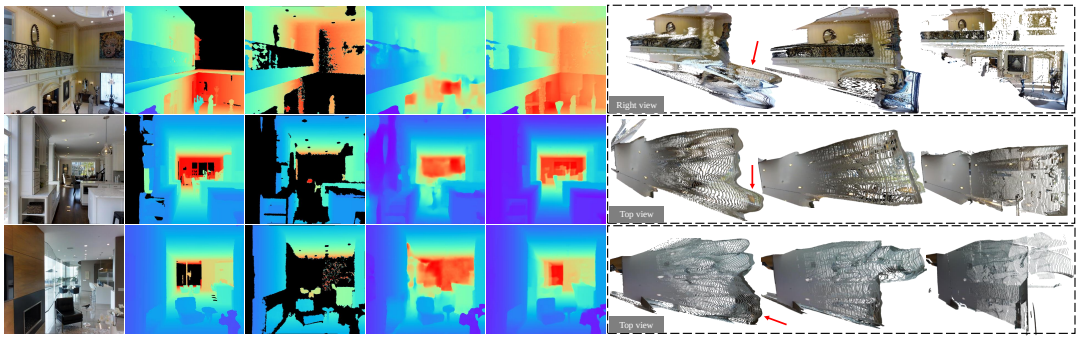


FIGURE 5.12. Qualitative comparison of the depth and reconstructed 3D shape. Our completed metric depth has finer details. The reconstructed 3D shape is more accurate than previous methods.

TABLE 5.11. Quantitative comparison of our depth completion method with state-of-the-art methods on the Matterport3D dataset. Note that we do not use any training data from the target Matterport3D dataset, while previous methods are trained on this dataset. Our method is on par with previous methods.

Methods	RMSE(m)↓	MAE(m)↓	$\delta_{1.05}$ ↑	$\delta_{1.1}$ ↑	δ_1 ↑	δ_2 ↑	δ_3 ↑
Huang *et al.* [55]	1.092	0.342	66.1	75.0	85.0	91.1	93.6
Zhang *et al.* [168]	1.316	0.461	65.7	70.8	78.1	85.1	88.8
Gansbeke *et al.* [134]	1.161	0.395	54.2	65.7	79.9	88.7	92.7
Li *et al.* [71]	1.054	0.397	50.8	63.1	77.5	87.4	92.0
Senushkin *et al.* [120]	**1.028**	**0.299**	**71.9**	**80.5**	**89.0**	93.2	95.0
Ours	2.35	0.574	68.9	79.2	88.1	**93.5**	**96.0**

which has missing depth on multiple large coherent regions.

We can see that although NLSP [96] and Senushkin *et al.* [120] can achieve state-of-the-art performance on NYU and Matterport3D dataset respectively, they cannot generalize to different types of sparse depth and unseen datsets. By contrast, our method can achieve comparable performance on different datasets. We conjunct taht our mix-data training strategy can significantly improve the model's robustness.

5.4.4 Applications

Monocular depth estimation can help many other tasks, such as image inpainting, objects removal and so on. Here we show an example, using our predicted depth to create 3D photo. Recently, several methods are proposed to synthesize a novel view from a single image, which need the monocular depth information as the guidance. We take the method of [122] to synthesize new views and the depth information is from our method or MiDaS [104]. From the synthesized new views, we randomly sample several views for comparison. Results are shown in Fig. 5.13. We can see that results with our provided depth have much fewer artifacts, see the woman's legs, the desk, and the white chair.

FIGURE 5.13. Comparison of synthesized new views with our depth and that of MiDaS. We can see that new views of ours show less artifacts and errors (see the woman's legs, the desk, and the chair).

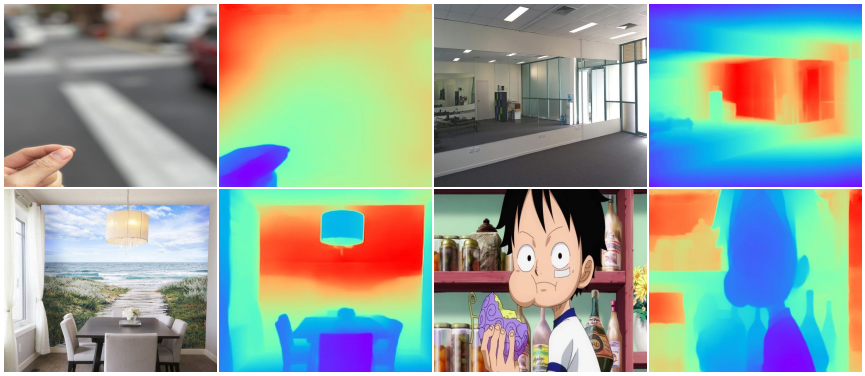


FIGURE 5.14. Failure cases of the monocular depth prediction module.

5.4.5 Limitations

We have observed a few limitations of our method. Here we analyze some failure cases of the depth prediction module and the point cloud module.

Failure cases of DPM. We show some typical failure cases of DPM in Fig. 5.14. 1) Out of focus. This may be due to the fact that our training images are all-in-focus. 2) Paintings (or mirrors) can cause ambiguity to the network. 3) Cartoons. Since the domain gap exists between cartoons and the real photos, the network does not work well. We believe that the above problems can be largely solved with more training data.

Failure cases of PCM. Our PCM cannot recover accurate focal length or depth shift when the scene does not have enough geometric cues, *e.g.*, when the whole image is mostly a wall or a sky region. The accuracy of our method will also decrease with images taken from uncommon view angles (*e.g.*, top-down) or extreme focal lengths. More diverse 3D training data may address these failure cases. In addition, our method does not model the effect of radial distortion from the camera and thus the reconstructed scene shape can be distorted in cases with severe radial distortion. Studying how to recover the radial distortion parameters using our PCM can be an interesting future direction.

5.5 Conclusion

In summary, we have presented, to our knowledge, the first fully data driven method that reconstructs 3D scene shapes from single monocular images. To recover the shift and focal length for 3D reconstruction, we have proposed to use point cloud networks trained on datasets with known global depth shifts and focal lengths. This approach has demonstrated strong generalization capabilities, and we are under the impression that it may be helpful for related depth and 3D reconstruction tasks. Our extensive experiments verify the effectiveness of our scene shape reconstruction method and the superior capability to generalize to unseen data.

Statement of Authorship

Title of Paper	Towards Domain-agnostic Depth Completion
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., & Shen, C. (2021). Learning to recover 3d scene shape from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 204-213).

Principal Author

Name of Principal Author (Candidate)	Wei Yin
Contribution to the Paper	Design new methods and conduct the experiments.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 10/13/2021

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Chunhua Shen
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Name of Co-Author	Oliver Wang
Contribution to the Paper	Discussion and review the paper.
Signature	Date 10/13/2021

Name of Co-Author	Simon Niklaus		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	October 13, 2021

Name of Co-Author	Simon Chen		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/13/2021

Name of Co-Author	Jianming Zhang		
Contribution to the Paper	Discussion and review the paper.		
Signature		Date	10/13/2021

Please cut and paste additional co-author panels here as required.

Chapter 6

Metric Scene Reconstruction With Sparse Points

6.1 Introduction

In last chapters, our proposed geometric constraints, learning objective, and two-stage reconstruction pipeline methods have solved the 3D scene reconstruction from monocular images. However, the predicted depth and reconstructed shape are scale-invariant. The metric information cannot be recovered.

In this chapter, we aim to recover the metric depth and do the metric reconstruction. As the single image input cannot provide enough metric information, we propose to combine the single image and a sparse depth map.

6.2 Background

Accurate metric depth is important for many computer vision applications, in particular 3D perception [121, 143] and reconstruction [94, 92]. Typically, depth is obtained by using direct range sensors such as LiDAR or Time-of-Flight (ToF) sensors included on modern mobile phones, or multi-view stereo methods [118, 158]. However, neither of these sources can provide dense depth from the perspective of the camera. For example, LiDAR sensors capture depth in a linear scanning pattern, ToF sensors are lower resolution and fail at specular or distant surfaces, and multi-view reconstruction methods [118, 158, 167] only provide confident depth at textured regions and are range limited by the camera baseline (the iPhone rear stereo camera has a maximum depth of 2.5 meters).

Existing algorithms that obtain dense depth from sparse depth inputs can be classified into two categories according to the input sparsity pattern: depth inpainting methods that fill large holes [168, 120, 55], and sparse depth completion methods that fill sparsely distributed depth measurements [22, 96, 153, 102, 21]. When working on a specific sparsity pattern, e.g., on either NYU [124] or KITTI [132], recent approaches [96, 22, 20, 102, 58] such as NLSPN [96] can obtain impressive performance. However, in real-world scenarios, the specific sparsity pattern may be unknown at

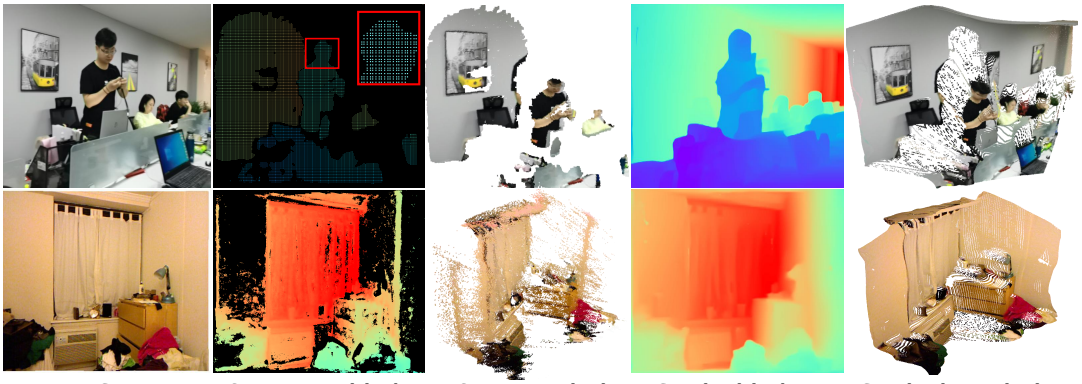


FIGURE 6.1. Our method fills in missing information in different types of sparse depth maps. A single model can be used to complete depth from a mobile phone Time-of-Flight sensor (top row), and a multi-view stereo algorithm [118] (bottom row).

training time, as it is a function of hardware, software, as well as the configuration of the scene itself.

In this paper, we revisit existing methods and analyze the gap between the performance on the training setup and downstream applications, and we find that existing depth completion methods suffer from the following key limitations. First, their methods work best on one specific sparsity pattern, while they poorly generalize to other types of sparse depth (e.g., from Kinect depths to smart phone depths). Second, they are sensitive to noise and outliers produced from the depth capture process. To address these issues, we propose a simple yet effective method towards robust depth completion.

Our method provides the following improvements. First, inspired by domain randomization methods [131, 130, 164], we analyze the existing set of common sparsity patterns and create a diverse set of synthetic sparsity patterns to train our model. To improve the cross-domain generalization ability, we follow recent monocular depth prediction methods [162, 104] and utilize a diverse training dataset which consist of multiple depth sources. Furthermore, to make our method robust to noise, we leverage the depth map predicted by a well-trained single image depth prediction method as a data-driven scene prior. Such approaches learn a strong prior on diverse scenes [104], but their predicted depths have unknown shift and scale due to the training data used (stereo images with unknown baseline). By incorporating sparse metric depth cues and a single image relative depth prior, our method is able to robustly produce a dense metric depth map. It can also be run in a recurrent manner to further refine the produced metric depth map.

We show that this simple approach improves upon the state of the art for depth completion, especially under the zero-shot cross-dataset generalization setting where the specific sparsity pattern is unknown during training. We also show that our method is more robust to noisy sparse depth measurements.

In conclusion, our main contributions are as follows.

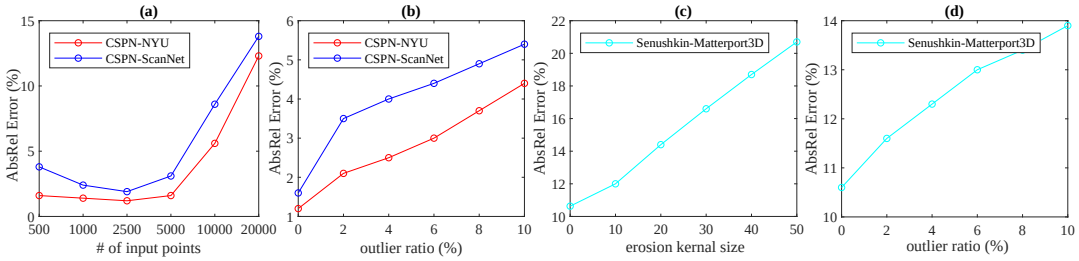


FIGURE 6.2. Robustness analysis. We analyze the performance of CSPN [22] (completion) and Senushkin *et al.* [120] (inpainting) in terms of input point numbers/patterns (a, c) and outlier ratios (b, d). CSPN is trained on NYU [124], and we evaluate it on both NYU and ScanNet [25]. Senushkin *et al.* is trained and evaluated on Matterport3D [11].

- We analyze existing depth completion methods in terms of generalization across different scenes and robustness to noise.

- We propose a new method that incorporates a data driven single-image prior and effective data augmentation techniques for domain-agnostic depth completion. To the best of our knowledge, we are the first to target the cross-domain depth completion problem. We show that our method generalizes well to various types of depth inputs, including those from special range sensors, mobile phones, and stereo methods.

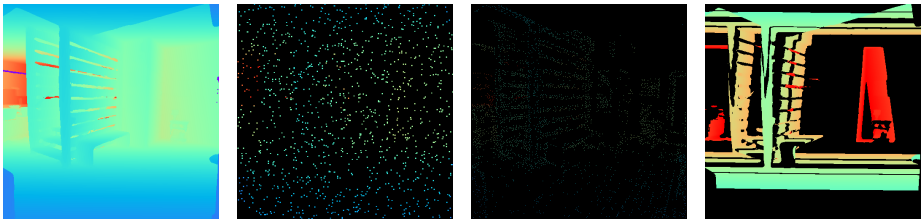
6.3 Analysis of Existing Methods

In this section, we evaluate two state-of-the-art depth completion methods in terms of their performance with different sparsity patterns, dataset generalization, and robustness to noise. To do this, we perturb the sparsity pattern of the input depth in various ways and add noise to it. We also evaluate methods on their zero-shot cross-dataset generalization performance [104] (evaluating on a different dataset than the models were trained on). We chose two methods trained on the NYU benchmark and the Matterport3D benchmark for this analysis, CSPN [21] and Senushkin *et al.* [120]. The former is designed to complete very sparse depth with only hundreds of sparse points, while the latter is designed to complete contiguous holes. We use the code and the model weights provided by the authors for this evaluation.

For the CSPN [21] method that is trained on NYU, we vary the number of measured/input points from 500 to 20000. Senushkin *et al.* [120] is trained on Matterport3D with the task of completing holes of depth maps. We erode the valid depth regions with different kernel sizes to control the number of valid points on Matterport3D. From Fig. 6.2 (a) and (c), we observe that their methods are sensitive to the variation of the number of valid points. Besides, as outliers are unavoidable in many applications, we also simulate depth noise by sampling $0\% - 10\%$ points from the sparse depth and multiplying the original depth with a random factor from $0.1 - 2$.

TABLE 6.1. Robustness to different sparse depth patterns (AbsRel Error).

	Uniform (2500 points)			Features (2500 points)		
	NYU	ScanNet	Matterport3D	NYU	ScanNet	Matterport3D
CSPN [20]	0.011	0.02	0.279	0.019	0.138	0.519
Senushkin *et al.* [120]	0.773	0.621	0.695	0.769	0.608	0.667



(A) (B) (C) (D)

FIGURE 6.3. Visualization of sampled sparse depths. We simulate three different patterns from the dense depth (a) to train models: random uniform sampling (b), feature point based sampling (c), and region-based sampling (d).

Fig. 6.2 (b) and (d) show that their performance decreases a lot with outliers in the input. Furthermore, to evaluate the generalization to other datasets, we also test CSPN on ScanNet (CSPN was trained on NYU). Fig. 6.2 shows the evaluation results.

Furthermore, we study the robustness to different sparse depth patterns. For both methods, we input two additional kinds of sparse depth, i.e. uniform sparse depth (Uniform) and the sparse depth whose valid points are detected by the FAST [106] feature detector (Features). It shows that CSPN is sensitive to the point distribution, while Senushkin *et al.* [120] fails on other kinds of sparse depths. Results are summarized in Tab. 6.1

6.4 Our Method

We now introduce our approach, which is designed to be robust to noise, applicable to different types of sparse depth, and to generalize well to unseen datasets.

6.4.1 Model architecture.

Our depth completion model takes as input the RGB image, sparse depth, and a guidance depth map, and it outputs a completed depth map. We use the ESANet-R34-NBt1D network with the ResNet-34 backbone proposed by Seichter *et al.* [119] for depth completion, and we use the affine-invariant depth predicted by the method from Yin *et al.* [162] as the guidance depth map. We sample 36000 images from Taskonomy [165], DIML [63], and TartainAir [141] as the training data.

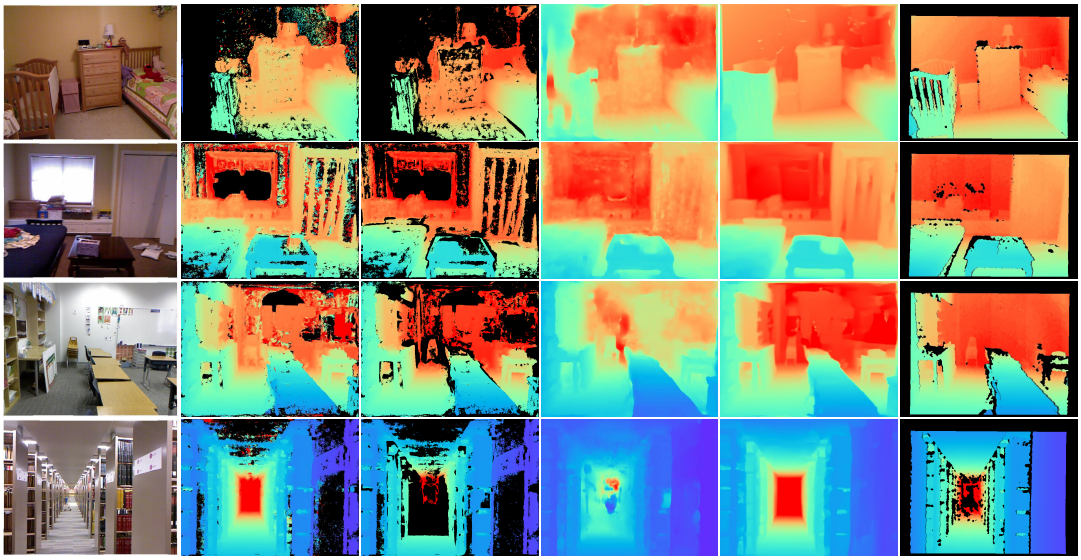


FIGURE 6.4. Qualitative comparison for completing noisy sparse depth. The noisy sparse depths are obtained by masking COLMAP [117] depths. Our completed results have less outliers and errors.

6.4.2 Training data generation.

As we cannot access enough data to cover the diverse sparsity patterns of all possible downstream applications during training, we instead simulate a set of various sparsity patterns. This approach is motivated by domain randomization [131, 130, 164] methods that train models on simulated data and show that the domain gap to real data can be reduced by randomizing the rendering in the simulator.

We categorize the sparse depth patterns into three main classes, which are illustrated in Fig. 6.3. During training, we sample sparse points from the dense ground truth depth and try to recover the dense depth map.

- **Uniform.** We sample uniformly distributed points, from hundreds to thousands of points, to simulate the sparsity pattern from the low-resolution depths, e.g., those captured by ToF sensors on mobile phones.

- **Features.** In order to simulate the sparsity pattern from structure-from-motion and multi-view stereo methods, where high confidence depth values are produced only at regions with distinct/matchable features, we apply the FAST [106] feature detector that samples points on textured regions and particularly image corners.

- **Holes.** Commodity-grade depth sensors cannot capture depth on bright, transparent, reflective and distant surfaces. Therefore, multiple large coherent regions may be missing. We simulate this by 1) masking the depth in a random polygonal region, 2) masking regions at a certain distance, or 3) masking the whole image with the exception of a polygonal region.

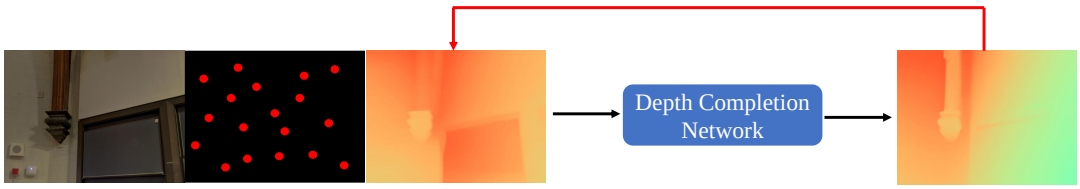


FIGURE 6.5. Our method takes an RGB image, sparse depth, and guidance map as input, and it outputs dense depth. We can iterate the network several times, replacing the guidance map with the output of the previous iteration.

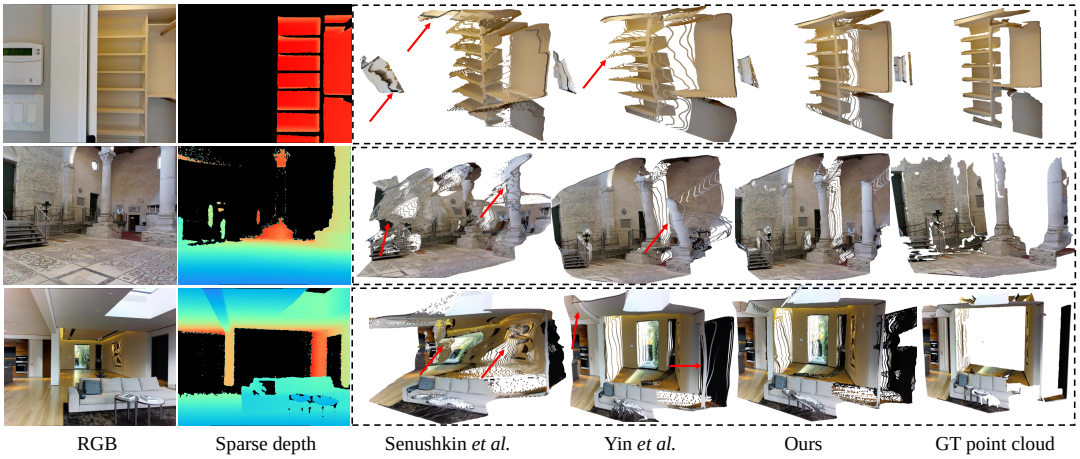


FIGURE 6.6. Qualitative comparison of depth and reconstructed 3D shape. Our completed metric depth has finer details and the reconstructed 3D shape is more accurate than previous methods.

To improve the diversity of these patterns, we augment each type of sparse depth by controlling the number of valid points, or dilating or eroding the valid regions with different kernel sizes, and combining sparsity patterns together.

6.4.3 Improving the robustness to outliers.

Outliers and depth sensor noise are unavoidable in any depth acquisition method. Most of previous methods only take an RGB image and a sparse depth as the input, and they do not have any extra source of information with which it could distinguish the outliers. However, our method leverages a data prior from the single image depth network, which can help resolve incorrect constraints. In order to encourage the network to learn this, we add outliers during training. Specifically, we randomly sample several points and scale their depth by a random factor from 0.1-2.

6.4.4 Iterative refinement.

As our method takes a guidance map as input, we can naturally extend it to an iterative refinement method, where we recursively feed the output of the network back into itself as the guidance map and compute a subsequent inference pass. We found that the method tends to converge in 3 iterations. Our framework is illustrated in Fig. 6.5.

TABLE 6.2. Depth completion results on the NYU dataset. Our method (not trained on NYU) shows on par performance with state-of-the-art methods that are trained on NYU.

Methods	RMSE(m)↓	AbsRel↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
S2D [89]	0.230	0.044	97.1	99.4	99.8
S2D+SPN [87]	0.172	0.031	98.3	99.7	99.9
DepthCoeff [59]	0.118	0.013	99.4	99.9	-
CSPN [20]	0.117	0.016	99.2	99.9	100.0
CSPN++ [22]	0.116	-	-	-	-
DeepLiDAR [102]	0.115	0.022	99.3	99.9	100.0
DepthNormal [153]	0.112	0.018	99.5	99.9	100.0
NLSP [96]	0.092	0.012	99.6	99.9	100.0
MiDaS [104](Guidance)	0.513	0.110	88.6	98.1	99.6
Yin *et al.* [162](Guidance)	0.402	0.090	91.3	98.0	99.5
Ours-baseline	0.210	0.036	98.4	99.6	99.9
Ours-W MiDaS [104]	0.199	0.024	98.6	99.6	99.9
Ours-W Yin *et al.* [162]	0.183	0.022	98.7	99.7	99.9

TABLE 6.3. Depth completion results on the Matterport3D dataset. Our method (not trained on Matterport3D) is comparable with state-of-the-art methods that are trained on Matterport3D.

Methods	RMSE(m)↓	MAE(m)↓	$\delta_{1.05}$ ↑	$\delta_{1.1}$ ↑	δ_1 ↑	δ_2 ↑	δ_3 ↑
Huang *et al.* [55]	1.092	0.342	66.1	75.0	85.0	91.1	93.6
Zhang *et al.* [168]	1.316	0.461	65.7	70.8	78.1	85.1	88.8
Senushkin *et al.* [120]	1.028	0.299	71.9	80.5	89.0	93.2	95.0
Yin *et al.* [162](Guidance)	2.06	1.13	17.9	29.8	50.7	72.3	83.4
MiDaS [104] (Guidance)	3.45	2.01	13.2	21.8	37.5	54.8	66.4
Ours-baseline	2.35	0.574	68.9	79.2	88.1	93.5	96.0
Ours-W MiDaS [104]	1.49	0.448	67.8	76.3	85.0	91.0	94.5
Ours-W Yin *et al.* [162]	1.03	0.320	71.2	79.0	87.1	93.1	96.0

6.5 Experiments

In this section, we conduct several experiments to demonstrate the effectiveness of our approach. We evaluate the robustness of our method to noisy sparse depth, the generalization to different sparse depth patterns on zero-shot testing data, and the effectiveness of recurrent refinement. We include a baseline method (Ours-baseline) that trains a model without the guidance image, that is, directly predicting the complete depth from RGB and sparse depth input.

6.5.1 Comparison with State-of-the-art Depth Completion Methods

Quantitative comparisons with current state-of-the-art methods on the NYU [124] and the Matterport3D [11] benchmark are summarized in Tab. 6.2 and Tab. 6.3. Note that our model has not been trained on these two datasets. Although these two benchmarks have two different sparse depth types, we use a single model for evaluation on both of them. On NYU, we can see that our method can achieve performance on par with previous methods. We can achieve comparable performance with previous methods, and better accuracy than the baseline. Comparing with our baseline (Ours-baseline),

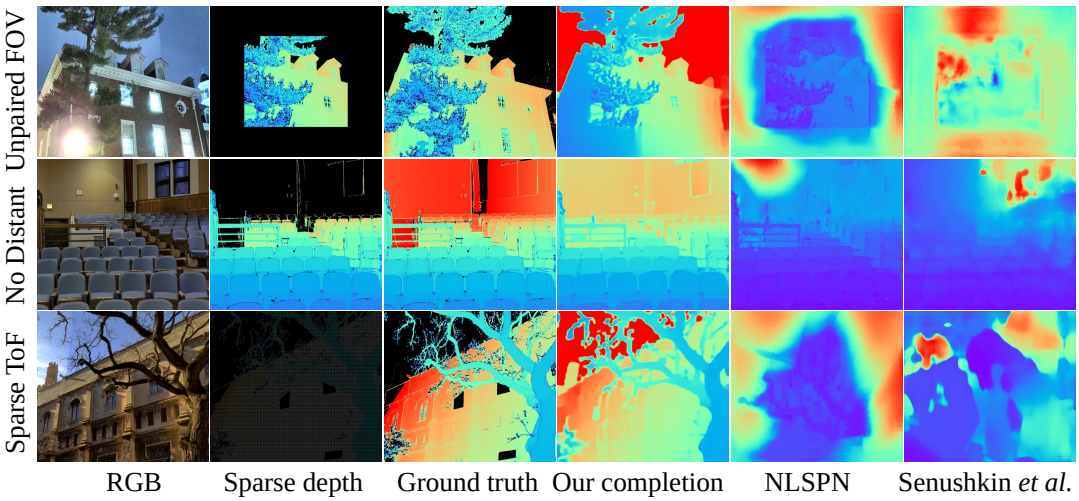


FIGURE 6.7. Qualitative completion results on the DIODE [135] dataset. Note that none of the methods is trained on this dataset. We compare our method with Senushkin *et al.* [120] and NLSP [96] using 3 different unseen sparsity patterns.

using a pretrained guidance map shows improved performance. When compared to Yin [162], which is our input guidance map, we see that even with only 500 sparse points, the performance is improved significantly. Furthermore, we analyze the effect of different guidance map inputs. We employ two monocular depth estimation methods to create the guidance map, i.e. MiDaS [104] and Yin *et al.* [162]. Note that 'Ours-W MiDaS' only takes the MiDaS depth as the input during test. We find that they both work well as the guidance map generator in our system. As depths from Yin *et al.* [162] are employed as the guidance map during training, taking their depth can achieve better results in test.

Moreover, the qualitative comparison on Matterport3D is illustrated in Fig. 6.6. Although Senushkin *et al.* [120] can achieve better accuracy than ours (it was trained on the same data), we find that our reconstructed scene structure is better. Yin *et al.* [162] is the single image depth map with a predicted focal length and shift. Although their reconstructed structure distorts in some regions (see the regions highlighted by red arrows), our completion network can rectify it by leveraging the sparse depth input.

6.5.2 Generalization

To evaluate the generalization to zero-shot testing data and different sparse depth types, we compare our methods with current state-of-the-art methods on 3 different datasets with 3 different sparse depth patterns. We simulate 3 new sparse depth types: 1) Unpaired FOV. We mask 25% region around 4 borders of the ground truth depth. 2) Sparse ToF. We simulate Huawei ToF sensor to sample the depth from the ground truth every 3 pixels horizontally and vertically. 3) No Distant. We mask the most 50% distant regions. Note that all these simulated patterns have not be utilized in our training.

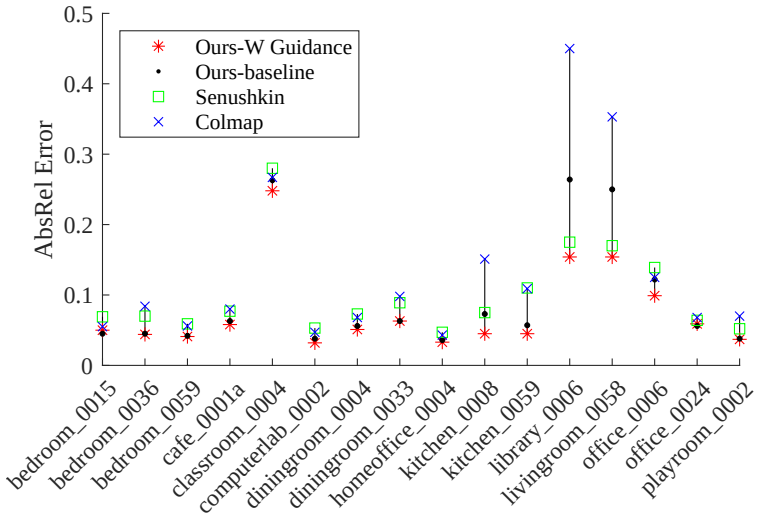


FIGURE 6.8. Completion results on 16 scenes sampled from NYU [124]. The input depth is noisy, which is generated using COLMAP [117].

Results are summarized in Tab. 6.4. We can find that although NLSP [96] and Senushkin *et al.* [120] can achieve state-of-the-art performance on NYU and Matterport3D respectively, they cannot generalize to different types of sparse depth and other datasets. We conjecture that it is because they are trained on one type only. By contrast, our method can achieve comparable performance on different datasets. Furthermore, comparing to our baseline method, using the guidance map can consistently improve the performance over all datasets and sparse depth types. The quantitative comparison of completing 3 different patterns on the DIODE dataset is demonstrated in Fig. 6.7.

TABLE 6.4. Comparison of state-of-the-art methods on zero-shot test datasets. We use 3 different patterns as the model input to analyze the robustness of depth completion methods.

	NYU			ScanNet			DIODE		
	AbsRel↓			AbsRel↓			AbsRel↓		
	Unpaired	FOV Sparse	ToF No Distant	Unpaired	FOV Sparse	ToF No Distant	Unpaired	FOV Sparse	ToF No Distant
NLSP [96]	0.150	0.190	0.114	0.716	1.413	0.202	6.684	11.370	1.005
Senushkin *et al.* [120]	0.224	0.615	0.093	0.255	0.793	0.166	0.687	6.120	0.623
Ours-baseline	0.046	0.018	0.041	0.049	0.022	0.047	0.150	0.143	0.144
Ours-W Guidance	0.031	0.013	0.030	0.028	0.014	0.037	0.139	0.111	0.137

6.5.3 Completing Noisy Depths

In order to obtain real world noisy data, we use COLMAP [117] to densely reconstruct a scene. We sample 16 scenes from NYU with over 4000 images, and we use Yin *et al.* [162] to mask the most noisy regions from the COLMAP depths.

As the noisy sparse depth pattern most resemble the 'Holes' type, we compare to Senushkin *et al.* [120]. Quantitative results are illustrated in Fig. 6.8. We can see that our method consistently performs the best on all test scenes. Furthermore, comparing with the baseline (no guidance map), our approach is more robust to noise (see 'library_0006' and 'living_room_0058'). Moreover, the qualitative comparisons are illustrated in Fig. 6.4. We can see that our completed depths have much less outliers and noise (see the wall).

6.5.4 Effectiveness of Recurrent Refinement

In this section, we demonstrate the effectiveness of our proposed iterative refinement. We recursively replace the guidance map with the previous stage output as the current stage guidance map. The quantitative evaluation is summarized in Tab. 6.5. 'W refine' feeds inputs in the original resolution and recursively refines outputs, while 'W/o refine' means that we don't employ the recursive refinement. We can see that our proposed refinement method can improve the depth accuracy and edges quality slightly (see the comparison on iBims-1). Furthermore, the qualitative results are shown in Fig. 6.9, in which we can see for example that the fine scale structure, such as the branches of the tree, are improved after several iterations.

TABLE 6.5. Effectiveness of the refinement. We refine the depth 3 times, which leads to the improved accuracy on full images (the first three datasets) and edge regions (the ibims-1 dataset).

	NYU				ScanNet				DIODE				iBims-1			
	Uniform		Holes		Uniform		Holes		Uniform		Holes		Uniform		Holes	
	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	$\varepsilon^{\text{acc}}_{\text{DBE}} \downarrow$	$\varepsilon^{\text{comp}}_{\text{DBE}} \downarrow$	$\varepsilon^{\text{acc}}_{\text{DBE}} \downarrow$	$\varepsilon^{\text{comp}}_{\text{DBE}} \downarrow$
W/o refine	0.024	0.99	0.015	0.99	0.021	0.99	0.014	0.99	0.22	0.88	0.14	0.92	2.20	4.66	1.34	3.85
W refine	0.022	0.99	0.015	0.99	0.018	0.99	0.012	0.99	0.21	0.88	0.13	0.92	2.07	4.43	1.25	3.77

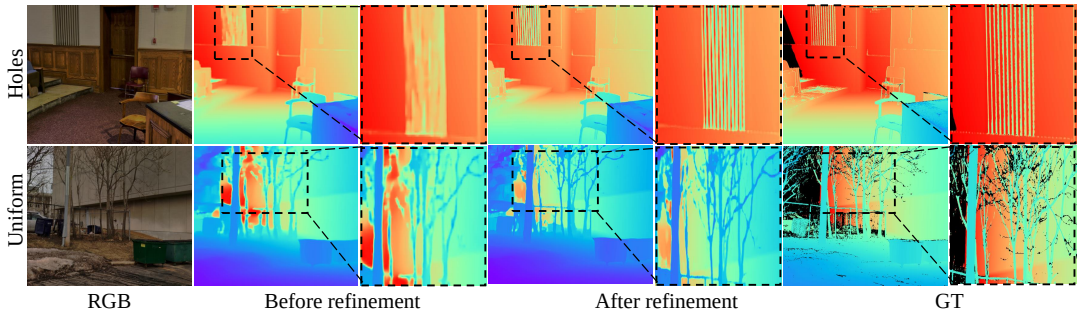


FIGURE 6.9. Qualitative results of the proposed iterative refinement.

6.5.5 Completing Mobile Phone Sensor Depth

Many smartphones have been equipped with cheap 3D sensors. To evaluate the robustness of our completion method for noisy phone-captured depths, we test on 3 different phones, i.e. iPhone 12, Huawei Mate 30, and Google Pixel 3. In each case, the acquired depths are different. iPhone uses stereo matching method to obtain the depth, which is normalized to $0-255$ and saved as the inverse depth. Huawei provides a low-resolution dense depth, i.e. 180×240 pixels, which is captured by a ToF sensor. Pixel uses the dual-pixel sensor, which has a very small baseline, to capture the depth. For the Pixel captured depth, we apply the provided confidence map to filter the most noisy regions and only rely on confidence depth values. The qualitative comparison is illustrated in Fig. 6.10. We can see that our completed depth has much less outliers than current methods.

6.5.6 Ablation of Synthetic Sparsity Patterns

This study aims to investigate the effectiveness of different simulated sparsity patterns. We remove one of the proposed patterns during training and evaluate them on 3 zero-shot datasets with different patterns. On NYU and ScanNet, we simulate the pattern 'Sparse ToF' and 'Unpaired FOV', while we use the provided sparse depth on Matterport3D. All models have been trained with the same number of epochs, and identical parameters other than the sparsity patterns. Results are summarized in Tab. 6.6. We observe that when missing the simulated sparse depth pattern, the performance on the most-related testing data will decrease. For example, the model trained without the 'Holes' pattern has worse performance than others on ScanNet and Matterport3D. Therefore, our proposed sparse depth generation method can improve the cross-domain generalization.

6.6 Limitations

We have observed a few limitations of our method. Here we analyze some failure cases. Our method takes as input a depth map from the monocular depth estimation model as the guidance. Although rare, when the guidance depth map has significant errors, it will inevitably have adverse effects on the depth completion. Furthermore, although

TABLE 6.6. Effects of different simulated sparsity patterns. The model is trained on the simulated patterns except the one specified by 'W/o' and evaluated on zero-shot datasets.

	AbsRel % ↓		
	NYU (Sparse ToF)	ScanNet (Unpaired FOV)	Matterport3D (Original pattern)
W/o Features	0.036	0.047	0.122
W/o Uniform	0.021	0.044	0.121
W/o Holes	0.012	0.12	0.165
Ours-Full	0.013	0.028	0.120

our method is robust to outliers and noises, the completion quality will decrease if over 50% of sparse depths are outliers.

6.7 Conclusion

In this paper, we proposed a simple system for depth completion. Our method leverages a single image depth prior and allows for dense metric scene reconstruction when sparse depth sensors are available. Our approach generates results that are valid over a variety of sensor types and is robust to the presence of sensor noise. Our method is also able to refine the details of the completed depth map through an iterative process. Depth completion systems like ours have applications in mobile phones which are increasingly commonly being shipped with sparse depth capture devices.

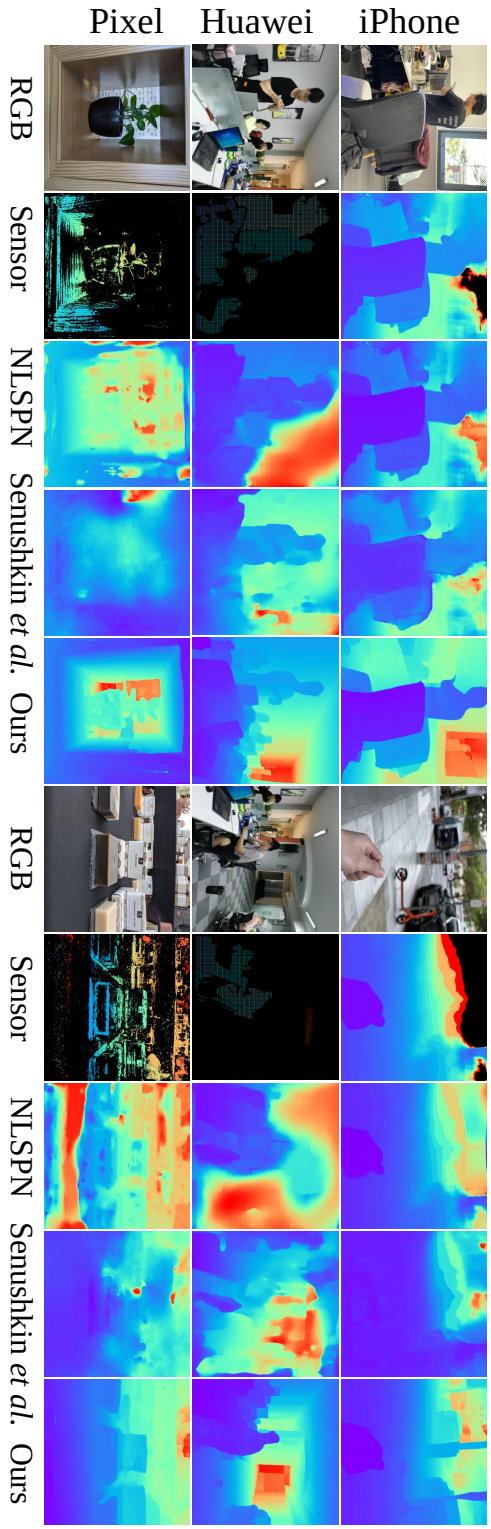


FIGURE 6.10. Completion of the phone-captured depths. Our method is more robust to different depth sensors than previous state-of-the-art methods.

Chapter 7

Conclusion and Future Directions

7.1 Conclusion

In this thesis, we propose a novel method to solve the problem of 3D reconstruction from a single in-the-wild image.

In Chapter 3, we propose a high-order geometric constraint for accurate monocular depth estimation, which not only boosts the depth accuracy significantly but also ensures high-quality 3D reconstruction from the depth. Although previous methods using the pixel-wise regression or classification loss can achieve high accuracy on depth, the reconstructed 3D point cloud from depth is far away from the original shape. To handle this challenge, we lift the depth to the 3D space and propose a global geometric constraint, termed virtual normal loss. Compared with other local geometric constraint, our method is more robust to noise.

In Chapter 4, we aim to solve the generalization issue of monocular depth estimation. Current learning metric depth methods can only work well on a specific scene, while learning relative depth methods cannot recover high-quality 3D shapes. To address this problem, we construct a large-scale and diverse dataset, and then propose to learn the affine-invariant depth on it. Our method ensures both high generalization and high-quality geometric shapes of scenes.

In Chapter 5, we propose a novel framework to solve the 3D reconstruction from a single image. In previous chapters, our proposed methods have solved the robust depth estimation on diverse scenes. To reconstruct the point cloud from depth, the depth shift and camera focal length should be recovered. We propose to use point cloud networks trained on datasets with known global depth shifts and focal lengths to predict them. Combined with the depth estimation stage, the 3D shape can be recovered from the single image. Our approach has demonstrated strong generalization capabilities on diverse scenes.

In Chapter 6, we propose a depth completion method to robustly recover metric depth. When the sparse depth is available from depth sensors, our method leverages a single image depth prior and allows for dense metric scene reconstruction. Compared with current state-of-the-art methods, our approach is more robust to a variety of sensor types and depth noise.

7.2 Future Directions

In this thesis, my proposed methods can solve the problem of recovering the 3D scene shape from a single in-the-wild image. Then I have started to think about how to obtain dense 3D reconstruction from a video or sparse views. Previous SLAM-based methods can only do the very sparse reconstruction on some feature points. And dynamic objects in the views will be ignored. How to densely reconstruct all of them from a free video will be the next challenge in the community. I hope my methods can serve as a strong scene shape prior for the video scene reconstruction.

Bibliography

[1] Ibraheem Alhashim and Peter Wonka. "High Quality Monocular Depth Estimation via Transfer Learning". In: abs/1812.11941 (2018).

[2] Iro Armeni et al. "Joint 2d-3d-semantic data for indoor scene understanding". In: *arXiv: Comp. Res. Repository* (2017), p. 1702.01105.

[3] Jonathan Barron and Jitendra Malik. "Shape, illumination, and reflectance from shading". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.8 (2014), pp. 1670–1687.

[4] Yoshua Bengio et al. "Curriculum learning". In: *Proc. Int. Conf. Mach. Learn.* ACM. 2009, pp. 41–48.

[5] Jia-Wang Bian et al. "Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video". In: *Adv. Neural Inform. Process. Syst.* 2019.

[6] Jiawang Bian et al. "Unsupervised scale-consistent depth and ego-motion learning from monocular video". In: *Adv. Neural Inform. Process. Syst.* 2019, pp. 35–45.

[7] D. J. Butler et al. "A naturalistic open source movie for optical flow evaluation". In: *Eur. Conf. Comput. Vis.* Springer. 2012, pp. 611–625.

[8] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. "Estimating depth from monocular images as classification using deep fully convolutional residual networks". In: *IEEE TCSVT* (2017).

[9] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. "Estimating depth from monocular images as classification using deep fully convolutional residual networks". In: *IEEE Trans. Circuits Syst. Video Technol.* 28.11 (2017), pp. 3174–3182.

[10] Ayan Chakrabarti, Jingyu Shao, and Greg Shakhnarovich. "Depth from a single image by harmonizing overcomplete local network predictions". In: *Adv. Neural Inform. Process. Syst.* 2016, pp. 2658–2666.

[11] Angel Chang et al. "Matterport3D: Learning from RGB-D Data in Indoor Environments". In: *Int. Conf. 3D. Vis.* (2017).

[12] Liang-Chieh Chen et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proc. Eur. Conf. Comp. Vis.* (2018).

[13] Liang-Chieh Chen et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *Eur. Conf. Comput. Vis.* 2018.

[14] Weifeng Chen, Shengyi Qian, and Jia Deng. "Learning single-image depth from videos using quality assessment networks". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 5604–5613.

[15] Weifeng Chen et al. "OASIS: A Large-Scale Dataset for Single Image 3D in the Wild". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020.

[16] Weifeng Chen et al. "OASIS: A Large-Scale Dataset for Single Image 3D in the Wild". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020, pp. 679–688.

[17] Weifeng Chen et al. "Single-image depth perception in the wild". In: *Adv. Neural Inform. Process. Syst.* 2016, pp. 730–738.

[18] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. "Structure-Aware Residual Pyramid Network for Monocular Depth Estimation". In: 2019.

[19] Zhao Chen et al. "Estimating depth from rgb and sparse sensing". In: *Eur. Conf. Comput. Vis.* 2018, pp. 167–182.

[20] Xinjing Cheng, Peng Wang, and Ruigang Yang. "Depth estimation via affinity learned with convolutional spatial propagation network". In: *Eur. Conf. Comput. Vis.* 2018, pp. 103–119.

[21] Xinjing Cheng, Peng Wang, and Ruigang Yang. "Learning Depth with Convolutional Spatial Propagation Network". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2019).

[22] Xinjing Cheng et al. "Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion". In: *AAAI Conf. Artificial Intelligence*. Vol. 34. 07. 2020, pp. 10615–10622.

[23] Jaehoon Cho et al. "A Large RGB-D Dataset for Semi-supervised Monocular Depth Estimation". In: abs/1904.10230 (2019).

[24] Christopher Choy, JunYoung Gwak, and Silvio Savarese. "4d spatio-temporal convnets: Minkowski convolutional neural networks". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 3075–3084.

[25] Angela Dai et al. "Scannet: Richly-annotated 3d reconstructions of indoor scenes". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2017, pp. 5828–5839.

[26] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *IEEE Conf. Comput. Vis. Pattern Recog.* Ieee. 2009, pp. 248–255.

[27] Raul Diaz and Amit Marathe. "Soft labels for ordinal regression". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 4738–4747.

[28] Tom van Dijk and Guido de Croon. "How Do Neural Networks See Depth in Single Images?" In: *Int. Conf. Comput. Vis.* 2019, pp. 2183–2191.

[29] David Eigen and Rob Fergus. "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture". In: *Int. Conf. Comput. Vis.* 2015, pp. 2650–2658.

[30] David Eigen, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network". In: *Adv. Neural Inform. Process. Syst.* 2014, pp. 2366–2374.

[31] Jose M Facil et al. "CAM-Convs: camera-aware multi-scale convolutions for single-view depth". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 11826–11835.

[32] Xiaohan Fei, Alex Wang, and Stefano Soatto. "Geo-Supervised Visual Depth Prediction". In: *arXiv: Comp. Res. Repository.* Vol. abs/1807.11130. 2018.

[33] David F Fouhey, Abhinav Gupta, and Martial Hebert. "Data-driven 3D primitives for single image understanding". In: *Int. Conf. Comput. Vis.* 2013, pp. 3392–3399.

[34] David Ford Fouhey, Abhinav Gupta, and Martial Hebert. "Unfolding an indoor origami world". In: *Eur. Conf. Comput. Vis.* Springer. 2014, pp. 687–702.

[35] Huan Fu et al. "Deep Ordinal Regression Network for Monocular Depth Estimation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* 2018.

[36] Huan Fu et al. "Deep Ordinal Regression Network for Monocular Depth Estimation". In: *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* 2018, pp. 2002–2011.

[37] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition.* Elsevier, 2013.

[38] Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining.* Springer, 2015.

[39] Ravi Garg et al. "Unsupervised cnn for single view depth estimation: Geometry to the rescue". In: *Eur. Conf. Comput. Vis.* Springer. 2016, pp. 740–756.

[40] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite". In: *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE. 2012, pp. 3354–3361.

[41] Andreas Geiger et al. "Vision meets robotics: The KITTI dataset". In: *Int. J. Robot. Res.* 32.11 (2013), pp. 1231–1237.

[42] Clément Godard et al. "Digging into self-supervised monocular depth estimation". In: *Int. Conf. Comput. Vis.* 2019, pp. 3828–3838.

[43] Clément Godard et al. "Digging into Self-Supervised Monocular Depth Prediction". In: *Int. Conf. Comput. Vis.* 2019.

[44] Xiaoyang Guo et al. "Learning monocular depth by distilling cross-domain stereo networks". In: *Eur. Conf. Comput. Vis.* 2018, pp. 484–500.

[45] Saurabh Gupta et al. "Learning rich features from RGB-D images for object detection and segmentation". In: *Eur. Conf. Comput. Vis.* Springer. 2014, pp. 345–360.

[46] Guy Hacohen and Daphna Weinshall. "On The Power of Curriculum Learning in Training Deep Networks". In: *Proc. Int. Conf. Mach. Learn.* 2019, pp. 2535–2544.

[47] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[48] Kaiming He et al. "Deep residual learning for image recognition". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2016, pp. 770–778.

[49] Daniel Herrera, Juho Kannala, Janne Heikkilä, et al. "Depth map inpainting under a second-order smoothness prior". In: *Scandinavian Conference on Image Analysis*. Springer. 2013, pp. 555–566.

[50] Stefan Hinterstoisser et al. "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes". In: *Int. Conf. Comput. Vis.* IEEE. 2011, pp. 858–865.

[51] Heiko Hirschmuller. "Stereo processing by semiglobal matching and mutual information". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 30.2 (2007), pp. 328–341.

[52] Yannick Hold-Geoffroy et al. "A perceptual measure for deep single image camera calibration". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 2354–2363.

[53] Junjie Hu et al. "Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps With Accurate Object Boundaries". In: *Proc. Winter Conf. on Appl. of Comp0 Vis.* 2019.

[54] Yiwen Hua et al. "Holopix50k: A Large-Scale In-the-wild Stereo Image Dataset". In: *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.* 2020.

[55] Yu-Kai Huang et al. "Indoor depth completion with boundary consistency and self-attention". In: *Proc. Workshop of IEEE Int. Conf. Comp. Vis.* 2019, pp. 0–0.

[56] Lam Huynh et al. "Guiding monocular depth estimation using depth-attention volume". In: *Eur. Conf. Comput. Vis.* Springer. 2020, pp. 581–597.

[57] Eddy Ilg et al. "Flownet 2.0: Evolution of optical flow estimation with deep networks". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2017, pp. 2462–2470.

[58] Saif Imran, Xiaoming Liu, and Daniel Morris. "Depth Completion with Twin Surface Extrapolation at Occlusion Boundaries". In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2021).

[59] Saif Imran et al. "Depth coefficients for depth completion". In: *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE. 2019, pp. 12438–12447.

[60] Jianbo Jiao et al. "Look Deeper into Depth: Monocular Depth Estimation with Semantic Booster and Attention-Driven Loss". In: *Eur. Conf. Comput. Vis.* 2018, pp. 53–69.

[61] Olga A Karpenko and John Hughes. "SmoothSketch: 3D free-form shapes from complex sketches". In: *ACM SIGGRAPH*. 2006, pp. 589–598.

[62] Kevin Karsch, Ce Liu, and Sing Bing Kang. "Depth transfer: Depth extraction from video using non-parametric sampling". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.11 (2014), pp. 2144–2158.

[63] Youngjung Kim et al. "Deep monocular depth estimation via integration of global and local predictions". In: *IEEE Trans. Image Process.* 27.8 (2018), pp. 4131–4144.

[64] Klaas Klasing et al. "Comparison of surface normal estimation methods for range sensing applications". In: *Proc. Int. Conf. on Robotics and Automation*. IEEE. 2009, pp. 3206–3211.

[65] Tobias Koch et al. "Evaluation of CNN-Based Single-Image Depth Estimation Methods". In: *Proc. Workshop of Eur. Conf. Comp. Vis.* 2018, pp. 331–348.

[66] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. "Semi-supervised deep learning for monocular depth map prediction". In: *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE. 2017, pp. 2215–2223.

[67] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. "Pulling things out of perspective". In: *IEEE Conf. Comput. Vis. Pattern Recog.*

[68] Iro Laina et al. "Deeper depth prediction with fully convolutional residual networks". In: *Proc. Int. Conf. 3D Vision (3DV)*. IEEE. 2016, pp. 239–248.

[69] Manuel Lang et al. "Nonlinear disparity mapping for stereoscopic 3D". In: *ACM Trans. Graph.* 29.4 (2010), pp. 1–10.

[70] Katrin Lasinger et al. "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer". In: abs/1907.01341 (2019).

[71] Ang Li et al. "A Multi-Scale Guided Cascade Hourglass Network for Depth Completion". In: *IEEE Wint. Conf. on Appli. of Comput. Vis.* 2020, pp. 32–40.

[72] Bo Li et al. "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2015, pp. 1119–1127.

[73] Hanhan Li et al. "Unsupervised Monocular Depth Learning in Dynamic Scenes". In: *arXiv: Comp. Res. Repository* (2020).

[74] Jun Li, Reinhard Klein, and Angela Yao. "A two-streamed network for estimating fine-scaled depth maps from single rgb images". In: *Int. Conf. Comput. Vis.* 2017, pp. 22–29.

[75] Ruibo Li et al. "Deep attention-based classification network for robust depth prediction". In: *Proc. Asian Conf. Comp. Vis.* 2018.

[76] Yangyan Li et al. "Pointcnn: Convolution on x-transformed points". In: *Adv. Neural Inform. Process. Syst.* 31 (2018), pp. 820–830.

[77] Zhengqi Li and Noah Snavely. "Megadepth: Learning single-view depth prediction from internet photos". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 2041–2050.

[78] Zhengqi Li et al. "Learning the Depths of Moving People by Watching Frozen People". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 4521–4530.

[79] Beyang Liu, Stephen Gould, and Daphne Koller. "Single image depth estimation from predicted semantic labels". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2010.

[80] Fayao Liu, Chunhua Shen, and Guosheng Lin. "Deep convolutional neural fields for depth estimation from a single image". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2015, pp. 5162–5170.

[81] Fayao Liu et al. "Learning depth from single monocular images using deep convolutional neural fields". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 38.10 (2015), pp. 2024–2039.

[82] Fayao Liu et al. "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2016).

[83] Fayao Liu et al. "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 38.10 (2016), pp. 2024–2039.

[84] Miaomiao Liu, Mathieu Salzmann, and Xuming He. "Discrete-continuous depth estimation from a single image". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2014, pp. 716–723.

[85] Zhijian Liu et al. "Point-Voxel CNN for Efficient 3D Deep Learning". In: *Adv. Neural Inform. Process. Syst.* 2019.

[86] Zhijian Liu et al. "Point-Voxel CNN for Efficient 3D Deep Learning". In: *Adv. Neural Inform. Process. Syst.* 32 (2019), pp. 965–975.

[87] Zhuang Liu et al. "Learning efficient convolutional networks through network slimming". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2736–2744.

[88] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *Int. J. Comput. Vis.* 60.2 (2004), pp. 91–110.

[89] Fangchang Ma and Sertac Karaman. "Sparse-to-dense: Depth prediction from sparse depth samples and a single image". In: *Proc. Int. Conf. on Robotics and Automation*. IEEE. 2018, pp. 4796–4803.

[90] Kiyoshi Matsuo and Yoshimitsu Aoki. "Depth image enhancement using local tangent plane approximations". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2015, pp. 3574–3583.

[91] Daniel Maturana and Sebastian Scherer. "Voxnet: A 3d convolutional neural network for real-time object recognition". In: *Proc. IEEE Int. Conf. Intell. Robots & Syst.* IEEE. 2015, pp. 922–928.

[92] Raúl Mur-Artal and Juan D. Tardós. "ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras". In: *IEEE Trans. Robot.* 33.5 (2017), pp. 1255–1262. DOI: 10.1109/TRO.2017.2705103.

[93] Vladimir Nekrasov et al. "Real-Time Joint Semantic Segmentation and Depth Estimation Using Asymmetric Annotations". In: *arXiv: Comp. Res. Repository.* Vol. abs/1809.04766. 2018.

[94] Richard A Newcombe et al. "Kinectfusion: Real-time dense surface mapping and tracking". In: *IEEE International Symposium on Mixed and Augmented Reality.* IEEE. 2011, pp. 127–136.

[95] Simon Niklaus et al. "3D Ken Burns Effect from a Single Image". In: *ACM Trans. Graph.* 38.6 (2019), 184:1–184:15.

[96] Jinsun Park et al. "Non-Local Spatial Propagation Network for Depth Completion". In: *Eur. Conf. Comput. Vis.* European Conference on Computer Vision. 2020.

[97] Emmanuel Prados and Olivier Faugeras. "Shape from shading: a well-posed problem?" In: *IEEE Conf. Comput. Vis. Pattern Recog.* Vol. 2. IEEE. 2005, pp. 870–877.

[98] Charles R Qi et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2017, pp. 652–660.

[99] Charles R Qi et al. "Volumetric and multi-view cnns for object classification on 3d data". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2016, pp. 5648–5656.

[100] Charles Ruizhongtai Qi et al. "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space". In: *Adv. Neural Inform. Process. Syst.* 30 (2017).

[101] Xiaojuan Qi et al. "GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 283–291.

[102] Jiaxiong Qiu et al. "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 3313–3322.

[103] Raul de Queiroz Mendes et al. "On deep learning techniques to boost monocular depth estimation for autonomous navigation". In: *Robotics and Autonomous Systems* 136 (2021), p. 103701.

[104] René Ranftl et al. “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2020).

[105] Anurag Ranjan et al. “Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 12240–12249.

[106] Edward Rosten and Tom Drummond. “Machine learning for high-speed corner detection”. In: *Eur. Conf. Comput. Vis.* Springer. 2006, pp. 430–443.

[107] Anirban Roy and Sinisa Todorovic. “Monocular depth estimation using neural regression forest”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2016, pp. 5506–5514.

[108] Ethan Rublee et al. “ORB: An efficient alternative to SIFT or SURF”. In: *Int. Conf. Comput. Vis.* IEEE. 2011, pp. 2564–2571.

[109] Radu Bogdan Rusu et al. “Aligning point cloud views using persistent feature histograms”. In: *Int. Conf. on Intell. Robots and Sys.* IEEE. 2008, pp. 3384–3391.

[110] Shunsuke Saito et al. “PiFu: Pixel-aligned implicit function for high-resolution clothed human digitization”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 2304–2314.

[111] Shunsuke Saito et al. “PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020, pp. 84–93.

[112] Mark Sandler et al. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 4510–4520.

[113] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. “Learning depth from single monocular images”. In: *Adv. Neural Inform. Process. Syst.* 2006, pp. 1161–1168.

[114] Ashutosh Saxena, Min Sun, and Andrew Y Ng. “Make3d: Learning 3d scene structure from a single still image”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 31.5 (2008), pp. 824–840.

[115] Ashutosh Saxena, Min Sun, and Andrew Y Ng. “Make3d: Learning 3d scene structure from a single still image”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 31.5 (2009), pp. 824–840.

[116] Johannes L Schönberger et al. “Pixelwise view selection for unstructured multi-view stereo”. In: *Eur. Conf. Comput. Vis.* Springer. 2016, pp. 501–518.

[117] Johannes Lutz Schönberger et al. “Pixelwise View Selection for Unstructured Multi-View Stereo”. In: *Eur. Conf. Comput. Vis.* 2016.

[118] Thomas Schops et al. "A multi-view stereo benchmark with high-resolution images and multi-camera videos". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2017, pp. 3260–3269.

[119] Daniel Seichter et al. "Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis". In: *arXiv: Comp. Res. Repository* (2020).

[120] Dmitry Senushkin, Ilia Belikov, and Anton Konushin. "Decoder Modulation for Indoor Depth Completion". In: *arXiv: Comp. Res. Repository* (2020), p. 2005.08607.

[121] Shaoshuai Shi et al. "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020, pp. 10529–10538.

[122] Meng-Li Shih et al. "3D Photography using Context-aware Layered Depth Inpainting". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020.

[123] Chang Shu et al. "Feature-metric loss for self-supervised learning of depth and egomotion". In: *Eur. Conf. Comput. Vis.* 2020, pp. 572–588.

[124] Nathan Silberman et al. "Indoor segmentation and support inference from rgbd images". In: *Eur. Conf. Comput. Vis.* 2012, pp. 746–760.

[125] Dalwinder Singh and Birmohan Singh. "Investigating the impact of data normalization on classification performance". In: *Applied Soft Computing* (2019), p. 105524.

[126] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. "Sun rgb-d: A rgb-d scene understanding benchmark suite". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2015, pp. 567–576.

[127] Andrew Spek, Thanuja Dharmasiri, and Tom Drummond. "CReaM: Condensed Real-time Models for Depth Prediction using Convolutional Neural Networks". In: *Int. Conf. on Intell. Robots and Sys.* IEEE. 2018, pp. 540–547.

[128] J. Sturm et al. "A Benchmark for the Evaluation of RGB-D SLAM Systems". In: *Proc. IEEE Int. Conf. Intell. Robots & Syst.* 2012.

[129] Jürgen Sturm et al. "A benchmark for the evaluation of RGB-D SLAM systems". In: *Proc. IEEE Int. Conf. Intell. Robots & Syst.* IEEE. 2012, pp. 573–580.

[130] Josh Tobin et al. "Domain randomization and generative models for robotic grasping". In: *Proc. IEEE Int. Conf. Intell. Robots & Syst.* IEEE. 2018, pp. 3482–3489.

[131] Josh Tobin et al. "Domain randomization for transferring deep neural networks from simulation to the real world". In: *Proc. IEEE Int. Conf. Intell. Robots & Syst.* IEEE. 2017, pp. 23–30.

[132] Jonas Uhrig et al. "Sparsity Invariant CNNs". In: *Int. Conf. 3D. Vis.*

[133] Benjamin Ummenhofer et al. "Demon: Depth and motion network for learning monocular stereo". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2017, pp. 5038–5047.

[134] Wouter Van Gansbeke et al. “Sparse and Noisy LiDAR Completion with RGB Guidance and Uncertainty”. In: *Proc. IEEE Int. Conf. Mach. Vis. Appli.* IEEE. 2019, pp. 1–6.

[135] Igor Vasiljevic et al. “DIODE: A Dense Indoor and Outdoor DEpth Dataset”. In: *arXiv: Comp. Res. Repository* (2019), p. 1908.00463.

[136] Igor Vasiljevic et al. “DIODE: A Dense Indoor and Outdoor DEpth Dataset”. In: abs/1908.00463 (2019).

[137] Chaoyang Wang et al. “Web stereo video supervision for depth prediction from dynamic scenes”. In: *Int. Conf. 3D. Vis.* IEEE. 2019, pp. 348–357.

[138] Nanyang Wang et al. “Pixel2mesh: Generating 3d mesh models from single RGB images”. In: *Eur. Conf. Comput. Vis.* 2018, pp. 52–67.

[139] Peng Wang et al. “Towards unified depth and semantic prediction from a single image”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2015, pp. 2800–2809.

[140] Peng-Shuai Wang et al. “O-cnn: Octree-based convolutional neural networks for 3d shape analysis”. In: *ACM Trans. Graph.* 36.4 (2017), pp. 1–11.

[141] Wenshan Wang et al. “Tartanair: A dataset to push the limits of visual slam”. In: *arXiv: Comp. Res. Repository* (2020).

[142] Xiaolong Wang, David Fouhey, and Abhinav Gupta. “Designing deep networks for surface normal estimation”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2015, pp. 539–547.

[143] Yan Wang et al. “Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 8445–8453.

[144] Yue Wang et al. “Dynamic graph cnn for learning on point clouds”. In: *ACM Trans. Graph.* 38.5 (2019), pp. 1–12.

[145] Zongji Wang and Feng Lu. “VoxSegNet: Volumetric CNNs for semantic part segmentation of 3D shapes”. In: *IEEE transactions on visualization and computer graphics* 26.9 (2019), pp. 2919–2930.

[146] Yin Wei et al. “Enforcing geometric constraints of virtual normal for depth prediction”. In: *Proc. IEEE Int. Conf. Comp. Vis.* (2019).

[147] Daphna Weinshall, Gad Cohen, and Dan Amir. “Curriculum Learning by Transfer Learning: Theory and Experiments with Deep Networks”. In: *Proc. Int. Conf. Mach. Learn.* 2018, pp. 5235–5243.

[148] Jiajun Wu et al. “Learning shape priors for single-view 3d completion and reconstruction”. In: *Eur. Conf. Comput. Vis.* 2018, pp. 646–662.

[149] Ke Xian et al. “Monocular relative depth perception with web stereo data supervision”. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 311–320.

[150] Ke Xian et al. "Structure-Guided Ranking Loss for Single Image Depth Prediction". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2020, pp. 611–620.

[151] Jianxiong Xiao et al. "Recognizing scene viewpoint using panoramic place representation". In: *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE. 2012, pp. 2695–2702.

[152] Saining Xie et al. "Aggregated residual transformations for deep neural networks". In: *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE. 2017, pp. 5987–5995.

[153] Yan Xu et al. "Depth completion from sparse lidar data with depth-normal constraints". In: *Int. Conf. Comput. Vis.* 2019, pp. 2811–2820.

[154] Yifan Xu et al. "Spidercnn: Deep learning on point sets with parameterized convolutional filters". In: *Eur. Conf. Comput. Vis.* 2018, pp. 87–102.

[155] Yanchao Yang, Alex Wong, and Stefano Soatto. "Dense depth posterior (ddp) from single image and sparse range". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 3353–3362.

[156] Zhenheng Yang et al. "LEGO: Learning Edge with Geometry all at Once by Watching Videos". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 225–234.

[157] Zhenheng Yang et al. "Unsupervised Learning of Geometry From Videos With Edge-Aware Depth-Normal Consistency". In: *AAAI Conf. Artificial Intelligence*. 2018.

[158] Yao Yao et al. "BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks". In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2020).

[159] Wei Yin, Yifan Liu, and Chunhua Shen. "Virtual Normal: Enforcing Geometric Constraints for Accurate and Robust Depth Prediction". In: *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).

[160] Wei Yin et al. "DiverseDepth: Affine-invariant Depth Prediction Using Diverse Data". In: *arXiv: Comp. Res. Repository* (2020), p. 2002.00569.

[161] Wei Yin et al. "Enforcing geometric constraints of virtual normal for depth prediction". In: *Int. Conf. Comput. Vis.* 2019.

[162] Wei Yin et al. "Learning to Recover 3D Scene Shape from a Single Image". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2021.

[163] Zhichao Yin and Jianping Shi. "Geonet: Unsupervised learning of dense depth, optical flow and camera pose". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 1983–1992.

[164] Sergey Zakharov, Wadim Kehl, and Slobodan Ilic. "Deceptionnet: Network-driven domain randomization". In: *Int. Conf. Comput. Vis.* 2019, pp. 532–541.

[165] Amir R Zamir et al. "Taskonomy: Disentangling task transfer learning". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2018, pp. 3712–3722.

[166] Bernhard Zeisl, Marc Pollefeys, et al. "Discriminatively trained dense surface normal estimation". In: *Eur. Conf. Comput. Vis.* Springer. 2014, pp. 468–484.

[167] Feihu Zhang et al. "GA-Net: Guided Aggregation Net for End-to-end Stereo Matching". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2019, pp. 185–194.

[168] Yinda Zhang and Thomas Funkhouser. "Deep Depth Completion of a Single RGB-D Image". In: *IEEE Conf. Comput. Vis. Pattern Recog.* (2018).

[169] Tinghui Zhou et al. "Unsupervised learning of depth and ego-motion from video". In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2017, pp. 1851–1858.

[170] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Int. Conf. Comput. Vis.* 2017, pp. 2223–2232.