Metric Monocular Reconstruction through Ordinal Depth

by

Mahesh Kumar Krishna Reddy

M.Sc., University of Manitoba, 2020
B.E., Visvesvaraya Technological University, 2016

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Fall 2022

Declaration of Committee

Name:	**Mahesh Kumar Krishna Reddy**
Degree:	**Master of Science**
Thesis title:	**Metric Monocular Reconstruction through Ordinal Depth**
Committee:	**Chair:** Tianzheng Wang Assistant Professor, Computing Science
	Yağız Aksoy Supervisor Assistant Professor, Computing Science
	Ke Li Committee Member Assistant Professor, Computing Science
	Hao (Richard) Zhang Examiner Professor, Computing Science

Abstract

Training a single network for high resolution and geometrically consistent monocular depth estimation is challenging due to varying scene complexities in the real world. To address this, we present a dual depth estimation setup to decompose the estimations into ordinal and metric depth. The goal of ordinal depth estimation is to leverage novel ordinal losses with relaxed geometric constraints to model local and global ordinal relations for capturing better high-frequency depth details and scene structure. However, ordinal depth inherently lacks geometric structure, and to resolve this, we introduce a metric depth estimation method to enforce geometric constraints on the prior ordinal depth estimations. The estimated scale-invariant metric depth achieves high resolution and is geometrically consistent in generating meaningful 3D point cloud representation for scene reconstruction. We demonstrate the effectiveness of our ordinal and metric networks by performing zero-shot and in-the-wild depth evaluations with state-of-the-art depth estimation networks.

Keywords: monocular depth estimation, point clouds, mesh generation

Dedication

To my parents and sister for their unconditional love and endless support!

Acknowledgements

First and foremost, I wish to convey my appreciation to my advisor Professor Yağız Aksoy, for his unrelenting encouragement, mentoring, and continuous motivation throughout my journey at Simon Fraser University, which was vital to the completion of my thesis. I want to thank Prof. Tianzheng Wang, Prof. Ke Li, and Prof. Hao (Richard) Zhang for their time and participation in the thesis defence. Furthermore, I thank Seyed Mahdi Hosseini Miangoleh and Long Mai for being fantastic collaborators on this project. I also want to thank Chris Careaga and Obumneme Stanley Dukor for their help and feedback on this project. Finally, I thank other members of the Computational Photography Lab at SFU for all the lively discussions and constructive feedback.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Figure 1.1: We propose a two-step depth estimation framework to generate high-resolution geometrically consistent metric monocular depth from a single image. Furthermore, the depth can be projected to dense and detailed 3D point clouds to recover the surface mesh for a diverse set of complex real-world scenes.

Depth estimation is an essential mid-level vision task with applications in computational photography, image editing, and 3D reconstruction. In the presence of two or more images of the same scene from different camera viewpoints, estimating depth is a simplified problem by employing techniques from epipolar geometry between any pair of images. However, inferring

Figure 1.2: Our high-resolution metric depth estimations can be projected to a geometric coherent 3D point cloud to generate detailed surface meshes for complex scenes and single objects.

strict scene geometry from a single or *monocular depth estimation* (MDE) is a challenging problem in the absence of multi-view information. Therefore, depth estimation can leverage different depth cues such as occlusions, relative sizes of objects, or perspective to reason about the geometric structure of the scene. Monocular depth enables structure-aware editing of still photographs (3D Photography [41], 3D Ken Burns effect [33] and Synthetic depth-of-field [48]), and 3D scene reconstruction. Despite the immediate photographic appeal of going into the 3D space through a single photograph, structure-aware computational photography and 3D rendering pipelines have yet to see a wide adoption in the artistic community. One significant limiting factor is that realistic 3D photography and rendering applications need high-resolution depth with high-frequency details and well-defined geometric structures for effective results.

With the rise in the usage of data-driven approaches in the computer vision community, there is wide adoption of convolutional neural networks (CNNs) for depth estimation [3, 12, 24, 25, 36, 53, 54, 56, 57, 59]. However, due to the limited capacity in CNNs, an inverse relationship exists between the structural consistency and the sharp depth discontinuities in depth estimation [32]. Most existing approaches address either scene structure [12, 35, 36, 59] or sharp depth boundaries [3, 53, 54] due to this limitation.

To estimate depth with coherent scene structure, recent methods [35, 36, 56, 57] collate data from multiple datasets (e.g., stereo pair datasets) containing diverse labeled depth representations. The images from different datasets are captured by camera setups with

Figure 1.3: We can project our high-resolution metric depth estimations to geometric and coherent point clouds to generate detailed surface meshes from single photographs with varying scene complexities.

varying camera intrinsic parameters. Therefore, to account for the unknown camera baseline between the two cameras in the stereo setup and the camera focal length across different images, these methods employ a scale and shift invariant (SSI) loss to estimate depth up to an affine transformation. Despite effective performance on different datasets, the SSI loss prioritizes scene structure over capturing sharp details.

Alternatively, other MDE approaches estimate sharp depth boundaries through *ordinal depth* at the cost of geometric structure. The motivation for ordinal depth is that it is much easier for a network to predict relative depth relationships between pixels, i.e., asking the network to ranking pixels based on their proximity to the camera rather than estimating their true metric depth distance [60]. The ordinal depth estimation networks depend on point pairs sampling to determine the depth ordinality through a ranking loss. The sampling can be either random [3, 53] or guided by the RGB input image gradients [54]. This results in better high-frequency depth details while lacking consistency in the global scene structure.

In addition to estimating depth for scene structure using SSI loss or sharp details through ranking loss, there is another stream of prior approaches [12, 24, 25, 56] that estimate *metric depth* by training on metric ground-truth depth from Kinect [12] or LiDaR [16]. The networks estimate depth up to a scale by employing scale-invariant loss formulations that enforce the depth estimation to have a similar ratio as the ground truth metric depth between any corresponding pair of pixels. However, a limitation of metric networks is that their training data distribution is either small [39, 12] or less diverse [16] due to complexity and

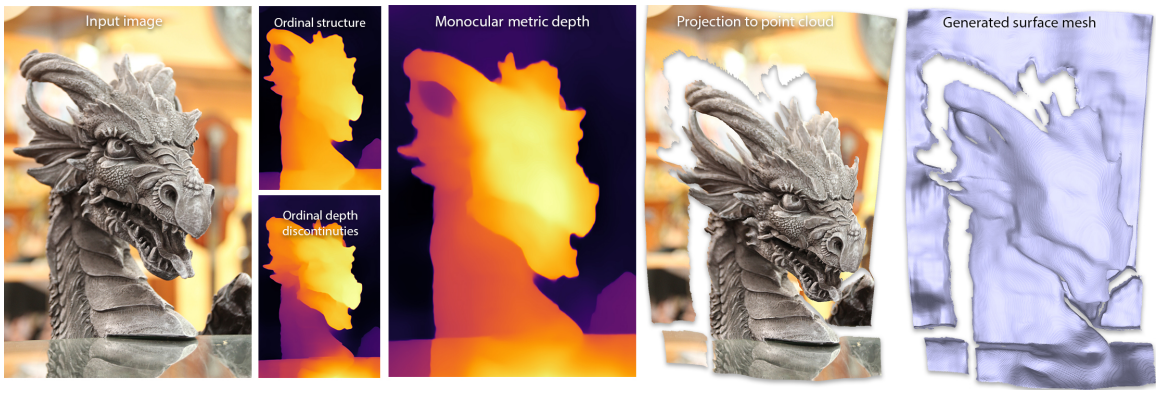

Figure 1.4: We consider an ordinal depth estimation network to capture local and global ordinal depth relations to infer scene structure and high-frequency depth details. The ordinal estimates are inputs to the metric depth estimation network to estimate geometrically consistent depth by enforcing geometric constraints. The estimated metric depth helps generate a dense 3D point cloud for scene reconstruction.

scalability issues in data capturing systems. Further, the data captured by these setups are either noisy (Kinect) due to the difficulties in matching pixel correspondences for complex objects and materials or sparse (LiDaR). Due to these data limitations and the challenging nature of the problem, the depth estimation is low-resolution and lacks high-frequency details, but it has a consistent geometric structure.

Modern MDE networks [35, 36, 54] struggle to generate high-resolution depth maps, either due to network capacity or receptive field size [32]. To overcome the limitation of earlier approaches [36, 54, 59] in estimating high-resolution depth estimation, Miangoleh *et al.* [32] propose a post-processing step to boost the depth estimation from off-the-shelf MDE networks [36, 54, 59]. The boosting framework performs a low-level merging of low-resolution and high-resolution depth estimations using a CNN to achieve high-resolution depth estimation. However, the boosting framework leverages depth estimations from pre-trained MDE networks [36, 54, 59] that results in estimating structurally inconsistent high-resolution depth. Therefore, training a single network to capture scene geometry and high-resolution details is challenging.

This thesis proposes a dual depth estimation framework to generate high-resolution and geometrically consistent monocular depth by decomposing the depth estimation into two steps: ordinal and metric depth. Based on our observations in Chapter 3, the SSI loss leads to consistent scene structure, while the ranking loss dominates in estimating sharp and accurate depth boundaries. Therefore, we improve the training setup of both SSI and ranking loss by combining them to estimate ordinal depth that shows both scene structure and sharp depth edges. This combined objective explicitly enables depth ordinality constraints to model local (details) and global (structure) depth relations. To further address

the lack of geometric structure consistency in the ordinal depth estimations, we leverage the ordinal estimations as inputs to our metric depth network. The ordinal inputs provide local and global depth relations. Our metric network estimates high-resolution metric depth with consistent geometric structure by employing sparse and dense geometry-aware losses. The high-resolution depth from our two-step approach can generate consistent 3D representations for in-the-wild complex scenes from a single image. Furthermore, we can produce dense and coherent point clouds by projecting the high-resolution metric depth estimation to 3D space. The dense point cloud further enables recovering a consistent surface mesh for diverse and complex real-world environments with intricate details, as shown in Figures 1.2 and 1.3. Finally, we provide an overview of the overall framework in Figure 1.4.

Chapter 2

Related Work

There is a proliferation in the monocular depth estimation (MDE) approaches due to its direct usage in numerous computer vision and computational photography applications [33, 41]. The traditional MDE approaches consider a prior on the scene structure for flat floors and straight vertical walls ("floor-wall" model) [8, 17], followed by generalizable approaches based on image superpixels to model the 3D structure of a scene [39] without any additional assumptions on the scene structure. More recently, with a burst of data-driven approaches, the convolutional neural network (CNN) based methods estimate dense depth from single images using labeled metric depth data [11, 12, 14, 23, 27, 38].

2.1 Monocular Ordinal Depth Estimation

Collecting large-scale metric datasets for depth estimation can be expensive; hence several CNN-based approaches solve depth estimation using synthetic datasets [31], sparse human labeled depth annotations [3], structure-from-motion (SfM) and multi-view stereo (MVS) on images collected from web [25], SfM and MVS on videos of people performing mannequin challenge [24], SfM on internet videos with an assessment network to consider only high-quality reconstructions, SfM on 3D movies [36, 49], ordinal annotations on stereo image pairs [53, 54], and steerable datasets from 3D scans of environments [10].

The stereo-based data collected from the internet have an unknown scale and shift due to the unknown camera baseline and focal length with respect to metric depth. Prior approaches use two main loss formulations to estimate depth: scale and shift invariant loss (SSI) [36, 49, 58], and ranking loss [3, 53, 54]. Wang *et al.* [49] account for an unknown shift in the disparity from web stereo videos by considering the gradient difference between the ground truth and rescaled disparity predictions at multiple scales. Ranftl *et al.* [36] propose a loss invariant to scale and shift to train on data collated from multiple datasets. Although the SSI loss [36] can estimate robust depth that captures global scene structure, it lacks high-resolution details compared to the networks using ranking loss [54]. Another type of supervision uses ordinal annotations to train MDE networks with the rank-

ing loss [3, 53, 54, 60]. A key component of ranking loss involves sampling image point pairs to apply the loss, and the prior approaches use different sampling strategies. Some sampling strategies are random [3], balanced random [53], and structure-guided [54]. In particular, Xian *et al.* [54] shows that sampling point pairs based on the image structure benefit the training in estimating better high-resolution details. However, a significant limitation of MDE approaches using ranking loss is the lack of coherent scene structure compared to networks using SSI loss [35, 36].

The key benefits of SSI and ranking losses are that they capture global scene structure and local high-frequency details, respectively. Inspired by their advantages, we define a depth space that obeys the **ordinal depth distribution** to combine the losses to complement each other in training an ordinal depth estimation network to estimate both scene structure and high-frequency details. Also, we propose a variant of the ranking loss, termed as **relaxed ranking loss** with a novel **triplet sampling** of points to create good harmony with the SSI loss. Through qualitative and quantitative evaluations, we demonstrate that our combined loss formulation can generate better details than SSI loss and better structure than ranking loss.

2.2 Monocular Metric Depth Estimation

The term "metric" depth indicates depth up to an unknown scale. Estimating metric depth from images is more challenging than the ordinal formulation [60] as the network needs to reason about the geometric structure of the scene. Moreover, the depth ratios greatly vary between pixels in images across different environments. For instance, the depth ratio between points a and b defined as $r = a/b$ indicates that point a is at a distance $r \cdot b$ further or closer than point b to the camera. The ratio between a pair of points denoting the foreground and background object in an indoor environment can be drastically smaller than the foreground and background points in an outdoor environment due to the difference in the metric depth distribution. Therefore, estimating these complex relations from a single input image is challenging. The early metric MDE approach by Eigen *et al.* [12] introduces a scale-invariant loss in the log-depth space to account for the unknown depth scale to estimate metric depth by matching the depth ratio of pixels in the depth estimation with the corresponding pixels in the ground truth depth. Li and Snavely [25] consider large-scale web data and derive ground truth using the SfM technique to train a metric MDE using scale-invariant loss formulation. Unlike ordinal depth estimation, the metric MDE approaches can derive consistent surface normals from metric depth estimations due to the geometric constraints by the scale-invariant loss. Chen *et al.* [4] propose to use surface normals to improve metric depth by enforcing additional geometric constraints. Yin *et al.* [56] propose a normal loss based on the virtual plane constructed with points sampled in 3D point clouds to introduce long-range geometric constraints to supervise depth estimation. Yin *et al.* [59]

introduce a framework to recover 3D shape from monocular images by estimating metric depth and an additional point cloud-based network to recover the unknown scale and shift to fix the scale and shift of the estimated depth. Further, to train the depth network, they use the pseudo-ordinal and sparse surface normals loss on points sampled based on image gradients [54].

We take inspiration from the limitations of the prior metric MDE approaches that struggle at high-resolution depth estimation from a single input image. Therefore, we propose a **metric depth network** that combines the input image and a pair of ordinal depth estimations at different resolutions to estimate high-resolution metric depth with geometrical consistency. Specifically, we use the low and high-resolution depth estimations from our ordinal depth network as inputs along with original RGB images. The low-resolution depth estimates provide a rough scene structure, while the high-resolution estimates provide the context of sharp depth discontinuities. The ordinal depth inputs enable us to use different geometric loss formulations to apply depth and surface normal-based geometric constraints to train the metric depth network. We employ a **sparse ratio loss** along with a dense scale-invariant loss to preserve the metric depth relations in the estimation. Additionally, we add **dense surface normal losses** to enforce local geometric and smoothness constraints on the surface normals derived from the estimated metric depth. The high-resolution nature of our geometrically consistent depth maps enables the projection of coherent, dense 3D point clouds. Furthermore, we employ a state-of-the-art point cloud to mesh generation method by Chen *et al.* [5] as a black box for 3D scene reconstruction for complex in-the-wild monocular images with varying intricate details.

2.3 High-Resolution Monocular Depth Estimation

The prior metric MDE methods do not naturally generate high-resolution depth due to network capacity or scene complexity. Therefore, most approaches utilize images with small input resolutions to train the networks to estimate depth. However, several downstream applications on depth such as shallow depth-of-field [48, 50], 3D Ken Burns [33], 3D Photography [21, 41], or 3D rendering would benefit from high-resolution and detailed depth estimations. Niklaus *et al.* [33] propose guided up-sampling of low-resolution depth based on the high-resolution image. Lyu *et al.* [28] propose introducing redesigned skip-connections with a feature fusion approach. Miangoleh *et al.* [32] propose a more effective technique to perform low-level gradient transfer using a trained CNN between depth estimations at different resolutions from pre-trained depth networks. However, the pre-trained MDE networks [36, 54] do not reason about the geometric structure, producing contorted 3D representations of the scene.

Our metric network is inspired by the work by Miangoleh *et al.* [32] for high-resolution depth estimation. Instead of implementing a low-level gradient transfer method [34], our

network performs high-resolution metric depth estimation with high-frequency details. Our metric network uses the ordinal depth estimations as constraints along with the original input image, similar to the CRF-based optimization setup by Zoran *et al.* [60] to estimate metric depth from ordinal constraints.

Chapter 3

Preliminaries

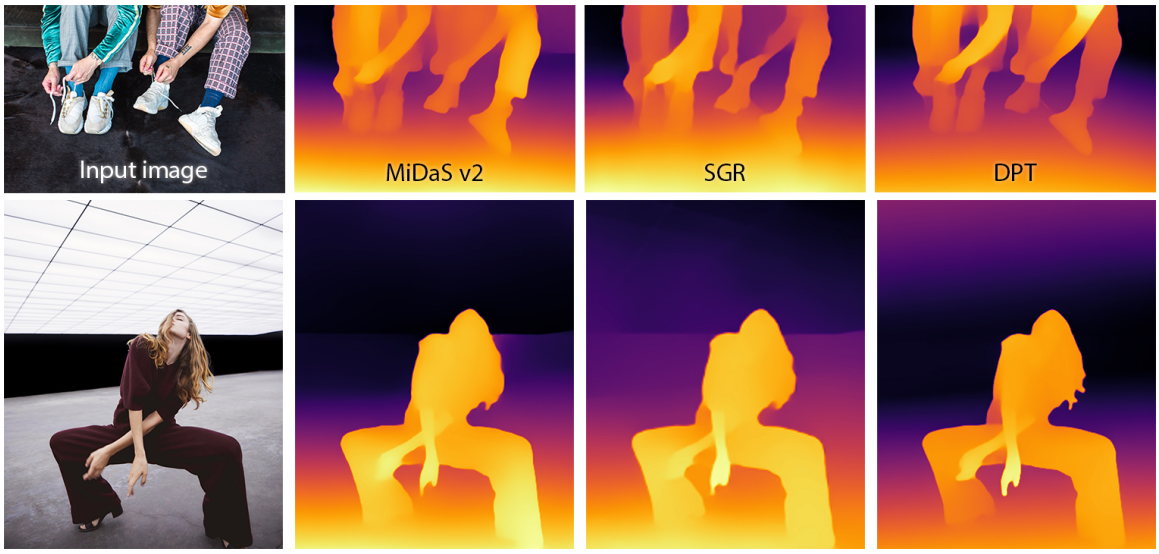

Figure 3.1: Depth estimations from recent state-of-the-art MDE methods. Ranftl *et al.* [36] introduce MiDaS v2 network using SSI loss for depth estimation. Xian *et al.* [54] propose structure-guided ranking (SGR) based on the ranking loss formulate for depth estimation. Ranftl *et al.* [35] uses transformers-based architecture for DPT using SSI loss. The SSI-based approaches are effective in capturing scene structure, while the ranking loss-based approaches capture better depth details.

Stereopsis is a significant contributor to depth perception of the environment through a binocular vision in humans [13]. However, in the absence of binocular vision, we can still perceive depth by leveraging depth cues such as perspective, shadows, or gradients to perceive depth through monocular vision. We can use convolutional neural networks (CNNs) for monocular depth estimation by training on labeled depth datasets. Monocular depth estimation (MDE) networks use different depth cues in the images to estimate depth. Based on the training objective, we can categorize the previous work in MDE into ordinal, pseudo-ordinal or metric networks. An ordinal depth network aims to estimate depth by reasoning

about the relative depth relations between pixels at the cost of geometric consistency. The ordinal depth setup relaxes the geometric constraints, which lets the network train from different datasets. The ordinal MDE majorly use ranking loss [3, 4, 53, 54] (§ 3.2) on sampled depth pixels. In addition, we can consider the scale and shift invariant loss (SSI) [35, 36, 57] as a dense pseudo-ordinal loss (§ 3.1).

The metric depth methods estimate consistent geometric depth that is a scale away from ground truth by employing scale-invariant loss [12] (§ 3.3). Estimating metric depth is a complex problem because the network must reason about the global depth ratio between pixels to preserve the scene geometry across diverse complex scenes with a wide range of depth values. Additional geometric supervision comes from surface normal-based angle loss [4] to align the surface orientations between the estimation and ground truth surface normals. A primary limitation of the metric depth is related to the quality of ground truth depth. For example, training on noisy or sparse metric depth ground truth can capture the geometric structure of the scene but not generate high-resolution details.

3.1 Pseudo-ordinal Scale and Scale Invariant Loss

Making MDE networks generalize to images in the wild requires training on large and diverse datasets. However, a key challenge is the limited number of sizeable metric depth datasets available to achieve it. Alternatively, prior approaches use large web-based stereo datasets to generate depth maps from the optical flow between images in the stereo pairs [18, 36, 49, 53, 54]. However, the web stereo datasets contain images from diverse cameras, introducing an unknown scale and shift ambiguity. The scale is due to the unknown focal length, which is the distance between the optical centre of the lens and the camera's image sensor. The unknown shift is due to the camera baseline, which is the distance between the two cameras in a stereo setup. To account for this ambiguity, recent work by Ranftl *et al.* [36] proposes a SSI loss to train the MDE networks. Before computing the loss, they align the depth estimation from the network with the ground truth depth using the least-squares criterion to recover the scale and shift parameters to account for the ambiguity. The scale and shift parameters $\mathbf{h} = (scale, shift)^T$ is given by:

$$\mathbf{h} = \Big(\sum_i \vec{\mathbf{d}}_i \vec{\mathbf{d}}_i^T \Big)^{-1} \Big(\sum_i \vec{\mathbf{d}}_i \mathbf{d}_i^* \Big), \quad \vec{\mathbf{d}}_i = (\mathbf{d}_i, 1)^T, \tag{3.1}$$

where $\mathbf{d}_i^*$ and $\mathbf{d}_i$ indicate the depth values for the pixel i in the ground truth and aligned depth estimation, respectively. We can define the SSI loss between the aligned depth estimation and the ground truth image as:

$$\mathcal{L}_{ssi}(\mathbf{d}, \mathbf{d}^*) = \frac{1}{2M} \sum_i (\vec{\mathbf{d}}_i^T \mathbf{h} - \mathbf{d}_i^*)^2, \tag{3.2}$$

where M is the number of pixels.

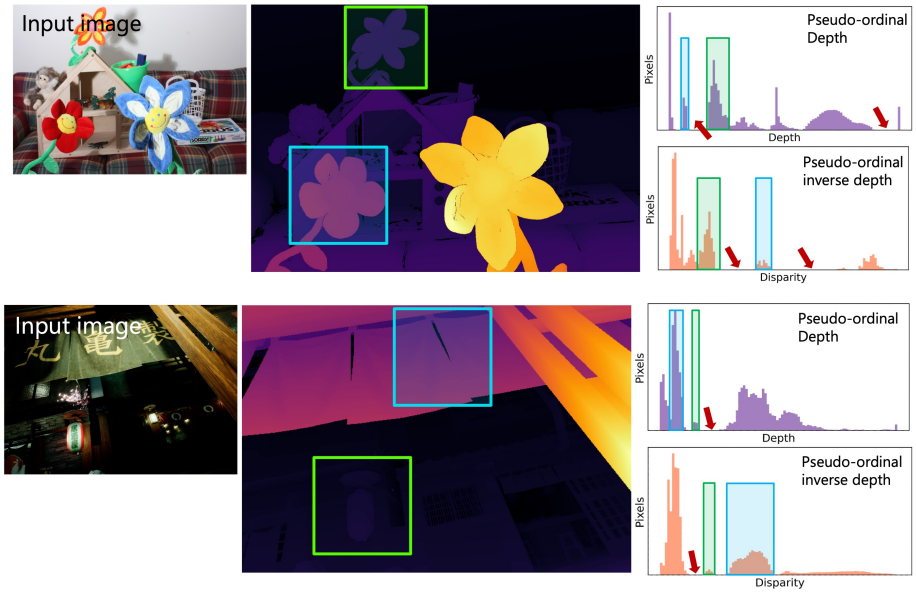

Figure 3.2: The scene complexity of images captured in the real world can have high variations or *gaps* in the depth distribution. This geometrical complexity limits the network from estimating high-resolution details while preserving the geometric structure. For example, for an indoor image (top row), the gaps shown by red arrows in both the depth and disparity (inverse depth) histograms indicate the depth difference between the flowers. On the other hand, the depth difference is more drastic in outdoor images (bottom row).

Pseudo-Ordinal Loss

We can consider the SSI loss as a *pseudo-ordinal loss* as it is neither completely ordinal nor metric. It is not ordinal as it does not explicitly solve for depth order like the ranking loss using ordinal constraints. In contrast, it does not enforce the depth ratio for corresponding pixels between depth estimation and ground truth depth to be the same as the metric loss. This is because the pseudo-ordinal depth is an affine transformation away from the metric depth.

In addition, as the pseudo-ordinal loss considers ground depth derived from web-stereo pairs, the generated ground truth is an affine transformation away from the true metric depth. Although the depth is not truly metric, as discussed earlier, it still captures some traits of metric depth distribution. A key observation is that most datasets cover a broad range of depth values in indoor or outdoor scenes. However, only a subset of the distinct depth values captures the essential geometric information. In Figure 3.2, the indoor scene (top row) shows the depth distribution with multiple clustered bins with some space (or *gaps*) between them. The gap in the histogram indicates a significant difference in the depth levels in the ground-truth maps. The gaps become more evident in outdoor scenes (bottom row), as the foreground and background objects are usually far apart. The gaps lead to significantly fewer information clusters with depth. Therefore, reasoning about these

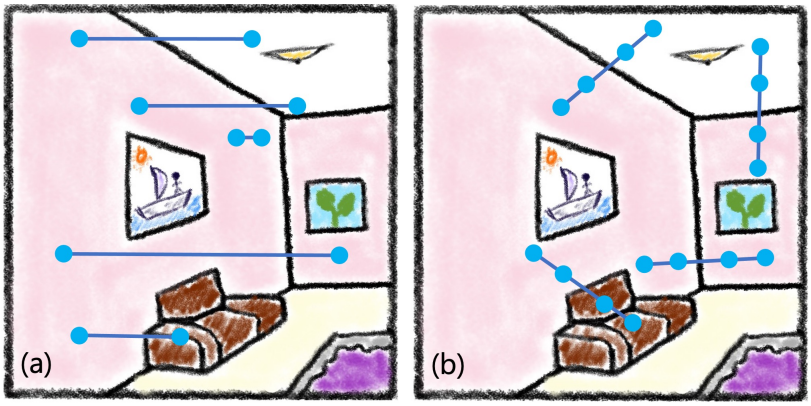

Figure 3.3: We present an illustration of different point pairs sampling techniques employed in monocular depth training pipelines for ranking loss. The random sampling [3] in (a) randomly picks pixel pairs along a horizontal line in the image. Xian *et al.* [54] propose a structure-guided sampling by leveraging the image gradients to sample along the edges in (b) and combined random pairs sampling and segmentation mask-based sampling. The previous sampling techniques leveraged only pixel pairs and generated imbalanced ordinal pairs.

varying depth distributions while preserving the metric characteristics is challenging for pseudo-ordinal loss based MDE networks and estimates monocular depth that lacks high-resolution details at the cost of scene structure.

3.2 Ordinal Ranking Loss

Unlike the pseudo-ordinal loss that applies a dense loss on the pixels, an alternate approach for MDE is applying the ranking loss on sparse pixels to estimate ordinal depth. The ordinal depth relaxes the geometric reasoning constraints on the network to estimate monocular depth by relative ranking of image pixels as described by Zoran *et al.* [60]. Most ordinal MDE approaches employ the ranking loss [3] that enforces ordinal constraints to rank a sparse set of pixel pairs.

For a given pair of points a_{gt} and b_{gt} in the ground truth depth map, their depth ordinality relation is given by:

$$\phi = \begin{cases} +1 & \text{if } \mathrm{a}_{gt} \;/\; b_{gt} > 1 + \delta \\ -1 & \text{if } \mathrm{b}_{gt} \;/\; a_{gt} > 1 + \delta \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

where δ is a depth tolerance threshold. The ranking loss [3] for the corresponding points a_{pred} and b_{pred} in the predicted depth maps based on the ground truth ordinal relation ϕ is given by:

$$\mathcal{L}_{rl}(a_{pred}, b_{pred}) = \begin{cases} \texttt{log}(1 + \texttt{exp}(-\phi \cdot (a_{pred} - b_{pred}))) & \text{if } \phi \neq 0 \\ (a_{pred} - b_{pred})^2 & \text{if } \phi = 0 \end{cases} \tag{3.4}$$

The goal of the above loss is to pull points a_{pred} and b_{pred} closer if their corresponding ground truth points a_{gt} and b_{gt} have an ordinal equality relation by optimizing the squared error. However, it pushes the points in the predicted depth far apart to have different depth values for inequality ground truth ordinal relations.

Pair sampling

The effectiveness of the ranking loss to estimate ordinal depth relies heavily on the sampling strategy. Prior MDE approaches use different sampling rules to enforce different training objectives. The sampling technique by Chen *et al.* [3] considers a combination of "unconstrained" and "symmetric" pairs that contribute almost equal point pairs. More specifically, they uniformly sample point pairs along a random horizontal line in the unconstrained setup but uniformly sample two symmetric points based on the centre of the random horizontal line. Xian *et al.* [53] propose another random sampling strategy through online random sampling of point pairs within each training mini-batch. They consider removing 25% the point pairs with unequal ordinal relations if their depth difference is large. This additional heuristic stabilizes the training by ignoring significant outlier errors and reducing the count of a large number of pairs with unequal ordinality. Furthermore, Xian *et al.* [54] propose a novel sampling strategy to estimate depth that captures high-frequency details. To achieve this, they use RGB image gradients as a proxy for the depth discontinuities. The sampling operation uses the gradients to guide the random sampling to choose four points for every edge point computed from the image gradients. To further augment the sampling, they utilize semantic maps of objects to sample inside and outside the semantic map. In addition, they randomly sample from the image to create long-range ordinal depth relations. However, a limitation of random-based sampling of *point pairs* in all three previous strategies is the selection of an imbalanced set of point pairs. Specifically, the number of equality point pairs is lesser than the inequality ordinal pairs.

While choosing point pairs, these sampling techniques randomly sample points and create pairs. However, the depth threshold (δ) to define the ordinal equality relation between a pair of points using the depth ratio is small. Since the points are randomly selected, it is highly unlikely that many points with a depth ratio less than the depth threshold are picked. Due to this reason, these techniques generate more point pairs with inequality relations than equality relations. We illustrate different prior sampling strategies (a and b) in Figure 3.3, where `blue` points indicate the sampled points and the line indicates the point pairs.

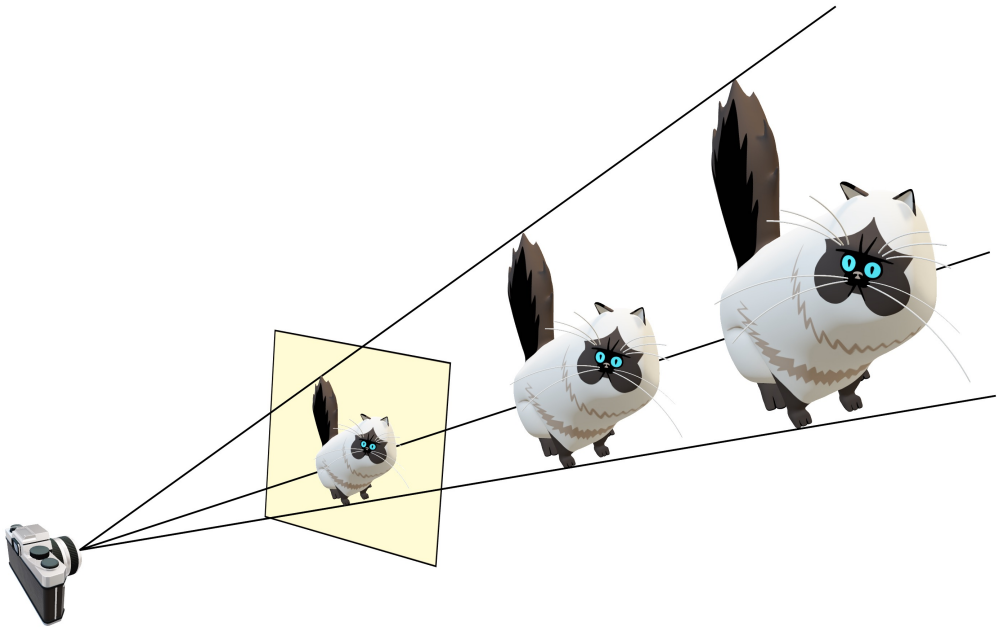

Figure 3.4: Illustration to demonstrate the issue of scale ambiguity in 3D to 2D perspective projection. Specifically, the ambiguity arises from the fact that we can project infinite 3D scenes with different scales to the exact 2D representation.

3.3 Metric Scale-Invariant Loss

Estimating consistent geometric depth requires capturing accurate ground truth metric datasets with devices like Kinect or LiDaR. However, due to the mechanics of 3D to 2D perspective projection, recovering the three-dimensional information from an input 2D input image through MDE is an ill-posed problem. Due to this ill-posed nature of the metric depth estimation problem, there are numerous geometrical solutions. More specifically, we can project numerous 3D scenes to the exact 2D representation as seen on the image plane in Figure 3.4. As seen in the figure, two 3D models of a cat with different scales can result in the same 2D image. To address this issue, most prior metric MDE approaches [12, 24, 25, 56] use a scale-invariant loss. Specifically, the goal of the loss is to match the depth ratio between pixels in the depth estimation to the ratio between the corresponding points in the ground-truth metric depth. Considering the depth ratio allows the loss to compute the error to optimize the metric MDE network while not accounting for the scale discrepancy between the estimation and ground truth. A vital benefit of the scale-invariant loss is preserving the geometric constraints by matching the ratios. The mathematical expression for the standard scale-invariant mean squared error [12] in the prior metric MDE approaches is given by:

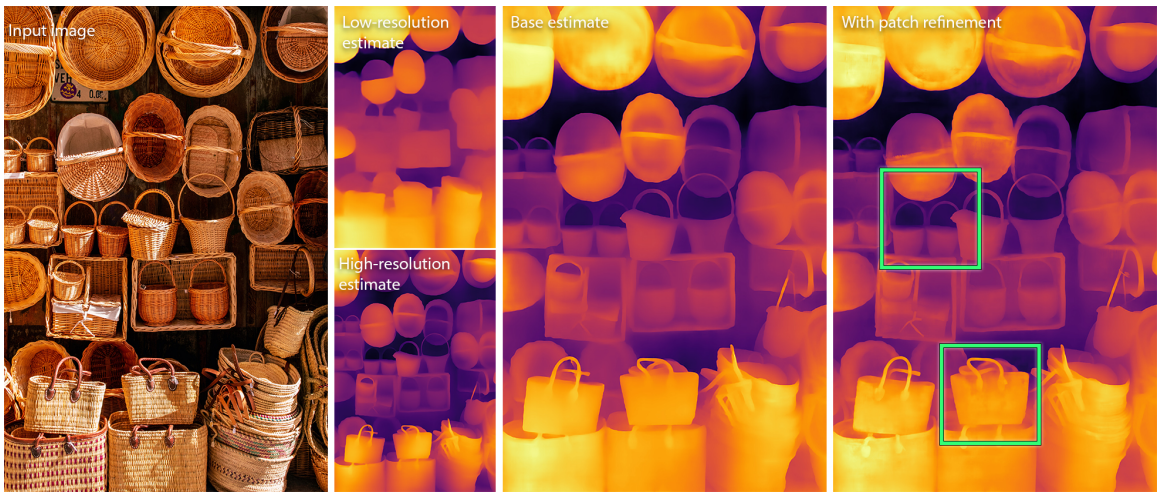

Figure 3.5: Overview of the depth estimation pipeline from boosting framework [32] using MiDaS depth network. For a given image, depth estimations at two different resolutions (receptive field size and R_{20}) capture scene structure and high-frequency details. Then, the merging network combines the two estimations to generate a high-resolution base estimate. To further improve the intricate details in the base estimate, a patch refinement approach is employed at the image patch level, as seen in the green squares.

$$\mathcal{L}_{si} = \frac{1}{n}\sum_{i=1}(r_i)^2 - \frac{1}{n^2}\left(\sum_{i=1} r_i\right)^2 \tag{3.5}$$

where $r_i = d_i - d_i^*$ and n is the total image pixels. The loss calculates the squared difference between two pixels in the prediction (d) and the same two points in the ground truth (d^*) averaged over all the points as described by Li *et al.* [24].

3.4 High-Resolution Monocular Depth Estimation

Depth from most prior MDE methods lacks high-resolution details essential to unleashing powerful downstream applications [33, 41]. This limitation stems from the trade-off between capturing scene structure and sharp high-frequency details. As a result, few MDE approaches estimate high-resolution depth. Niklaus *et al.* [33] resort to the depth refinement technique by using a high-resolution image to guide the depth upsampling through a neural network. Despite generating sharp object boundaries for large objects, it lacks capturing intricate high-frequency details for smaller objects. Recently, a depth boosting framework by Miangoleh *et al.* [32] generates high-resolution depth for images in the wild. The boosting depth is a plugin framework on the pre-trained MDE networks. Specifically, they demonstrate some critical behaviours of pre-trained MDE networks. First, they observe that the characteristics of the depth estimation vary by changing the resolution of the input

RGB image. For low-resolution input images similar to the training resolution of the MDE network, the depth estimations generate consistent scene structure but not high-frequency details. In contrast, gradually increasing the input resolution generates depth that shows an increase in high-frequency details while degrading the scene structure. They show that dual behaviour in pre-trained MDE networks is due to the network's limited capacity of CNNs and receptive field size. To leverage the unique characteristics of depth estimations at different resolutions, they introduce a *merging network* to generate high-resolution depth. The goal of the merging network is to fuse the consistent structure from the low-resolution depth with the high-frequency details from the high-resolution depth. The low-resolution image has the same dimension as the receptive field of the pre-trained MDE network. However, to determine the optimal dimension for the high-resolution image denoted by R_x, they consider the contextual cues by constructing an edge map from the input image's gradients. The notation R_x indicates that $x\%$ of pixels do not receive any contextual information at a given image resolution. Through experimental analysis, they determine R_{20} to be the optimal dimension to estimate the high-resolution depth and note that R_{20} can be much large than the original input resolution. Then they successfully merge the low-resolution depth with a spatial dimension similar to the receptive field size and the high-resolution depth at the R_{20} size through the merging network by resizing the input depth images to the training size of the merging network. They name the output from the merging network on the whole image as *base estimate*.

Additionally, they make use of the second observation that the output behaviour of the pre-trained MDE networks for inputs at different resolutions is also related to the density of the depth cues present in the image. The further away the contextual cues are from the receptive field, the structural inconsistencies are higher in the depth estimations. This observation also indicates that determining the optimal high-resolution size for an image is driven by the regions with the lowest contextual cue density. However, the high-density cues can still benefit from high-resolution estimations. To formalize this observation, they propose a patch selection mechanism to generate high-resolution depth for different patch regions in the image having high contextual cue density. They further merge the patch estimation onto the *base estimation* from the first step. The patch-based refinement on the base estimate further improves the high-frequency depth details in the depth estimation. We show the overall pipeline to demonstrate the double estimation to generate the base estimate and the patch refinement to generate the final high-resolution depth with sharp object boundaries in Figure 3.5.

Chapter 4

Ordinal Depth Estimation

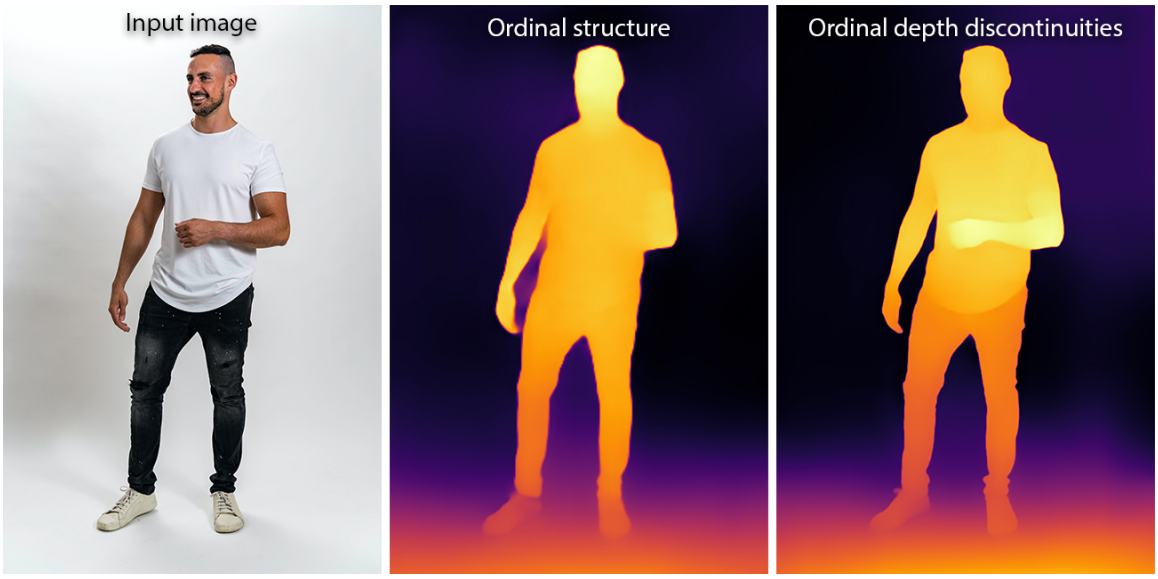

Figure 4.1: Overview of the depth estimations from our ordinal depth for an image in the wild. The depth estimation shows consistent scene structure at low input resolution, while the high input resolution results in depth estimation containing high-frequency depth details.

Following our observation in Chapter 3, to estimate depth with consistent global structure and sharp depth details, we need a depth estimation network that effectively uses both the sparse ranking and dense scale and shift invariant (SSI) based loss formulations. However, naively combining the SSI and ranking loss to estimate depth can negatively impact the network due to the design limitations of the sparse ranking loss and the pseudo-ordinal nature of the SSI loss. The sampling strategy is another key component to provide a more balanced setup for sparse ranking-based loss for effectively estimating depth details.

To overcome the above challenges, we propose an ordinal depth estimation network that combines the sparse and dense losses with good harmony to estimate depth with consistent

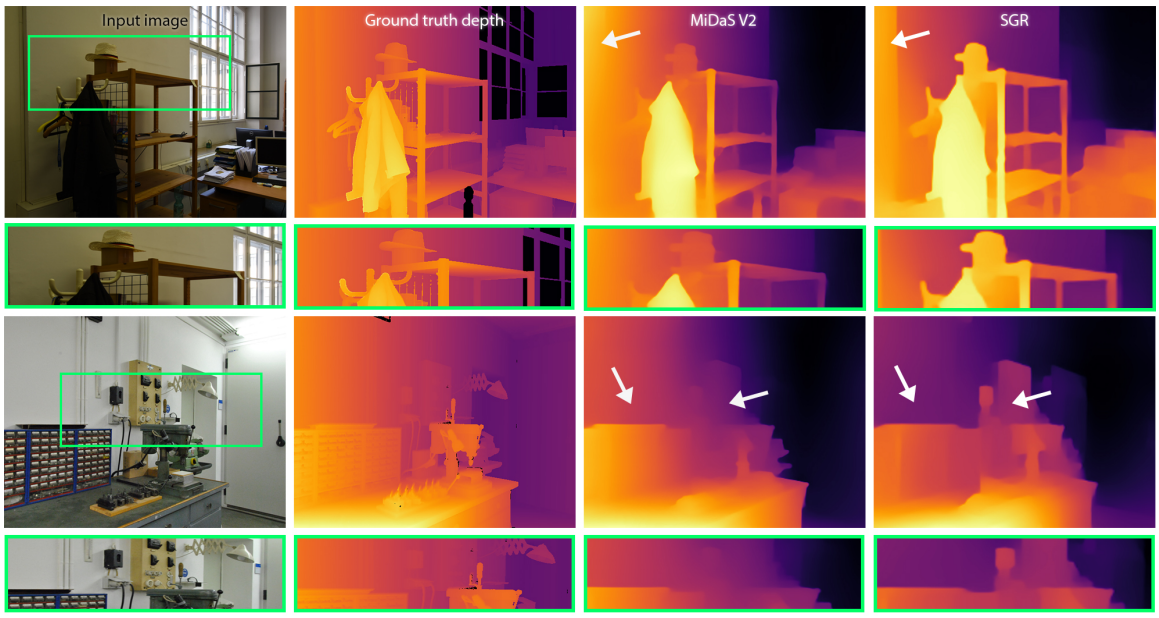

Figure 4.2: Overview of the trade-off between dense and sparse depth loss formulations. The dense pseudo-ordinal loss used in MiDaS [36] estimates overall scene structure but lacks geometric details, as indicated by white arrows. On the other hand, the sparse ordinal loss used in SGR [54] is better are capturing sharp edges but lacks structural consistency, as seen in smooth surfaces.

global structure and sharp depth boundaries. We introduce several key changes to MDE networks to harmonize the dense and sparse losses effectively. Specifically, we introduce an ordinal depth space that removes the geometric constraints to use fully ordinal dense loss. We propose a relaxed ranking loss to make informed penalization for point pairs that do not satisfy the ground truth depth ordinality relation. Further, we propose a triplet sampling of sparse points to provide better ordinal depth context for the sparse relaxed ranking loss and balanced sampling of equality and inequality ordinal pairs.

Algorithm 1 Algorithm to generate ordinal depth space

Data: depth map d
Result: inverse ordinal depth map d_o
Step 1: Compute the depth histogram with bins $\rho = 100$ in the depth space
Step 2: Remove bins with occupancy less than a threshold $\tau = (0.02 \times H \times W)$
Step 3: Transform the ordinal depth to inverse depth (disparity) space
Step 4: Min-max normalize the inverse ordinal depth to constrain the space to $[0, 1]$ to train the MDE networks

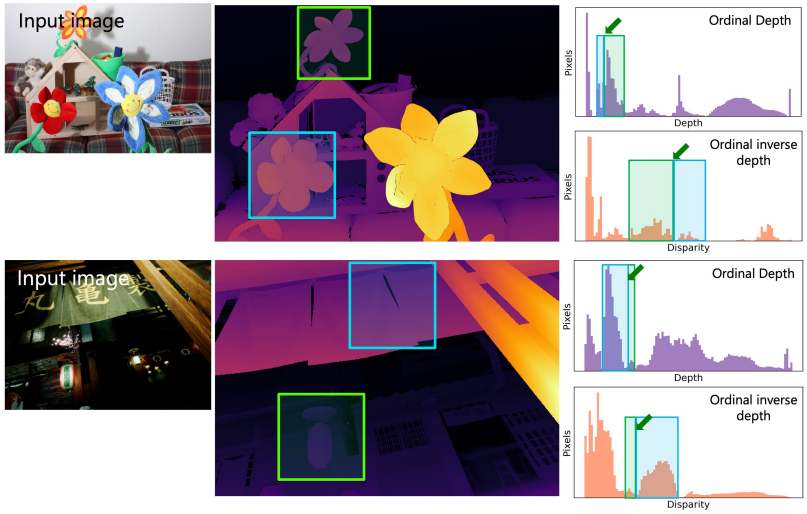

Figure 4.3: Our observation in Figure 3.2 shows that the ground truth depth distribution (or space) with geometric constraints represented by *gaps* limits pseudo-ordinal SSI loss from effective depth estimation. To overcome this, we generate ordinal depth space by removing the geometric constraints in the ground truth depth space while maintaining the depth ordering. Specifically, we remove the bins in the depth histogram that do not meet a minimum occupancy threshold and make the depth distribution more connected as indicated by `green` arrows in both indoor (top row) and outdoor (bottom row) images.

4.1 Ordinal Depth Space and Dense Ordinal Loss

A histogram can represent a depth space to show the distribution of pixels based on their depth values. Computing the histogram for a metric depth map shows similar depth values forming connected clusters. A *gap* in the depth distribution between the depth clusters indicates a drastic change in the corresponding metric depth values. Additionally, the gaps are the basis of the geometric structure in metric depth.

To make the SSI loss completely ordinal, we remove the geometric constraints present in the metric-based depth space (§ 3.1) by proposing an **ordinal depth space**. To perform this transformation to ordinal depth space, we evenly redistribute the depth values into a fixed range. We introduce a technique to efficiently discard gaps in the original depth space shown in Figure 3.2 and generate ordinal depth space that preserves the depth order but not the geometric information. An advantage of this change is that it retains all the rich, sharp depth discontinuities. More formally, we compute the depth histogram for the ground-truth depth and remove bins not meeting a minimum occupancy threshold while retaining the depth order. Next, we normalize the ordinal depth values and generate the inverse ordinal depth. Finally, we constrain the output space to be in $[0, 1]$ to train the MDE models. In Algorithm 1, we describe the steps to compute the ordinal depth space. In Figure 4.3, we show the ordinal depth for the indoor and outdoor images without any gaps in their respective depth distribution.

The ordinal depth is invariant to any strictly increasing function that retains the ordinal relationship between the pixels. Instead of estimating ordinal depth with a fixed scale, we encourage ordinal behaviour by aligning the ground-truth ordinal with the ordinal network estimations using a monotonically increasing affine function. Our dense ordinal loss resembles the SSI loss formulation (§ 3.1) but optimizes in the ordinal depth space. We can estimate the scale and shift in the inverse ordinal space using the least-squares criterion:

$$\mathbf{h} = \Big(\sum_i \vec{\mathbf{o}}_i \vec{\mathbf{o}}_i^T\Big)^{-1}\Big(\sum_i \vec{\mathbf{o}}_i \mathbf{o}_i^*\Big), \quad \vec{\mathbf{o}}_i = (\mathbf{o}_i, 1)^T, \tag{4.1}$$

where $\mathbf{h} = (scale, shift)^T$, $\mathbf{o}_i^*$ and $\mathbf{o}_i$ indicate the ground-truth inverse ordinal depth and estimated inverse depth values for the pixel i, respectively.

We can define the dense ordinal loss between the aligned ordinal depth estimation and the ordinal ground truth image as:

$$\mathcal{L}_{ord}(\mathbf{o}, \mathbf{o}^*) = \frac{1}{2M}\sum_i (\vec{\mathbf{o}}_i^T \mathbf{h} - \mathbf{o}_i^*)^2, \tag{4.2}$$

The dense loss formulation does not penalize the estimations from the network if the linear ordering of the depth values in the ordinal depth estimations is correct. Furthermore, the ordinal depth estimations are a scale and shift away from the ground truth ordinal depth.

4.2 Relaxed Ranking Loss

The goal of sparse ranking loss is to push pixels with ordinal inequality relations far apart and bring pixels with ordinal equality relations closer to each other. To that end, we propose a sparse loss called **relaxed ranking loss** that introduces a simplified ordinal relation between pairs of pixels. It discourages any loss between point pairs if their difference satisfies a minimum depth threshold ($\lambda_{thresh} > 0$) or if they are separated by a distance of λ_{thresh}. Additionally, we use a ReLU function to compute the loss for point pairs that do not satisfy the depth threshold or ground-truth ordinal relation. We show the loss curve for the relaxed ranking loss in Figure 4.4. Unlike the ranking loss curve, our loss does not penalize the points as long as they respect the ground-truth ordinal relation. Another benefit of the relaxed ranking loss is that it combines well with other losses (e.g., dense ordinal, dense pseudo-ordinal, or geometric loss), as it does not generate conflicting signals for point pairs that already obey the depth ordering. The relaxed ranking loss helps generate sharper details than the ranking loss discussed in Section 5.2.

For a pair of points a_{pred} and b_{pred} in the predicted depth map, we can define the relaxed ranking loss as:

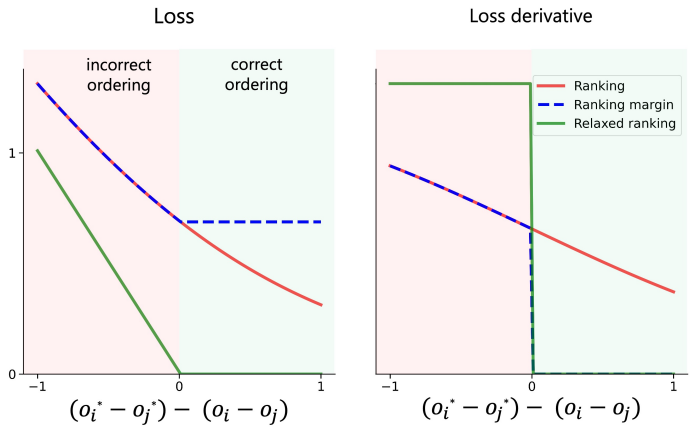

Figure 4.4: The diagram shows the loss (left) and loss derivative (right) curve for the ranking [3], relaxed margin [4], and our relaxed ranking loss. Compared to other losses, we do not penalize the point pairs as long as they satisfy the depth ordering or ground-truth ordinality, as seen in the graph.

$$\mathcal{L}_{rrl}(a_{pred}, b_{pred}) = \begin{cases} \texttt{max}(-\phi \cdot (a_{pred} - b_{pred}) + \lambda_{thresh}, 0) & \text{if } \phi \neq 0 \\ (a_{pred} - b_{pred})^2 & \text{if } \phi = 0 \end{cases} \tag{4.3}$$

We empirically determine $\lambda_{thresh} = 0.01$, and ϕ indicates the ordinal relation determined based on the corresponding pair of points in the ground truth depth with respect to a threshold $\delta = 0.01$. The ordinality relation (ϕ) for the corresponding ground truth points a_{gt} and b_{gt} is given by:

$$\phi = \begin{cases} +1 & \text{if } a_{gt} - b_{gt} > \delta \\ -1 & \text{if } a_{gt} - b_{gt} < \delta \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

Additionally, we can call the relaxed ranking loss as *sparse ordinal loss*. Our ordinal depth network combines sparse and dense ordinal loss to estimate ordinal depth with high-frequency details and consistent ordinal scene structure.

Our sparse ordinal loss is similar to the ranking loss [3] in applying sparse supervision using the sampled points to estimate depth. However, a unique trait of the ranking loss is that it continues to compute a high gradient for pixels despite the estimated ordinality relation matching the ground truth for sampled point pairs. Specifically, the ranking loss encourages the depth of the sampled point pairs to be very different if their ground truth ordinal relation is inequality. Our sparse ordinal loss overcomes this issue by only penalizing the points if their depth difference is not greater than λ_{thresh} for ground truth inequality

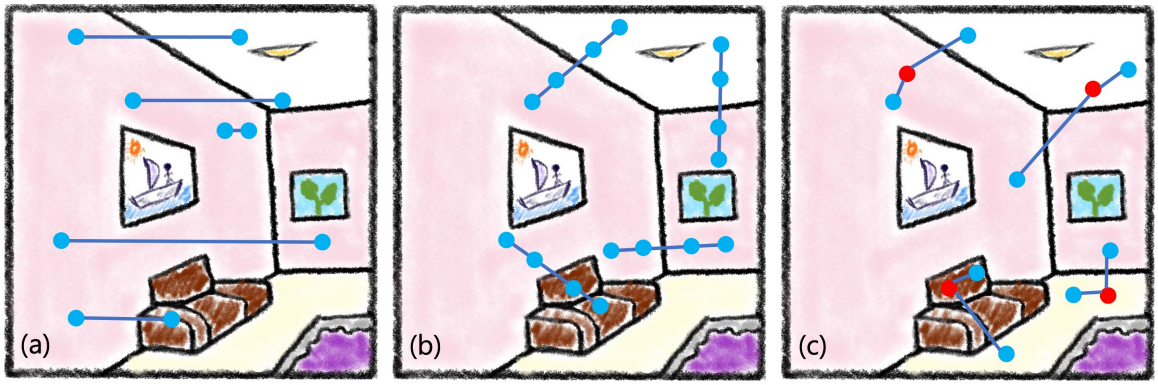

Figure 4.5: We present an illustration to compare our triplet sampling strategy with the point pairs sampling adopted in previous monocular depth estimation networks with ordinal loss formulations. For triplet sampling, we choose an anchor point (in red circles) and randomly select positive (same depth) and negative (different depth) ordinal points (d).

ordinal relations. We show this phenomenon (red line) of the ranking loss in Figure 4.4. Additionally, this characteristic of the ranking loss negatively affects the combined training setup with the dense ordinal loss. A similar observation was shown by Chen *et al.* [4] when combing the ranking loss and dense geometric loss.

4.3 Triplet Sampling

Most prior MDE approaches [3, 53, 54] with ranking loss formulations consider sampling pixel pairs to determine the depth ordinality. However, the sampling techniques in prior approaches suffer from imbalanced point pairs. Specifically, the number of point pairs with inequality relation is more than the equality constraint. Moreover, the ranking loss requires only point pairs to compute the loss. Hence, they do not have an additional reference depth value to guide the pushing and pulling of point pairs. Specifically, if one of the points in the pair has incorrect depth estimation, then both points will be pulled closer together. This phenomenon can generate gradients that conflict with other dense losses, as some correct pixel estimations continue to receive a high loss and gradient. Therefore, there are more efficient ways to use in sparse ordinal type loss formulations than sampling point pairs. To address this limitation, we consider sampling pixel triplets instead of pixel pairs [3, 53, 54]. Specifically, we randomly sample an anchor point and select two other pixels based on a depth threshold following the previous MDE approaches: one with positive ordinality relation and another with negative ordinality relation with the anchor point. The pixel with positive ordinality has the same depth as the anchor point, and the pixel with negative ordinality has a different depth value (either higher or lower) from the anchor point.

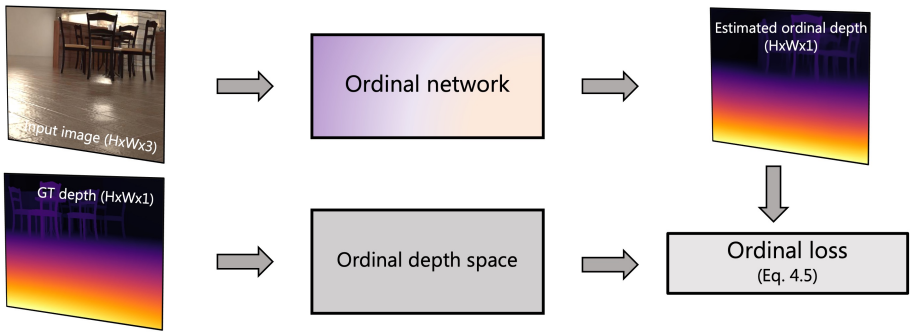

Figure 4.6: We provide an overview of the training pipeline of the ordinal network. First, for an input RGB image and ground-truth depth map, we pre-process the ground-truth depth to generate the ordinal depth. Then, the ordinal network estimates the depth for an input image, and the loss is computed based on Eq. 4.5 between the estimated and ground-truth ordinal depth as described in Section 4.4.

The sampling generates balanced ordinal positive and negative values, unlike the prior pixel pair sampling mechanisms [3]. We illustrate our triplet sampling procedure in Figure 4.5.

The depth threshold we use is uniform across all images and generates reliable and effective ordinal depth as described in Section 5.2. However, there are two extreme scenarios where triplets cannot be constructed. Only the dense ordinal and multi-scale depth gradient losses (§ 4.4) will be computed in these scenarios to train the network. One such scenario has a slight variation in depth, or the depth difference between points is less than the uniform threshold (e.g., flat wall facing the camera). In this case, it is not possible to sample points with an ordinal inequality relation with the anchor point. Similarly, the second scenario with significant variations in depth or the depth difference between any point pair is more significant than the uniform threshold (e.g., hallway), and in this, it is not possible to find points with an equality relation for the anchor.

4.4 Ordinal Network

Due to the complementing properties of the dense ordinal loss and our sparse ordinal loss, we combine them to generate ordinal depth with better details than dense-only training and a better structure than sparse-only training. Our final loss for the ordinal training combines dense ordinal, sparse ordinal, and multi-scale gradient loss [25] to train the ordinal depth network in the inverse ordinal depth space. The goal of the multi-scale gradient loss ($\mathcal{L}_{msg}$) is to enforce local depth smoothness.

Finally, we define the overall combined loss for training the ordinal depth network as:

$$\mathcal{L} = \lambda_{ord}\mathcal{L}_{ord} + \lambda_{rrl}\mathcal{L}_{rrl} + \lambda_{msg}\mathcal{L}_{msg} \tag{4.5}$$

We set the loss weights $\lambda_{ord} = 10$, $\lambda_{rrl} = 10$, and $\lambda_{msg} = 0.5$ when training ordinal depth network. We show the overview of our ordinal training setup in Figure 4.6.

4.5 Datasets

We follow the recent monocular depth estimation approaches [35, 36, 58] to train our *ordinal depth network* on a diverse set of datasets for better generalization. Specifically, we train our model on the combination of the following datasets:

Omnidata [10]: Omnidata proposes a framework to generate "steerable" datasets from 3D scans of environments. In particular, we use the Hypersim [37], Replica [43], and Replica [43] + Google Scanned Objects [7] datasets from Omnidata.

The **Hypersim** dataset from Omnidata [10] consists of diverse photo-realistic synthetic images rendered for 461 complex 3D indoor scenes [37] with significant depth variations. Each scene contains 100 images rendered from different camera positions with an image resolution of 768×1024. We consider the official train split from [37] for training the ordinal depth network.

The **Replica** dataset contains 18 real-world indoor scenes with an image resolution of 512×512. A custom RGBD camera setup with an infrared (IR) projector captures all the data. In addition, the Replica dataset provides high-quality 3D meshes by fixing the planar surface for holes in the scanning process. Omnidata uses the 3D meshes to render $\sim$ 57K training images from different camera poses.

The **Replica+GSO** is a dataset rendered in Omnidata setup by scattering the 3D objects [7] in Replica [43] indoor scenes. In addition, the GSO dataset captures over 10K 3D scans of real-world individual objects using a standard off-the-shelf RGBD capture setup. Finally, the omnidata framework renders a total of $\sim$ 108K training images.

OpenRooms [26]: OpenRooms dataset renders photo-realistic synthetic data from 3D scans collated from existing open repositories of indoor scenes with different lighting and material setups. The dataset generates $\sim$ 100K images at 480×640 image resolution.

TartanAir [51]: TartanAir dataset contains photo-realistic scenes rendered from the Unreal Engine using the AirSim plugin. The dataset comprises 30 diverse indoor and outdoor scenes captured from a drone camera with complex trajectories to capture diverse views of the scene with different lighting conditions and detailed scenes.

FSVG [22]: The FSVG dataset contains images captured from Grand Theft Auto V [15]. The dataset consists of $\sim$ 100K training images by implementing an autonomous agent to navigate the scene. The capture starts at a random location in the environment and ends after collecting 30 seconds of data before resetting.

HRWSI [54]: The HRWSI comprises $\sim$ 21K web stereo pairs, and an optical flow network is used to generate the disparity maps based on the forward-backward consistency

check to use reliable values. A segmentation network sets the sky region in an image to have the minimum disparity value.

Holopix50K [18]: The Holopix50k dataset consists of $\sim$ 50K image pairs captured from the Holopix platform. We compute the disparity maps from RAFT [45] and consider sky segments from Mask2Former [6] to set corresponding disparity values to zero.

We perform zero-shot depth evaluation on the four datasets not seen during training similar to prior MDE approaches [36, 54]. We consider DIODE [47], IBims-1 [20], Middlebury [40], and KITTI [16] datasets. In DIODE, we use the validation set of 771 images, including indoor and outdoor images. In IBims-1, we use the official test split of 100 images for the evaluation. To evaluate Middlebury, we consider all the training and test images. Finally, we consider the Eigen split of 697 images for evaluation in KITTI.

4.6 Implementation Details

For our *ordinal depth network* experiments, we follow the network setup of [54] with a ResNeXt101 [55] feature extractor. We initialize the feature extractor ResNeXt101 [55] with WSL weights [29] similar to the setup in [36] trained in a weakly-supervised learning setup on a large corpus of images, with further fine-tuning on ImageNet1K dataset. We use Sigmoid to restrict the output depth values to [0, 1] during training. We use a learning rate of 10^{-5} with an Adam optimizer. For datasets with outdoor images, we set the disparity of the sky to zero in every image. To better stabilize the ordinal training, we consider warm-starting the network by considering $\sim 5K$ images by randomly sampling 20 images from every scene in the Hypersim [37] dataset and training the network with an L1 loss for one epoch. We then continue the training on the Hypersim [37] and OpenRooms [26]. We then incrementally add the other datasets in the following order: Replica [43], Replica [43]+GSO [7], FSVG [22], TartanAir [51], HRWSI [54], and Holopix50K [18]. We construct a batch size by uniformly sampling the images from every dataset and setting the batch size to 16.

To crop the input image for training, we follow the setup from Miangoleh *et al.* [32] to compute the $\mathcal{R}_0$ size such that no pixel in the image is far away from the contextual cues. Since we upper bound the cropping resolution to R_0, any resolution less than that ensures every pixel receives context information. This cropping mechanism helps the network fit the entire image with contextual cues into the receptive field and thus helps the network estimate consistent global ordinal structure with details. We then randomly crop from [384, $\mathcal{R}_0$] and resize to 384×384. We then randomly apply horizontal flip, color jitter, Gaussian blur, and grayscale data augmentation operations for better generalization. During training, we adjust the ground-truth distribution in the depth space as described in Section 4.1 and compute our overall ordinal loss in Section 4.4. We match the scale and shift of the predicted ordinal estimate with the ground truth ordinal depth using the least-squares criterion before computing the combined ordinal loss (§ 4.4).

Chapter 5

Ordinal Depth Evaluation

This chapter evaluates our ordinal depth network trained with sparse and dense ordinal losses. We show quantitative and qualitative results of our ablation study to demonstrate the effectiveness of different components in our ordinal depth network training setup. We compare the ordinal depth network with the state-of-the-art ordinal networks to show the improvement in performance on novel datasets not seen during training. Finally, we present the results of our ordinal network with boosting framework [32] to show the strong generalization capabilities for photographs in the wild.

5.1 Evaluation Metrics

We consider several evaluation metrics to evaluate different attributes of depth estimation. First, we use the ordinal (Ord.) loss to evaluate the ordinality between depth points by randomly sampling 10K points in an image. We use D^3R [32] to measure the high-frequency details in the estimated depth. To reduce the inference time, we resize the images with a larger axis equal to the network's training size and adapt the smaller size to preserve the aspect ratio for depth prediction. Additionally, we align the predictions from our ordinal network with the ground-truth depth in the inverse-depth space using the least-squares criterion.

The Ordinal evaluation metric is Ord. $= \sum_i^n w_i \mathbb{I}(l_i \neq l_i^*))/\sum_i^n w_i$, where n is the total number of sampled pixel pairs and w is set to 1. l indicates the determined ordinal relation for a pair i. To quantify the quality of edge accuracy, we use D^3R from Miangoleh *et al.* [32]. Specifically, to calculate D^3R, the first step is to compute superpixel segments on the ground truth depth and construct pixel pairs based on the neighbouring segment centres. The pixel pairs indicate the depth accuracy for object boundaries and drastic changes to depth. The error metric on the n pixel pairs is given by: $D^3R = (\sum_i^n |l_i - l_i^*|)/n$.

To further measure the edge accuracy, we employ the Soft Edge Error (SEE) by Chen *et al.* [2]. The SEE metric computes the absolute difference in error between a value in the estimated depth and a local patch of size $k \times k$ around a point at the same location in the

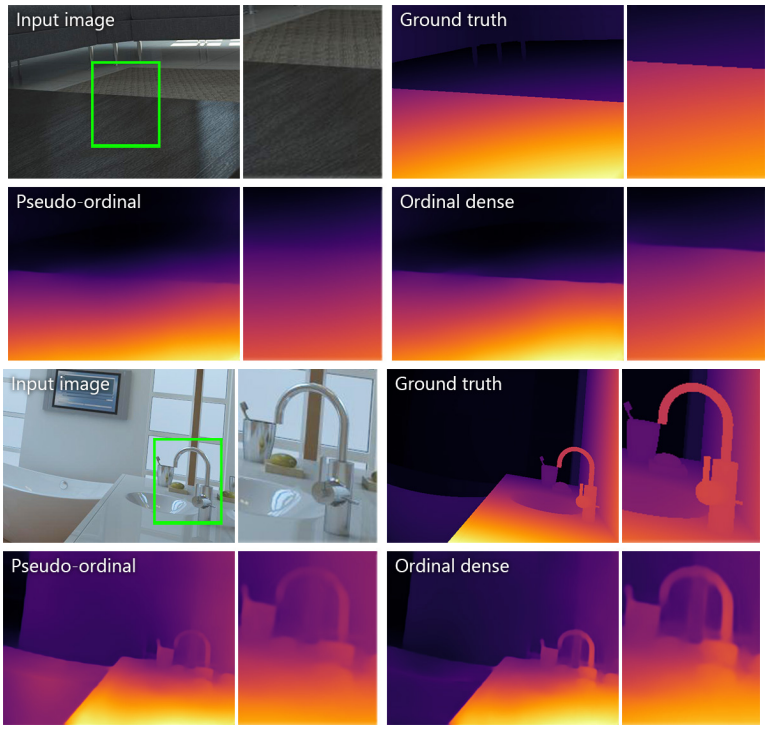

Figure 5.1: We present an example from our controlled experiment in Section 5.2 where we compare the pseudo-ordinal dense loss by [36] with our dense ordinal loss. The fully ordinal formulation of our loss allows the network to generate sharper details.

ground-truth depth. The SEE metric is defined as $1/n \sum_i^n see_k(d_i - d_i^*)$, where $i \in Edge(d^*)$ and $see_k = \texttt{min}(|d_i - d_j^*|), j \in n_k(i)$. In SEE, n_k indicates the $k \times k$ neighbourhood patch for the point i.

During training and inference, we resize the images to the corresponding network's training resolution such that the larger axis matches the training size and adapt the minor axis to maintain the aspect ratio. Specifically, we resize the images for our network to ensure the larger axis is set to 384 while adapting the minor axis. For KITTI, with a wide aspect ratio, we follow the setup by Ranftl *et al.* [36] to set the minor axis to 384 and adapt the larger axis to handle the aspect ratio. Finally, all the estimations are rescaled to the original image size to compute the evaluation metrics, except for D^3R.

5.2 Ablation Study

In this section, we evaluate the effectiveness of our proposed ordinal depth training strategies and inspect the contribution of each of our proposed components. Specifically, we analyze different losses for training monocular ordinal depth estimation, the effect of combining the sparse and dense ordinal loss, and different ordinal sampling strategies.

Table 5.1: We perform an ablation study to evaluate the effectiveness of the ordinal depth network on IBims-1 [20], Middlebury [40], and Hypersim [37] datasets. We compare our ordinal loss with scale and shift invariant (SSI) [36] loss and sparse Ranking loss [3]. Additionally, we compare our triplet point sampling with random [3], random balanced [53], and structure-guided sampling [54].

Methods	IBims-1		Middlebury		Hypersim	
	Ord. ↓	RMSE ↓	Ord. ↓	RMSE ↓	Ord. ↓	RMSE ↓
Dense Loss						
Pseudo-ord SSI	0.150	0.604	**0.161**	0.202	0.138	0.427
Ordinal SSI	0.143	0.570	0.170	<u>0.199</u>	0.136	0.428
Sparse Loss						
Random + Ranking	0.179	0.768	0.182	0.200	0.176	0.585
Random + Ordinal	**0.126**	<u>0.553</u>	0.182	0.209	**0.126**	<u>0.424</u>
Random Balanced + Ordinal	0.139	0.605	0.195	0.215	**0.126**	0.426
SGR + Ordinal	0.139	0.574	0.187	0.212	**0.126**	0.427
Triplet + Ordinal	0.136	0.558	0.189	0.219	0.129	0.434
Combined Loss						
SSI + Triplet + Ordinal	<u>0.135</u>	**0.537**	<u>0.167</u>	**0.193**	<u>0.128</u>	**0.413**

We experiment with our method on the Hypersim dataset [37] from Omnidata [10]. First, we construct the training set by sampling 10K images and their associated depth maps from the provided train split [37] of the data. In particular, we sample 20 images for each scene in the Hypersim training data. Then, we train our depth estimation models on the constructed training data and evaluate the depth estimation performance on Hypersim [37], Middlebury [40], and IBims-1 [20] using standard depth estimation metrics, including SSI-RMSE, and Ord.

Table 5.1 compares the overall depth estimation for different ordinal loss setups. We first compare the model trained using SSI loss in the pseudo-ordinal depth space and our ordinal depth space. The results in rows 1 & 2 in Table 5.1 across different datasets and metrics show that using ordinal depth space with dense SSI loss effectively estimates ordinal depth. Furthermore, removing metric constraints to generate ordinal depth space is beneficial in learning sharp depth edges, as shown in Figure 5.1. In contrast, the pseudo-ordinal trained network generates smooth edges and lacks the object's structural details. In rows 3 & 4, we compare the ranking loss and our relaxed ranking loss. Our relaxed variant effectively improves the overall numerical results for all the datasets, indicating it works well for diverse datasets. This is because our sparse-only training with relaxed ranking loss does not push the depth values very far apart as long as they satisfy the minimum threshold described in Section 4.2. In contrast, the ranking loss negatively impacts the training by pushing the objects very far from each other despite the point pairs satisfying the ground truth

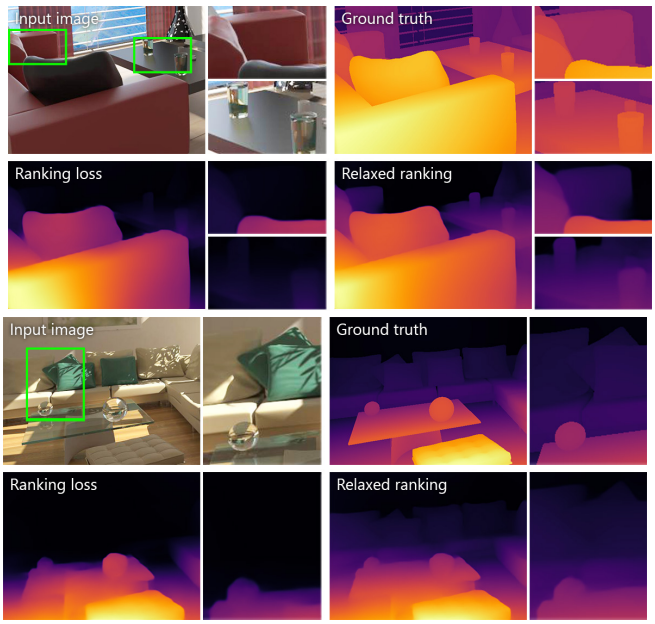

Figure 5.2: We showcase the qualitative results from our controlled experiment described in Section 5.2 to compare the ranking [3] and our relaxed ranking loss. Our relaxed ranking loss captures sharper details compared to the ranking loss. Additionally, our loss does not push the objects in the background too far from the foreground as long as the depth order is satisfied, which is not the case in the ranking loss.

ordinality relation, as described in Figure 5.2. In addition to positively impacting training, the relaxed ranking loss also improves the sharp details in the depth estimation.

In rows 5 to 7, we study the effect of different point sampling strategies for our relaxed ranking loss. Numerically, our triplet sampling is on par with other mechanisms and performs well on the *Ord.* metric to show that using the triplet sampling of sparse points in the relaxed ranking loss estimates better ordinal depth. Compared to the other sampling techniques (random, balanced random, and structured-guided samplings), our triplet sampling generates better depth details (see Figure 5.3). Specifically, creating a reference depth for the positive and negative sampled points helps establish a better context for the foreground and background objects, leading to better overall results.

To get the best of all the different components, we combine them in the last row in Table 5.1. Specifically, the final setup involves the dense ordinal loss and our sparse relaxed ranking loss with triplet sampling. Numerically, the combined setup achieves the best overall results across all datasets. The *RMSE* points to the global structure of the depth estimation effectively estimated by the combined loss. In addition, the *Ord.* metric indicates that the complete ordinal dense and sparse training is effective in estimating ordinal depth. In Figure 5.4, we compare the results from our combined loss with the sparse and dense-only loss setups. The combined loss displays consistent overall structure as the dense-only loss and sharp depth details as the sparse-only loss. This indicates that our dense and sparse

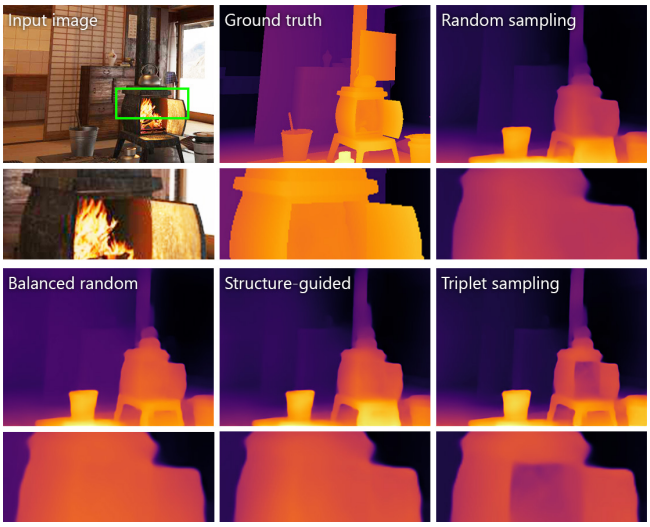

Figure 5.3: In this figure, we present the qualitative results of the experiment to study the impact of the point pair sampling technique on the ordinal depth estimations. We compare the random [3], balanced random [53], structure-guided sampling [54] and our triplet-based sampling techniques. Our sampling technique helps to infer sharper depth discontinuities compared to others.

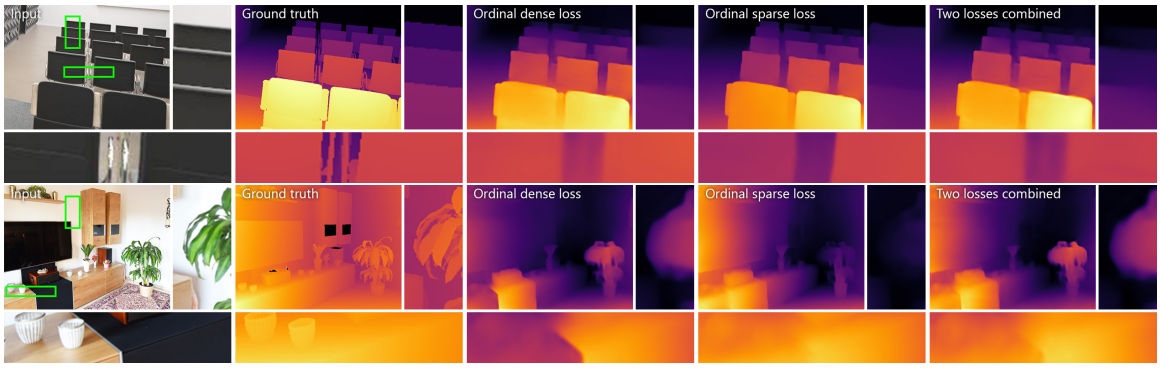

Figure 5.4: We present an example from our controlled experiment in Section 5.2, where we show that our dense and sparse losses harmonize well to generate better structure and details.

Table 5.2: We consider a zero-shot quantitative evaluation of our proposed ordinal network "Ours-ordinal" with other state-of-the-art ordinal depth networks using depth ordinal and edge accuracy metrics. Specifically, we compare with VN TPAMI [56], SGR [54], LeReS [59], MiDaS [36], DPT [35], 3D KenBurns [33]. Additionally, we compare our boosted [32] ordinal depth "Ours-boosted" with SGR and MiDaS.

Methods	Middlebury				IBims-1						DIODE				KITTI	
	$SEE_{k3}\downarrow$	$SEE_{k5}\downarrow$	Ord. $\downarrow$	$D^3R\downarrow$	$SEE_{k3}\downarrow$	$SEE_{k5}\downarrow$	Ord. $\downarrow$	$D^3R\downarrow$	$\varepsilon^{acc}_{DBE}\downarrow$	ε^{comp}_{DBE}	$SEE_{k3}\downarrow$	$SEE_{k5}\downarrow$	Ord. $\downarrow$	$D^3R\downarrow$	Ord. $\downarrow$	$D^3R\downarrow$
VN TPAMI	0.120	0.115	0.211	0.600	0.157	0.149	0.140	0.683	4.795	79.595	0.227	0.219	0.207	0.944	0.150	0.085
SGR	0.169	0.163	0.220	0.516	0.191	0.182	0.199	0.579	2.127	48.068	**0.130**	**0.124**	0.247	0.899	0.166	0.060
SGR-bmd	0.160	0.154	0.210	0.304	0.188	0.179	0.196	0.489	2.049	28.367	0.140	0.134	0.236	0.906	0.161	**0.052**
LeReS-ordinal	0.150	0.144	0.197	0.456	**0.101**	**0.096**	0.107	0.502	2.399	23.940	0.221	0.213	0.204	0.916	0.159	0.071
MDS	0.111	0.106	0.176	0.451	0.135	0.128	0.127	0.502	**1.874**	45.755	0.174	0.167	0.175	0.888	0.120	0.071
MDS-bmd	0.102	0.098	0.164	0.223	0.132	0.125	0.126	0.438	1.901	33.415	0.180	0.174	0.182	0.892	**0.117**	0.069
DPT	0.104	0.099	**0.161**	0.369	0.143	0.136	**0.100**	0.436	2.001	29.740	0.176	0.169	**0.166**	0.884	0.186	0.091
KenBurns	0.134	0.129	0.219	0.525	0.132	0.125	0.122	0.674	2.190	22.840	0.183	0.177	0.237	0.941	0.136	0.078
Ours-ordinal	0.098	0.093	0.196	0.404	0.111	0.105	0.113	0.439	1.878	25.404	0.142	0.136	0.202	0.901	0.129	0.082
Ours-bmd	**0.094**	**0.090**	0.185	**0.170**	0.114	0.109	0.125	**0.324**	1.914	**12.709**	0.155	0.149	0.222	**0.880**	0.127	0.063

losses complement each other in the combined setup to achieve the best result for ordinal depth estimation.

5.3 Comparison with Ordinal State-of-the-art

We perform the zero-shot evaluation of our ordinal depth network following evaluation setup by Ranftl *et al.* [36]. We compare our results with several prior baselines to compare the effectiveness of our ordinal network. The Virtual Normal (VN TPAMI) network by Yin *et al.* [56] combines a virtual normal on sparse normals with a dense scale and shift invariant (SSI) loss. The structured-guided ranking (SGR) approach by Xin *et al.* [54] propose to use a ranking loss with pair sampling based on the image gradients as a proxy for depth discontinuities to estimate better depth edges in sparse-only training. The LeReS [59] method uses a variation of dense SSI with normal-based sparse loss to estimate relative depth. The Midas [36] uses the SSI loss to train a depth network across diverse datasets that accounts for the unknown scale and shift in dense-only training. The KenBurns [33] estimates a coarse depth and an up-sampling module to estimate high-resolution depth with accurate boundaries. The DPT [35] uses the same training setup as Midas but replaces the convolutional architecture with the recent vision transformer [9] based architecture. Additionally, we consider the boosted versions of SGR (SGR-bmd) and Midas (Midas-bmd) using the depth boosting framework by Miangoleh *et al.* [32] to estimate high-resolution ordinal depth.

In Table 5.2, we compare our two approaches, "Ours-Ordinal" and "Ours-bmd", with all the prior methods. On the high-resolution Middlebury [40] dataset, the results from our boosted ordinal network show strong performance on all the edge-based metrics. This shows that our ordinal network is better at capturing high-frequency details for high-resolution input images than all the baselines. However, for the ordinal metric (Ord.), the DPT baseline shows strong results as it has a larger network capacity than our convolutional-based

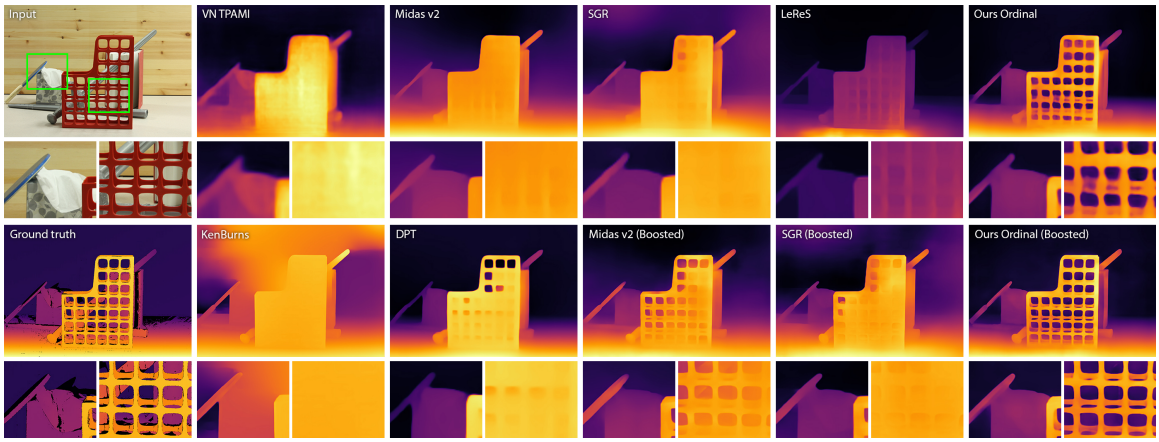

Figure 5.5: We present a zero-shot qualitative comparison against state-of-the-art ordinal and pseudo-ordinal depth estimation methods: VN [56], Midas [36], SGR [54], KenBurns [33], DPT [35] and our ordinal depth estimation method. We also show high-resolution results for Midas, SGR and Ours using the boosting method by Miangoleh *et al.* [32]. We detail our experimental setup in Section 4.6.

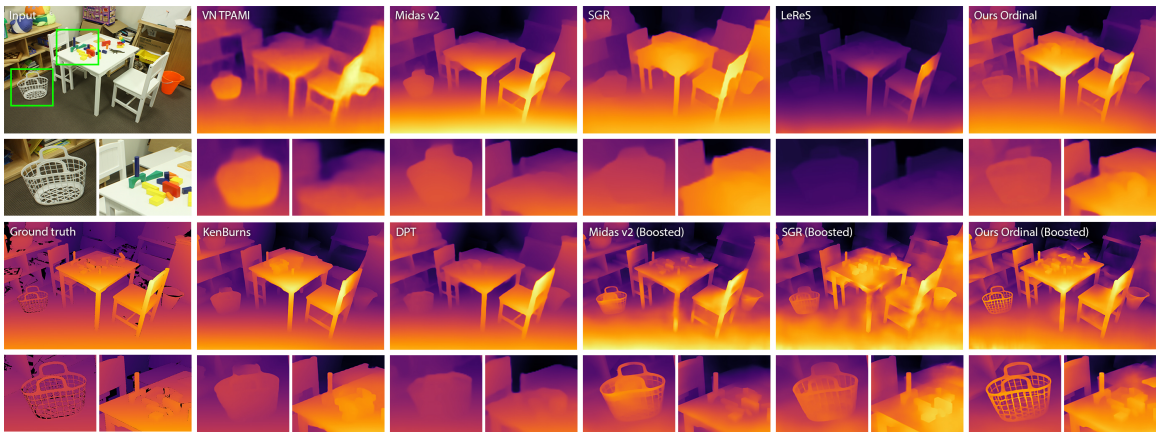

Figure 5.6: We compare our ordinal network with previous state-of-the-art ordinal and pseudo-ordinal depth estimation methods on Middlebury [40].

network to capture complex scenes. In Figure 5.5, we show qualitative results on the Middlebury dataset. Our ordinal networks outperform other methods in precisely capturing the depth edges to show holes in the foreground object. Additionally, our ordinal network is effective at discerning the information inside holes as background, which other approaches fail to capture. Further, in Figure 5.6, our approach captures the high-frequency details of intricate objects like a basket and toys on the table for a complex indoor image in Middlebury.

On IBims-1 [20] dataset, our ordinal networks show strong results on the edge accuracy metrics. However, LeReS has better performance on the SEE_k metrics. A possible explanation is that LeReS involves a sparse loss for surface normals on points sampled along strong gradient edges in the image using structure-guided sampling. This explicit supervision gives LeReS an advantage for the SEE_k metrics. Again, DPT shows better performance for the ordinal (Ord.) metric due to the large model capacity. In Figure 5.7, we compare the qualitative results on images from the IBims-1 dataset for all the methods. Our ordinal networks demonstrate the effectiveness of capturing intricate details in images. In particular, our approaches get precise and sharp depth boundary edges of the leaves and the table. Similarly, in Figure 5.8, our methods infer the shape of complex and small objects like the cables and the arm of the overhead projector.

On the DIODE [47] dataset, our ordinal networks perform better than most of the baselines apart from SGR and its boosted counterpart for the SEE_k metrics. Again, DPT shows a stronger result for the ordinal (Ord.) metric due to network capacity. For the D^3R metric, our networks show competitive results with respect to other models.

For the KITTI [16] dataset, as the ground truth is sparse, we only consider the metrics that consider a sparse set of points to evaluate a model. Our ordinal networks do not have the best results on KITTI as the input images are not high-resolution with unusual aspect ratios but show comparable results to the best-performing SGR model.

Our combined loss has a consistent ordinal structure than sparse-only methods and higher depth details than dense-only networks. Overall our combined ordinal dense and sparse loss with the multi-scale gradient depth loss does help the network to generate depth with both consistent global structure at low resolution and sharp high-frequency depth details at higher resolutions based on our boosted ordinal depth estimations.

5.4 In-the-wild Ordinal Depth Estimation

We train our ordinal network with the combined loss on a variety of synthetic and real-world datasets, covering a wide range of environments to make it generalize for in-the-wild images. To test this depth generalization hypothesis, we collected diverse images from Unsplash contributed by different photographers. The image set contains complex scenes with dynamic objects like people, birds, and animals. Further, we employ the boosted [32]

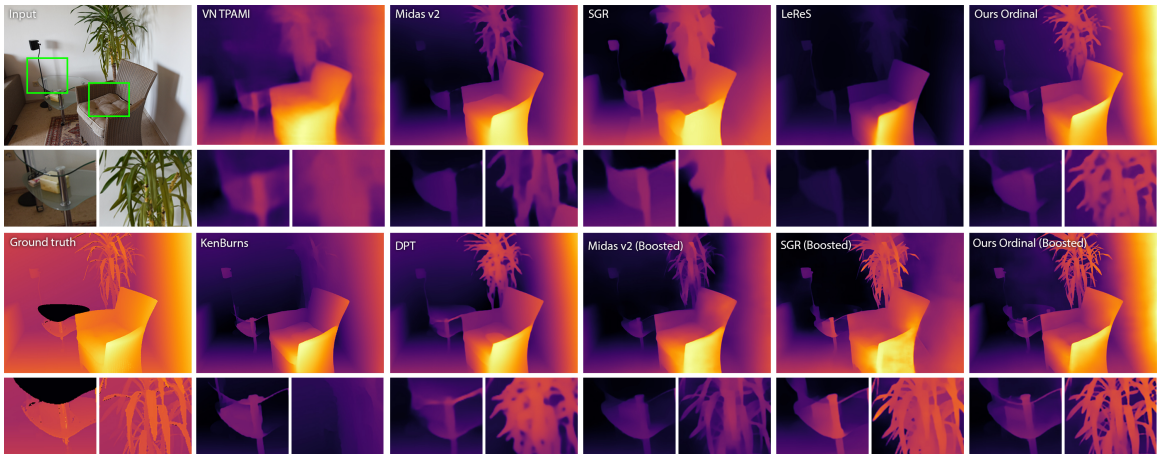

Figure 5.7: We compare our ordinal network with previous state-of-the-art ordinal and pseudo-ordinal depth estimation methods on IBims-1 [20].

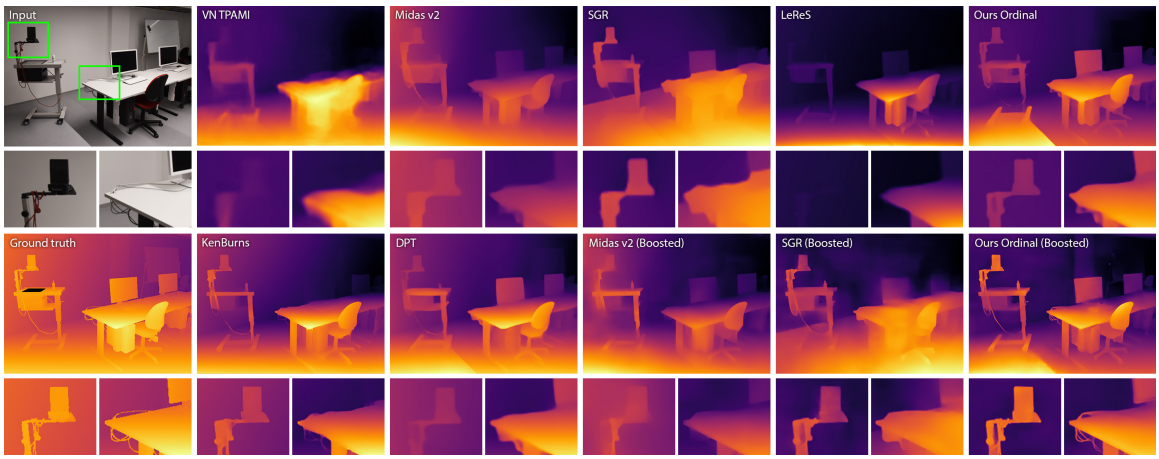

Figure 5.8: We compare our ordinal network with previous state-of-the-art ordinal and pseudo-ordinal depth estimation methods on IBims-1 [20].

Figure 5.9: We present the high-resolution qualitative results from our ordinal network with the boosting technique [32] for in-the-wild images capturing complex indoor and outdoor environments of different objects.

version of our ordinal depth network to generate high-resolution ordinal depth images. We show the results of this setup in Figure 5.9. Despite not training on datasets containing animals, our boosted ordinal results show consistent global ordinal structure with sharp depth boundaries for animals, as seen in result (a). For images with multiple objects of the same type, our result shows consistent object structure with clear depth boundaries to separate the objects, as seen in images (b and d). Additionally, our depth estimation shows a strong result on images with humans. Specifically, for a close-up shot of a human face with intricate hair structure, our results capture the subtle details of facial hair in the depth estimation, seen in result (c). Similar to the images with animals, our ordinal network can estimate reliable high-resolution ordinal depth for images containing birds with intricate features, as seen in result (e). Finally, for outdoor images with a group of people and trees, our depth estimation shows consistent global structure while capturing all the tiny details of the trees, as seen in result (f). These results thus prove our generalization hypothesis that our ordinal depth network can estimate reliable depth for photographs in the wild.

Chapter 6

Metric Depth Estimation

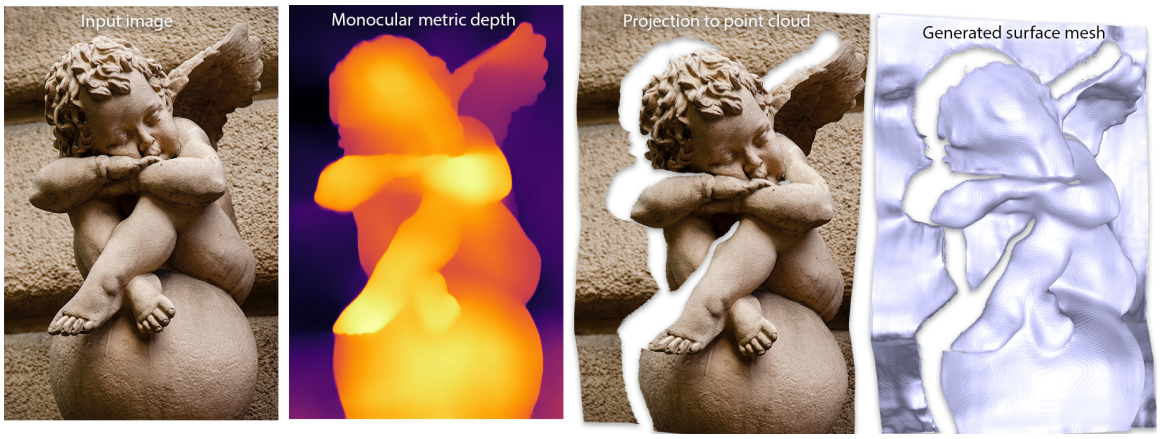

Figure 6.1: The depth estimation from our metric depth network on a photograph in the wild. Our depth estimation is high-resolution with geometric consistency. We can project the metric depth estimation to a dense 3D point cloud. Finally, we can leverage an off-the-shelf meshing network to recover the surface mesh from the 3D point clouds.

In Chapter 4, we introduce our fully ordinal depth estimation method. Although our approach achieves better depth details and structure in the ordinal space, as highlighted in Section 5.3, the depth estimation still lacks geometric structure, which can hinder its usage in different downstream applications [33, 41].

The convolutional nature of the ordinal depth network can be utilized to generate depth estimations for input resolutions greater than the training resolution of 384×384. Specifically, varying the input resolution of the image generates depth with different characteristics. Specifically, our ordinal depth estimation captures the overall ordinal scene structure for low-resolution (receptive field size) input, as the network can see the complete image. However, for high-resolution input, our ordinal network estimates depth with high-frequency local details but with inconsistent global scene structure. We provide an overview of these observations in Figure 6.2. A similar observation was shown by Miangoleh *et al.* [32] on

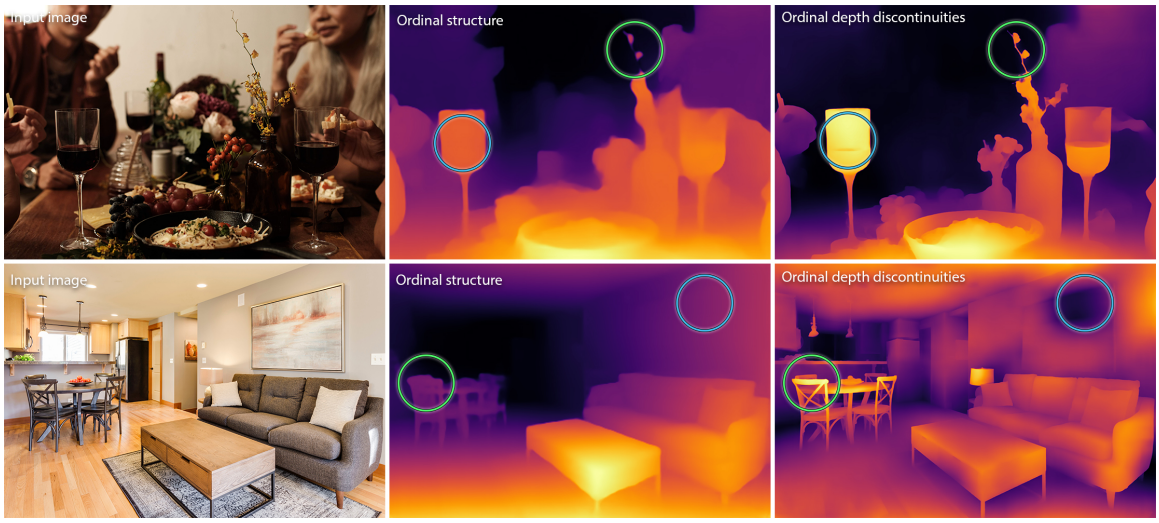

Figure 6.2: Overview of the characteristics of the low and high-resolution ordinal estimations. The low-resolution ordinal estimates capture the scene structure (green circles) but lack high-frequency details (blue circles). On the other hand, the high-resolution ordinal estimates are good at generating sharp depth discontinuities (green circles) but lack structural consistency (blue circles).

pre-trained depth networks and exploited this behaviour of the pre-trained depth networks to boost their depth estimations to high resolution.

We set up the high-resolution metric depth estimation problem with two ordinal depth inputs and an RGB image. The first ordinal depth estimation is at the network receptive field size of the ordinal network, 384×384, to capture the global ordinal relationships in the input scene. The second ordinal depth is a high-resolution estimation with resolution $\mathcal{R}_{20}$ as defined by Miangoleh *et al.* [32] (§ 3.4), which is adaptive to images based on their content and the edge distribution. The advantage of our setup is the readily available depth information for the metric network to estimate high-resolution metric depth. Instead of reasoning about the scene structure and depth details from a monocular image for high-resolution metric depth estimation, our metric depth network can leverage the local and global ordinal information provided in the inputs. In addition to the rich information in the input data, we apply different geometric constraints on the network to explicitly fix the geometric structure of the depth estimation.

The high-resolution metric depth with two ordinal depth estimations setup resembles the mid-level depth estimation framework by Zoran *et al.* [60]. Specifically, Zoran *et al.* [60] leverages the local and global context in the images to estimate sparse ordinal depth relations. The local context captures local image formations, while the global context captures the overall image structure. The sparse ordinal depth relations are then utilized in a CRF-based optimization to generate dense depth. In contrast to Zoran *et al.* [60], instead

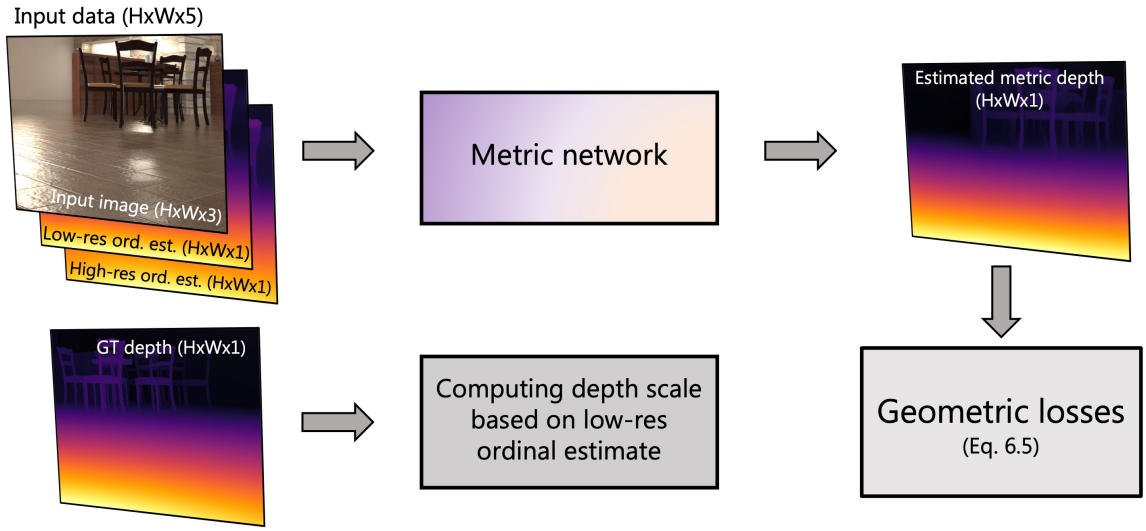

Figure 6.3: Overview of the high-resolution metric depth training pipeline. Given an input image and the ground-truth depth, we first generate low and high-resolution ordinal estimates from the pre-trained ordinal network. Then, we channel-wise concatenate the input image, low-res, and high-res ordinal estimates of dimension $H \times W \times 5$ as input to the metric network. We also fix the ground-truth depth scale based on the low-resolution ordinal estimate using the least-squares criterion. Finally, we compute the scale-invariant loss on the estimated metric depth and the scaled-fixed ground-truth depth to penalize the network during training. The scale-invariant loss in Eq. 6.5 is a combination of sparse ratio, multi-scale gradient, surface normal, and multi-scale surface normal gradient loss as described in Section 6.5.

of estimating sparse ordinal depth relations, we train the ordinal depth network in the ordinal depth space to estimate two dense ordinal depth estimations to capture the local and global contextual information. Specifically, the global context is captured by our ordinal estimation for the input image at the receptive field size. Meanwhile, the local context is provided by our ordinal depth estimation for the input image with a resolution higher than the receptive field size. We then use our metric depth network to take the ordinal inputs along with the RGB to estimate dense metric depth, similar to the CRF-based optimization by Zoran *et al.* [60].

We use the different sparse and dense losses to effectively train the metric depth network to enforce geometric constraints in both the depth and the surface normal space. We use a dense loss, a sparse depth ratio loss, and a multi-scale depth gradient loss for the depth space. In addition, we use dense angle loss and a multi-scale normal gradient loss in the normal space.

6.1 Dense Loss

The "metric" depth is defined up to an unknown scale. The prior MDE methods [12, 24, 25, 56] adopt the scale-invariant loss formulation to estimate metric depth. During training, similar to the scale and shift-invariant (SSI) loss [36], the scale-invariant loss requires a least-squares fit between depth estimation and ground truth metric depth before computing the loss. However, following this approach to determine the scale can make our dense loss unstable due to inaccurate depth estimation from an under-trained metric network in the early training epochs. Computing the scale on these inaccurate depth estimations and the ground truth depth can negatively affect the least-squares fit, leading to high gradients and poor network convergence.

In our metric depth estimation setup with ordinal inputs, we can use the low-resolution ordinal estimate as a point of reference for our metric depth network. Specifically, the low-resolution ordinal estimate from the pre-trained ordinal depth network is stable. The least-squares fit between the low-resolution ordinal estimate and the ground truth produces reliable results. Therefore, we use our low-resolution ordinal depth estimation (o^L) to set the arbitrary scale in the ground truth depth (d^*) as given by:

$$c = \underset{c}{\operatorname{argmin}} \sum_i (c \cdot d_i^* - o_i^L)^2 \tag{6.1}$$

We can use this fixed scale c to define our ground-truth depth as $\hat{d}^* = c \cdot d^*$. We can also fix the arbitrary scale of the high-resolution ordinal depth estimate to that of the low-resolution ordinal estimation to ensure that both the inputs and output depth maps have the same overall scale. We can define the loss between the metric depth estimation d and the scaled fixed ground truth $\hat{d}^*$ as:

$$\mathcal{L}_{dsi} = \frac{1}{n} \sum_i |d_i - \hat{d}_i^*|, \tag{6.2}$$

where n is the total number of pixels in the image.

A limitation of this setup is that if the ordinal estimations are noisy, then the scale we determine for the ground truth depth can be unreliable. However, the ordinal network can generate noisy depth if the input images contain noise, as described in Section 9.1.

6.2 Sparse Ratio Loss over Triplets

In addition to the scale-invariant loss (§ 6.1), we define a sparse ratio loss over triplets to enforce an additional geometric constraint. This is based on the motivation that each pair of points in the depth estimation should differ by the same ratio as the corresponding points in the ground- truth, regardless of their global scale. Inspired by this setup, we define the loss as:

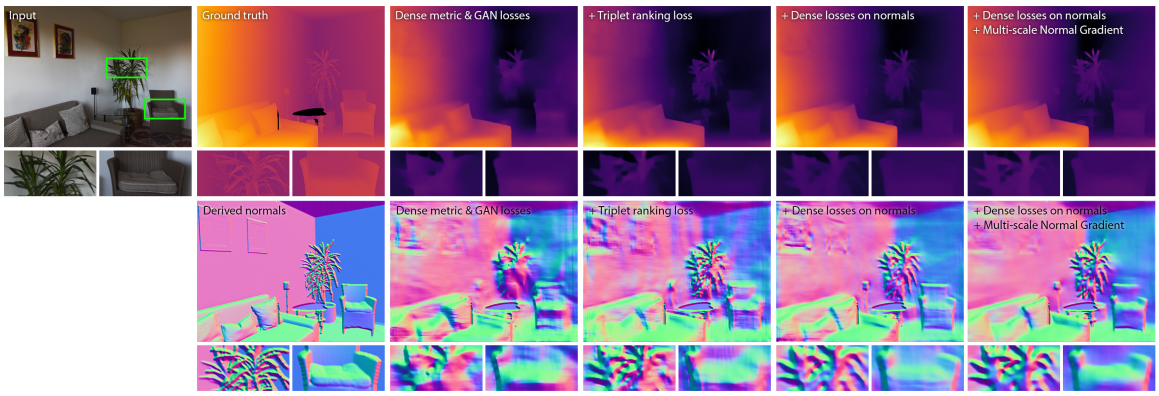

Figure 6.4: Our sparse metric loss helps the network generate sharper details, while our dense losses on normals make it possible to generate smooth surfaces on IBims-1 [20].

$$\mathcal{L}_{srl}(i,j) = \begin{cases} \left(\frac{\hat{d}_i^*}{\hat{d}_j^*} - \frac{d_i}{d_j}\right)^2 & \text{if } \hat{d}_j^* > \hat{d}_i^* \\ \left(\frac{\hat{d}_j^*}{\hat{d}_i^*} - \frac{d_j}{d_i}\right)^2 & \text{otherwise} \end{cases} \tag{6.3}$$

In the above equation, both cases ensure the defined loss is smaller than one, making the loss more stable on scenes with a large depth ratio between the foreground and background objects (e.g., outdoor images). We define the sparse depth ratio loss on the points selected with our triplet sampling. As seen in ordinal training with triplet sampling (§ 4.3), using it in metric training enhances the high-resolution details in the prediction, as discussed in Section 7.2.

6.3 Dense Surface Normal Loss

In addition to the depth-based losses to estimate global geometric structure, we impose constraints on the local surface geometry. To achieve this, we compare the surface normal maps derived from the estimated and ground-truth depth. We derive the surface normals by normalizing a vector $\mathbf{n} \in \mathbf{R}^3$, where $\mathbf{n} = [\nabla d_x, \nabla d_y, 1]^T$ and ∇ indicates the depth gradient. We can define the dense normal loss ($\mathcal{L}_{dsn}$) as:

$$\mathcal{L}_{dsn} = \sum_{i}^{n} (1 - \mathbf{n}_i \cdot \mathbf{n}_i^*) \tag{6.4}$$

where, n is the total number of image pixels. The loss in Eq. 6.4 is similar to the angle loss defined in Chen *et al.* [4].

6.4 Multi-Scale Normal Gradient Loss

Inspired by the multi-scale gradient loss [25] on depth estimations, we consider computing the multi-scale gradient loss on the surface normals. As the surface normals are derived from the depth maps, considering a multi-scale gradient would be equivalent to determining the second derivative of the depth. Therefore, the second derivative gives information on the local curvature of the depth. We show the effectiveness of using this loss in Figure 6.4 when applied to other geometric losses for training the scale-invariant network. We define the loss as $\mathcal{L}_{nmsg}$.

6.5 Overall Loss Function

To train the metric depth estimation network, we combine all the aforementioned losses to enforce local and global geometric constraints. We can define the overall metric loss as:

$$\mathcal{L}_{scale-invariant} = \lambda_{dsi}\mathcal{L}_{dsi} + \lambda_{srl}\mathcal{L}_{srl} + \lambda_{sn}\mathcal{L}_{sn} + \lambda_{nmsg}\mathcal{L}_{nmsg} + \lambda_{msg}\mathcal{L}_{msg} + \lambda_{lsgan}\mathcal{L}_{lsgan} \quad (6.5)$$

where, $\mathcal{L}_{dsi}$ is the dense scale-invariant loss, $\mathcal{L}_{srl}$ is the sparse depth ratio loss over triplets, $\mathcal{L}_{dsn}$ is the dense surface normal loss, $\mathcal{L}_{nmsg}$ is the multi-scale surface normal gradient loss, $\mathcal{L}_{msg}$ is the multi-scale depth gradient loss, and finally $\mathcal{L}_{lsgan}$ is the Least Squares GAN [30]. We set the loss weights λ_{dsi}, λ_{srl}, λ_{sn}, λ_{nmsg}, λ_{msg}, and λ_{lsgan} to 1000, 500, 100, 5, 500, and 1, respectively.

6.6 Datasets

To train our metric depth network, we use the following datasets: Hypersim [37], Sun RGBD [42], and TartanAir [51] dataset. The Hypersim and TartanAir datasets are described in the training datasets part of Section 4.5.

SunRGBD [42] dataset comprises $\sim$ 11K RGBD images from real-world environments of indoor scenes captured from four sensors. In addition, the depth maps from cameras are further improved to remove noise and holes by combining depth from nearby frames. We consider the Hypersim and Sun RGBD datasets to train the indoor metric network while training the outdoor network using the TartanAir dataset. The motivation behind training separate metric networks for indoor and outdoor scenes is due to the difference in the depth range between indoor and outdoor scenes. While our ordinal network can handle both environments by utilizing our ordinal depth space, which relaxes the metric constraints, it is challenging for the metric network. Therefore, we resort to training two models to handle this complexity in metric depth. For the indoor network, we utilize Hypersim and compute all the geometric losses (§ 6.5). However, for the Sun RGBD dataset, we only consider the

sparse depth ratio loss on the sampled points using our triplet sampling. We only consider the TartanAir dataset with all the geometric losses for training for the outdoor network.

To test the metric network, we follow the same zero-shot evaluation setup on four test datasets described for ordinal network evaluation in Section 4.5.

6.7 Implementation Details

To train the metric depth network, we consider the Pix2Pix [19] training setup. We use the EfficientNet [44] backbone network with ImageNet pretrained weights for the encoder. We modify the first layer to consider input with five channels. The input resolution is $1024 \times 1024 \times 5$, where 5 is the number of input channels. We channel-wise concatenate the RGB input with low and high-res ordinal estimates. We train the network with a batch size of 1 and randomly crop the image with size $(H/2 \times W/2)$ by resizing it to 1024×1024. Then, we use the metric geometric losses (§ 6.5) and the Least Squares GAN [30] loss with the Patch-GAN setup [19]. In Figure 6.3, we illustrate the overall training pipeline of the metric network. In Figures 7.2 and 7.3, we show the qualitative results from our metric depth network compared to the state-of-the-art baseline networks. In Figure 7.4, we project the depth estimations from the metric network on Middlebury [40] to a dense coherent 3D point cloud.

Chapter 7

Metric Depth Evaluation

In this chapter, we evaluate our metric depth estimation network trained with a combination of depth and surface normal based losses to introduce local and global geometric constraints. We show quantitative and qualitative results of our ablation study to demonstrate the effectiveness of each of the losses in our metric depth network training setup. We compare the metric depth network with the state-of-the-art metric networks to show the improvement in performance on novel datasets not seen during training. Finally, we present the results of our metric network to demonstrate the robust performance against a stereo-based approach that uses multi-view information to estimate depth.

7.1 Evaluation Metrics

To evaluate the metric network for zero-shot setup, we consider the metric depth evaluation metrics along with the ordinal Ord. and D^3R metrics. Specifically, we consider the scale-invariant root mean squared error (SI-RMSE), the absolute relative difference (Abs.), and the depth threshold (δ_t). The depth threshold δ_t metric is defined as $\delta_{1.25} = \max(d/d^*, d^*/d) < 1.25$, where d^* is the ground-truth depth and d is the estimated depth. We can define the absolute relative metric (Abs.) as $1/n \sum_i^n |d_i^* - d_i|/d_i^*$, where n is the total number of image pixels. Additionally, we use the $\varepsilon_{\text{DBE}}^{\text{acc}}$ and $\varepsilon_{\text{DBE}}^{\text{comp}}$ metrics from Koch *et al.* [20] as part of the IBims-1 dataset to measure the accuracy and completeness of the depth edges.

We consider SI-RMSE and Abs. for all datasets but use the depth threshold metric $\delta_{1.25}$ on all datasets except Middlebury [40]. For DIODE [47], IBims-1 [20], and KITTI [16], we compute the SI-RMSE, and absolute relative in the depth space, whereas ordinal and D^3R in the inverse depth space. For Middlebury, we compute all the metrics in the inverse-depth space since the ground-truth data denotes disparity. We consider the least squares criterion to fit the scale of the ground truth to the metric depth estimation before computing the evaluation metrics.

Table 7.1: We perform an ablation study on our metric network to study the impact of different geometric losses. We evaluate the networks on IBims-1 [20] dataset and consider both depth and surface normal based metrics.

Methods	Depth			Angle Distance		% Within t°		
	Ord. ↓	D^3R ↓	RMSE ↓	Mean ↓	Median ↓	11.25 ↑	22.5 ↑	30 ↑
$\mathcal{L}_{dsi} + \mathcal{L}_{lsgan} + \mathcal{L}_{msg}$	0.128	0.414	0.913	37.053	29.093	0.203	0.423	0.532
$+\mathcal{L}_{srl}$	0.134	0.404	0.783	35.317	27.797	0.218	0.440	0.549
$+\mathcal{L}_{srl} + \mathcal{L}_{ssn}$	0.137	0.413	0.839	35.671	27.786	0.218	0.439	0.549
$+\mathcal{L}_{dsi} + \mathcal{L}_{dsn}$	0.121	0.408	**0.760**	31.496	23.761	0.264	0.500	0.610
$+\mathcal{L}_{dsi} + \mathcal{L}_{dsn} + \mathcal{L}_{nmsg}$	**0.120**	**0.368**	0.780	**30.189**	**22.595**	**0.284**	**0.528**	**0.633**

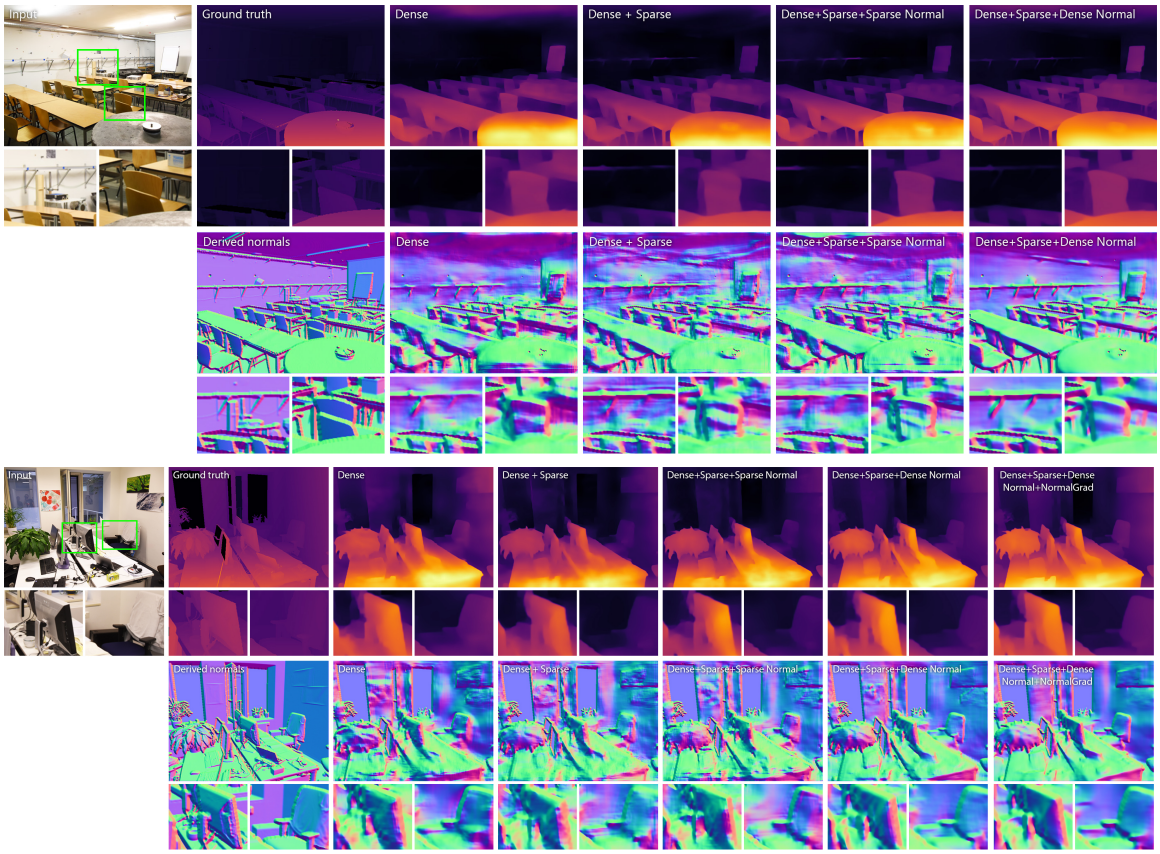

Figure 7.1: Our sparse metric loss helps the network generate sharper details, while our dense losses on normals make it possible to generate smooth surfaces on IBims-1 [20].

7.2 Ablation Study

We demonstrate the effectiveness of different proposed dense and sparse geometric losses in training the metric network. We consider a controlled experiment by training the network setup described in Section 6.7 on just the Hypersim [37] using the data preparation setup detailed in Section 5.2. We consider the standard depth quality metrics *RMSE*, *Ord.*, and D^3R introduced previously. In addition, we use two additional metrics, *Angle Distance* and % *Within t*, specifically designed to assess the geometric quality of depth estimations.

In Table 7.1, we show the quantitative evaluation of the metric network trained with different loss configurations. Training the network with a combination of the dense scale-invariant loss ($\mathcal{L}_{dsi}$), multi-scale gradient loss ($\mathcal{L}_{msg}$), and sparse ratio loss ($\mathcal{L}_{srl}$) generates depth with smooth depth boundaries and uneven surfaces as seen in the depth and derived surface normal qualitative results in Figures 6.4 and 7.1. By adding the sparse ratio loss, we can enforce sharper details and better geometric structure to preserve the depth ratio between pixels according to the ground truth ratio. However, adding the sparse depth ratio loss introduces spiky artifacts on smooth regions as the loss supervision is sparse. Further, adding the sparse surface normal ($\mathcal{L}_{ssn}$) loss to align the surface orientation of the prediction with ground truth does not improve the performance. A reasonable explanation for this behaviour is that sparse normal lacks local context around the sampled point to fix the orientation of the surfaces. An alternative is to use dense surface normal angle loss on all the image pixels. With the addition of dense surface normal loss ($\mathcal{L}_{dsn}$) and this does improve the overall quality of the generated depth maps in terms of sharper depth boundaries, and surfaces appear flatter. Finally, to further improve the local geometric details related to curvature, we apply multi-scale normal gradient loss ($\mathcal{L}_{nmsg}$). The network trained with all the losses (§ 6.5) solves the issues concerning distortions on flat surfaces as they appear smooth with sharper depth details. The combined loss also shows consistent improvement across all the metrics.

7.3 Comparison with State-of-the-art

We perform a zero-shot evaluation of our metric depth network following the evaluation setup by Ranftl [36]. We train two networks independently on indoor and outdoor datasets (§ 6.6) based on the implementation details in Section 6.7. We compare our results with several prior metric depth baselines to compare the effectiveness of our metric network. We compare with a MegaDepth (MD) [25] network trained on depth maps generated from a structure-from-motion (SfM) framework on web images. The network combines scale-invariant loss with sparse ordinal loss to estimate the metric depth. The baseline Mannequin Challenge (MC) [24] trains a network of web videos that represents the mannequin challenge of people. They again use the SfM framework for the ground-truth generation to train the depth network using scale-invariant losses. The Virtual Normal setup by Yin *et al.* [57]

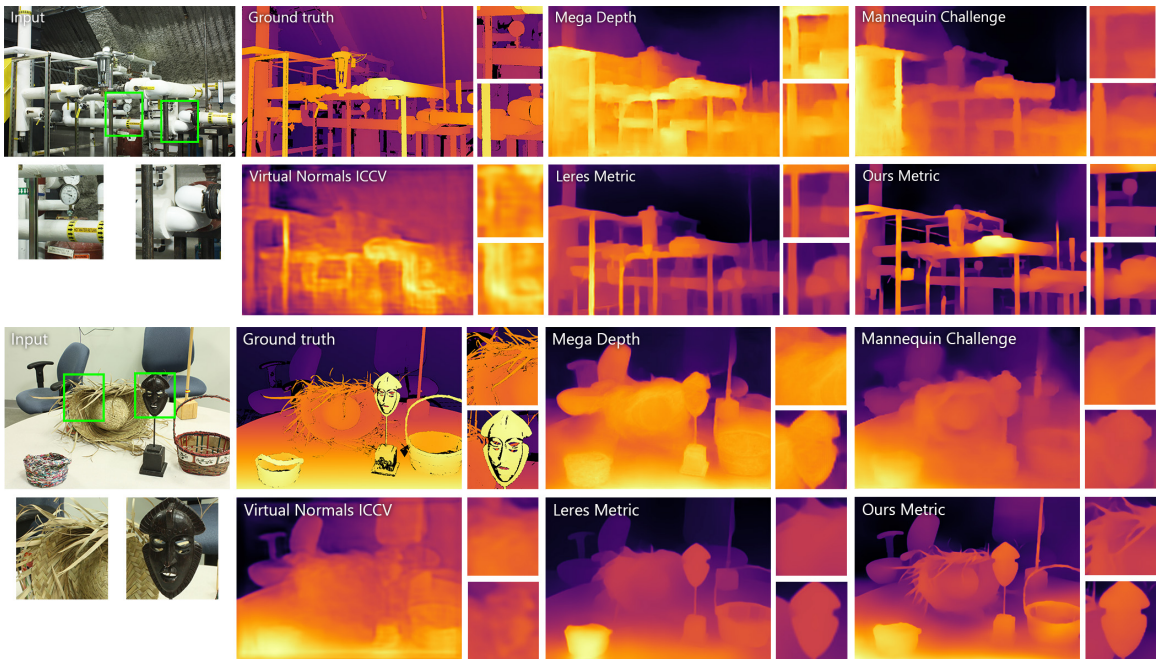

Figure 7.2: We compare the depth estimations across different metric depth networks on Middlebury [40] dataset.

Table 7.2: We consider quantitatively evaluating our metric network "Ours-si" with other prior depth networks. Specifically, we compare with MD [25], MC [24], Virtual Normals [57], and LeReS [59]. We use the scale-invariant depth metrics and edge metrics to compare the performance of different models.

Methods	Middlebury				IBims-1							DIODE					KITTI				
	RMSE ↓	δ_1 ↑	Ord. ↓	D^3R ↓	RMSE ↓	Abs. ↓	δ_1 ↑	Ord. ↓	D^3R ↓	$\varepsilon_{\text{DBE}}^{\text{acc}}$ ↓	$\varepsilon_{\text{DBE}}^{\text{comp}}$ ↓	RMSE ↓	Abs. ↓	δ_1 ↑	Ord. ↓	D^3R ↓	RMSE ↓	Abs. ↓	δ_1 ↑	Ord. ↓	D^3R ↓
MD	0.242	30.070	0.258	0.556	2.200	47.246	48.575	0.268	0.596	3.145	78.144	9.857	45.480	46.560	0.259	0.929	**9.055**	**23.636**	55.361	0.159	0.079
MC	0.229	28.395	0.274	0.690	1.067	22.719	60.580	0.255	0.724	4.083	57.348	10.215	44.546	42.171	0.307	0.945	11.904	31.784	41.137	0.237	0.096
VN ICCV	0.254	29.141	0.255	0.688	**0.738**	13.751	80.441	0.179	0.707	4.089	50.938	10.505	43.762	40.068	0.351	0.952	12.671	36.747	**34.599**	0.321	0.111
LeRes	0.169	28.160	0.199	0.434	0.877	20.155	68.753	**0.107**	0.431	2.252	20.046	**9.225**	40.604	**50.560**	**0.204**	0.911	10.642	25.233	53.826	0.160	0.08
Ours-si	**0.138**	**41.085**	**0.190**	**0.235**	0.766	**13.218**	**81.199**	0.126	**0.324**	**1.642**	**14.304**	10.092	**38.448**	50.359	0.226	**0.901**	11.432	26.502	50.800	**0.117**	**0.074**

employs sparse geometric losses. Finally, Yin *et al.* [59] propose a point-cloud network to estimate the unknown scale and shift for generating metric depth by fixing the ordinal input estimate.

We perform a zero-shot evaluation of our metric depth network following the evaluation setup by Ranftl [36]. We train two networks independently on indoor and outdoor datasets (§ 6.6) based on the implementation details in Section 6.7. We compare our results with several prior metric depth baselines to compare the effectiveness of our metric network. We compare with a MegaDepth (MD) [25] network trained on depth maps generated from a structure-from-motion (SfM) framework on web images. The network combines scale-invariant loss with sparse ordinal loss to estimate the metric depth. The baseline Mannequin Challenge (MC) [24] trains a network of web videos that represents the mannequin challenge of people. They again use the SfM framework for the ground-truth generation to train the

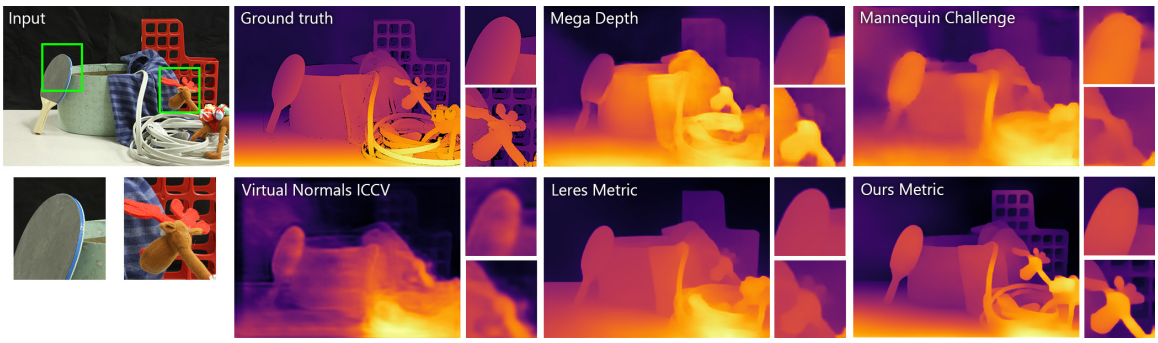

Figure 7.3: Comparison of metric depth estimation methods on Middlebury [40].

depth network using scale-invariant losses. The Virtual Normal setup by Yin *et al.* [57] employs sparse geometric losses. Finally, Yin *et al.* [59] propose a point-cloud network to estimate the unknown scale and shift for generating metric depth by fixing the ordinal input estimate.

In Table 7.2, we show the quantitative comparison of our metric depth network on zero-shot datasets with all the previous baselines. On Middlebury [40] dataset, our metric network "Ours-si" performs the best across all metrics for geometric structure and sharp edges on high-resolution input images. Similarly, on IBims-1, we show consistent improvement in results compared to other baseline metric networks. For DIODE [47] dataset, our metric depth network shows improvements on metrics like Abs. and D^3R. Moreover, it performs on par with other SoTA baselines on the other metrics, indicating that it is good at geometric structure and depth details. On KITTI, our metric network shows comparable results due to the low-resolution unconventional aspect ratio of the input images. We employ an indoor-only model on all datasets except the KITTI dataset. Note that the baseline methods either resort to using homogeneous datasets as used in MC [24] and Virtual Normal [57] or only outdoor landmark data as in MD [25], hence displaying strong quantitative results on datasets with similar data distribution. In addition to our geometric losses, the significant benefit that makes our metric regression easier is the availability of ordinal inputs to the network that assist with sharp depth edges and scene structure.

In Figures 7.2 and 7.3, we present the qualitative results of all the methods. Our metric depth network estimates depth with sharp details on challenging scenes with complex objects. The LeReS [59] first estimates the ordinal depth and then estimates the scale and shift parameters from a secondary point cloud-based network to produce metric depth. Although it can generate geometric scene structure, it does not estimate high-frequency scene details and depth edges as the depth estimation happen at low resolution or the receptive field size of the network. At receptive field size, the network only generates scene structure, with additional post-processing to fix the geometric structure with the point cloud network. In contrast, our network can capture intricate details of complex objects. In Figure 7.2 (top

Figure 7.4: We compare the point clouds generated from the depth estimated by our metric network with LeReS [59] on Middlebury [40] dataset.

row), our metric depth estimation captures high-frequency details of the pipes. Most prior approaches do not capture intricate details of the pipes and create noisy depth estimations. For the image in the bottom row, our metric depth estimation captures all the different objects on the table with sharp details (e.g., hat). In contrast, the baseline approaches wash out all details and create smooth depth for the objects. A major limitation of the baseline approaches is that the depth estimation happens at low resolution. In contrast, our approach leverages low and high-resolution ordinal estimates to provide context on the scene structure and depth details. Similarly, in Figure 7.3, our metric depth estimation is effective at inferring the details of the object in the background, whereas the baselines generate smooth depth of the background object.

In addition to the depth estimation, we also project our detailed depth maps to dense coherent 3D point clouds as shown in Figure 7.4 and compare the point clouds from LeReS. The results in 3D show the details of complex scene structures that our metric depth

Table 7.3: We compare our high-resolution metric with stereo depth SMD-Nets [46] on the UnRealStereo4K [46] dataset. Despite using a single input to estimate high-resolution depth, our network demonstrates strong results in capturing sharp edges. In addition, we consider the soft edge error (SEE) [2] to demonstrate the edge accuracy.

Methods	UnRealStereo4K (3840 × 2160)	
	$SEE_{k3} \downarrow$	$SEE_{k5} \downarrow$
SMD-Nets	66.41	64.97
Ours	**42.16**	**41.31**

captures, unlike the baseline method LeReS at capturing high-frequency details. Specifically, our approach can precisely separate the background seen through the holes in the foreground object (centre column) that the baseline method fails to achieve. A similar observation is in separating the yellow cup (first column) from the background wall in the point cloud generated from our metric depth estimation.

The high-resolution nature of our metric depth with sharp depth edges is also helpful for interactive image editing applications, and one such application is depth-based image segmentation. Specifically, we can perform simple depth thresholding to generate a binary mask to indicate object segment, as shown in Figure 7.5. With this approach, the obtained segments from our metric depth estimation show sharp object boundaries compared to other networks that either do not capture a complete object or contain incorrect boundaries. For instance, our depth can generate a detailed segment of the flower (middle row) compared to LeReS. Similarly, our depth can segment a complete dragonfly with its wings (last row) compared to the baseline LeReS network.

7.4 Comparison with Stereo Depth

We evaluate the metric network against a stereo-depth network that considers multi-view information to construct depth. Specifically, we consider the SMD-Nets [46] network that introduces a depth estimation method using stereo data to estimate high-resolution depth with sharp discontinuities. In Table 7.3, we provide the quantitative comparison using the SEE [2] metric on the depth edges on UnRealStereo4K [46] dataset. Despite considering only a single view (left image) in a zero-shot setup to estimate depth, our metric network shows a stronger result with a lower edge error compared to SMD-Nets [46] that use stereo data.

In Figure 7.6, we qualitatively demonstrate the high-resolution depth estimated from our network compared to SMD-Nets. Our network can better reason about the depth edges than the stereo depth network. In the top row, our network captures smaller objects like stones in greater detail. Similarly, in the bottom row, our network captures sharper details of metallic objects.

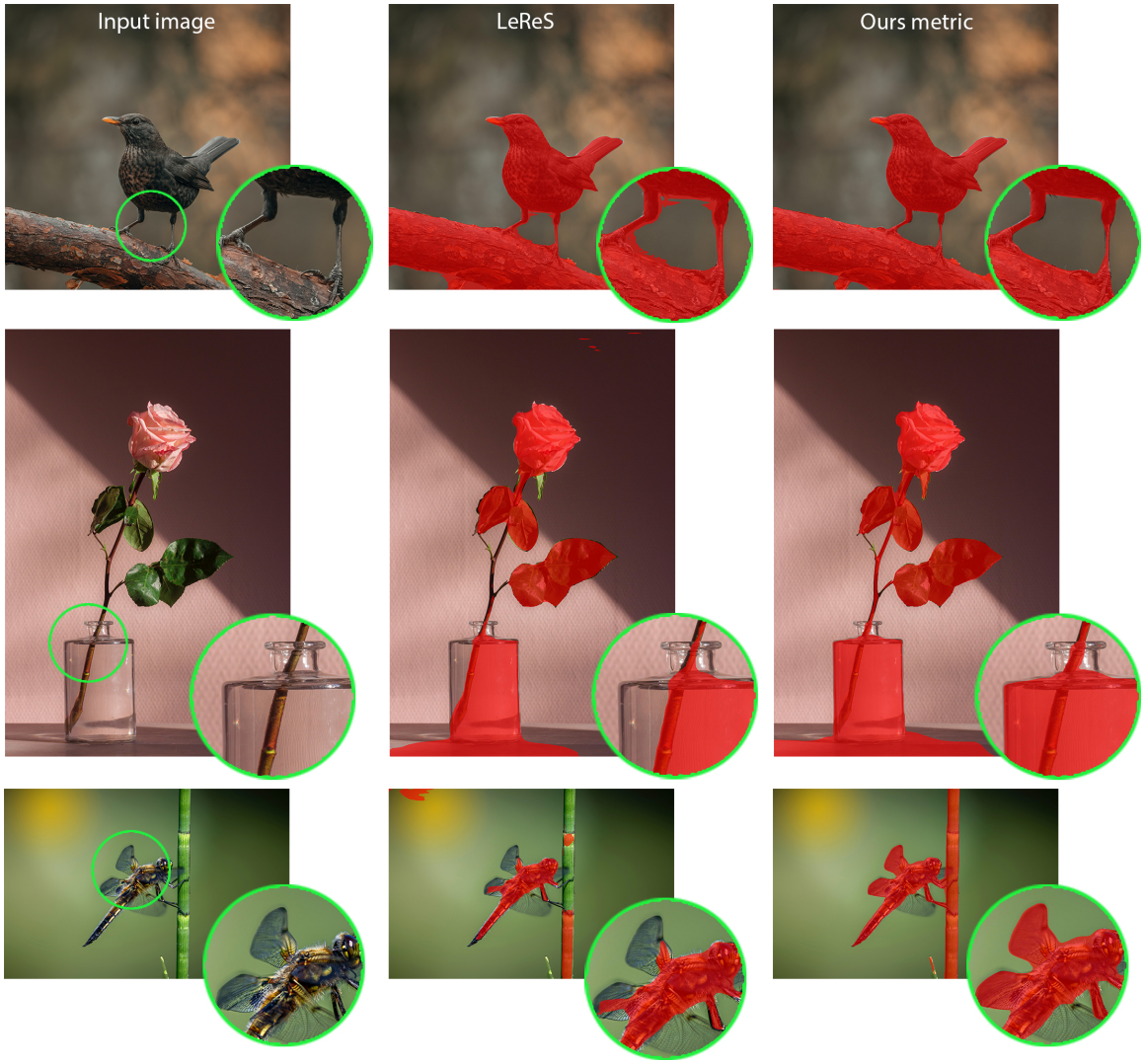

Figure 7.5: Overview of depth-based segmentation by thresholding depth values and overlaid on the input image. This illustration demonstrates the sharp depth discontinuities attained by our metric network. The results illustrate red segments based on the depth values. Our metric network captures complete objects in each segment compared to LeReS [59].

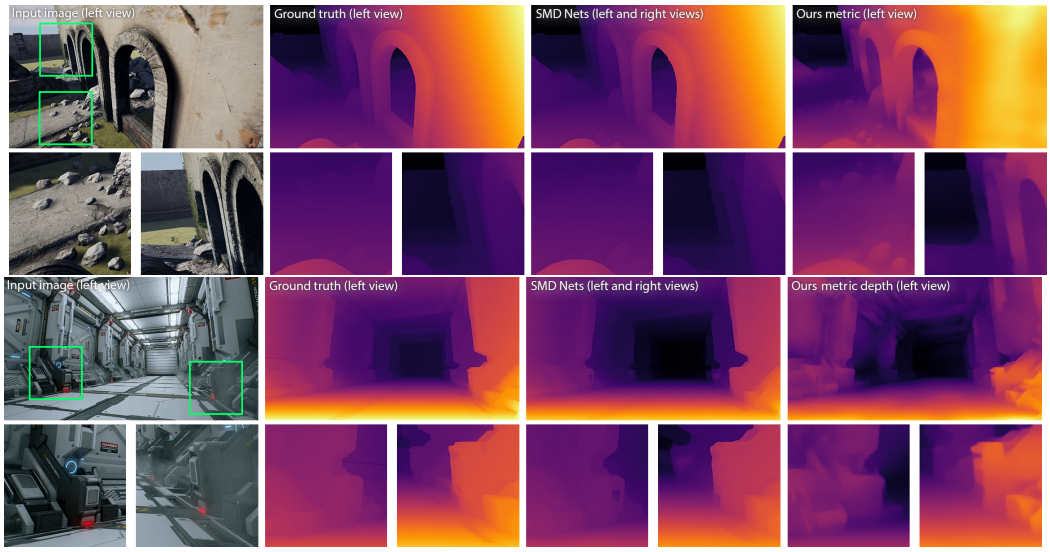

Figure 7.6: We compare our high-resolution metric with stereo depth SMD-Nets [46] on the UnRealStereo4K [46] dataset. Despite using only a single input to estimate high-resolution depth, our network demonstrates strong results in capturing high-frequency depth details compared to the stereo depth network that uses stereo input data, as shown in the insets.

Chapter 8

High-Resolution Metric Depth In-the-wild

In Chapter 4, we introduce the fully ordinal depth network, and in Chapter 6, we introduce our metric network that uses the ordinal estimates to generate high-resolution depth. We now demonstrate the strengths of our high-resolution estimations based on in-the-wild data.

Metric depth enables geometric reconstruction, so we consider projecting the dense, high-resolution depth onto a 3D point cloud. Additionally, we consider filtering out the high-gradient edges in depth to remove the floating points in the 3D representation. Finally, given this dense point cloud, we reconstruct the scene by recovering the surface mesh using the neural dual contouring (NDC) framework proposed by Chen *et al.* [5].

To demonstrate the strength of our metric depth estimation in generating a 3D dense point cloud and recovering geometric consistent 3D surface mesh, we compare it with another previous state-of-the-art metric network from Yin *et al.* [59] (LeReS). We also compare with the high-resolution ordinal depth from Ranftl [36] through boosting framework (BMD Midas v2). In Figure 8.1, we compare the results on images containing humans. The first image (top) contains a person on a couch with other intricate objects, and the second (bottom) captures a person with an empty background. Despite not training on human-only datasets, our metric network captures the consistent structure (such as walls, couch, floor) shown in red inset and high-frequency details (such as plant, human face) shown in green inset for both the humans and the background objects. Although LeReS is a metric depth network, it estimates low-resolution depth with smooth depth edges. In contrast, BMD Midas v2 estimates high-resolution ordinal depth with sharp edges but with a distorted 3D geometric structure as seen on walls (above) and human faces (below).

In Figure 8.2, we show a similar comparison of complex indoor scenes with intricate objects. The first image (top) captures an inside view of a solarium with many plants and lamps and the second (bottom) captures an inside view of a dentist's office. Compared to LeReS and BMD Midas v2, our metric network estimates depth that captures the high-frequency details (such as glass ceiling, plants, lamps, and chairs) shown in green insets

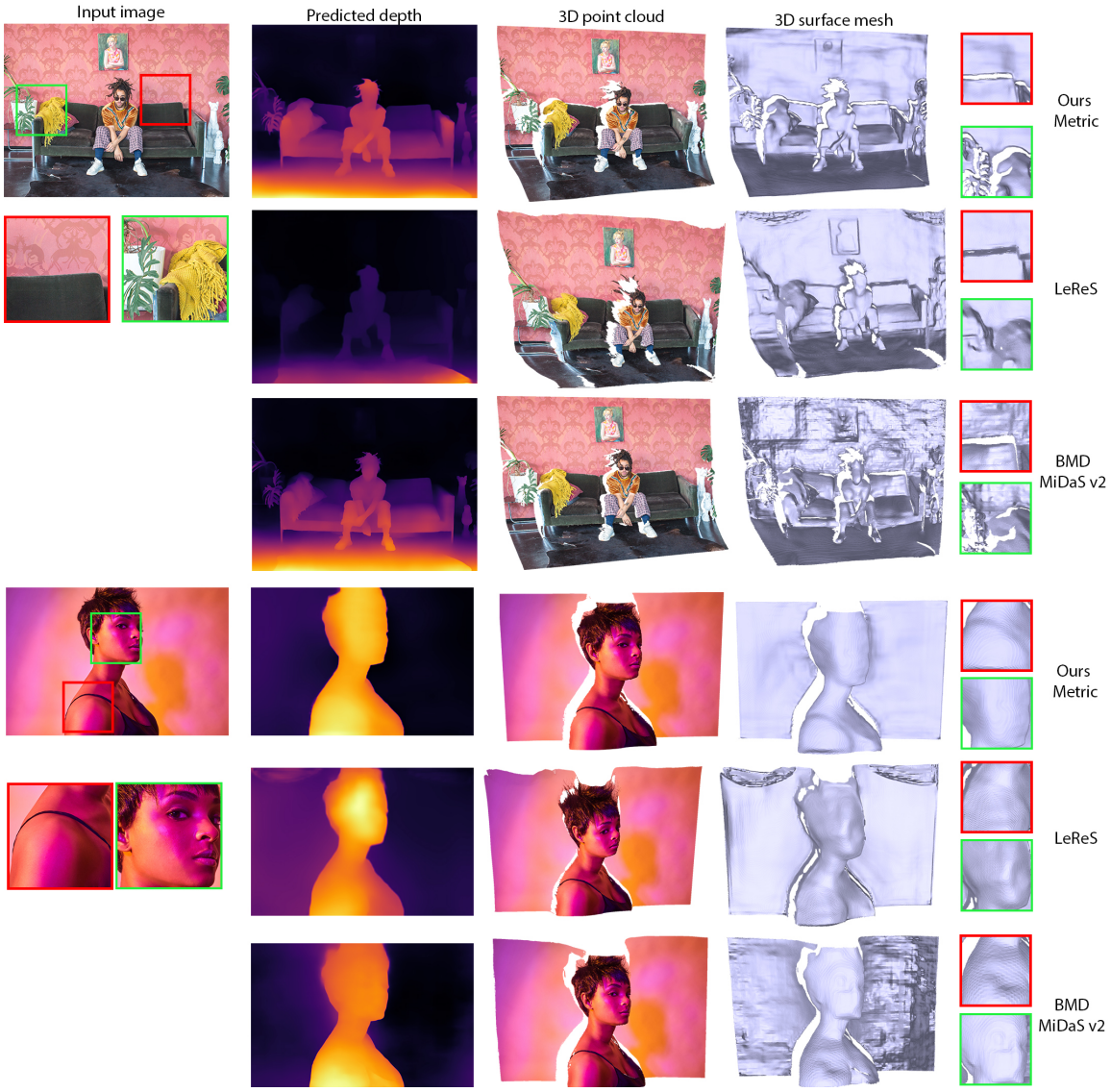

Figure 8.1: We compare the point clouds and recovered surface meshes from the high-resolution depth estimation by our metric network with LeReS [59] and Midas with boosting [32] framework for photographs in the wild containing people in diverse environments. LeReS [59] estimates metric depth but at low resolution with a consistent geometric structure. Midas with boosted depth produces high-resolution depth but with distorted scene structure. We highlight the geometric structure characteristics in red insets and the depth details in green insets. Our results contain both high-resolution details and consistent geometric structure.

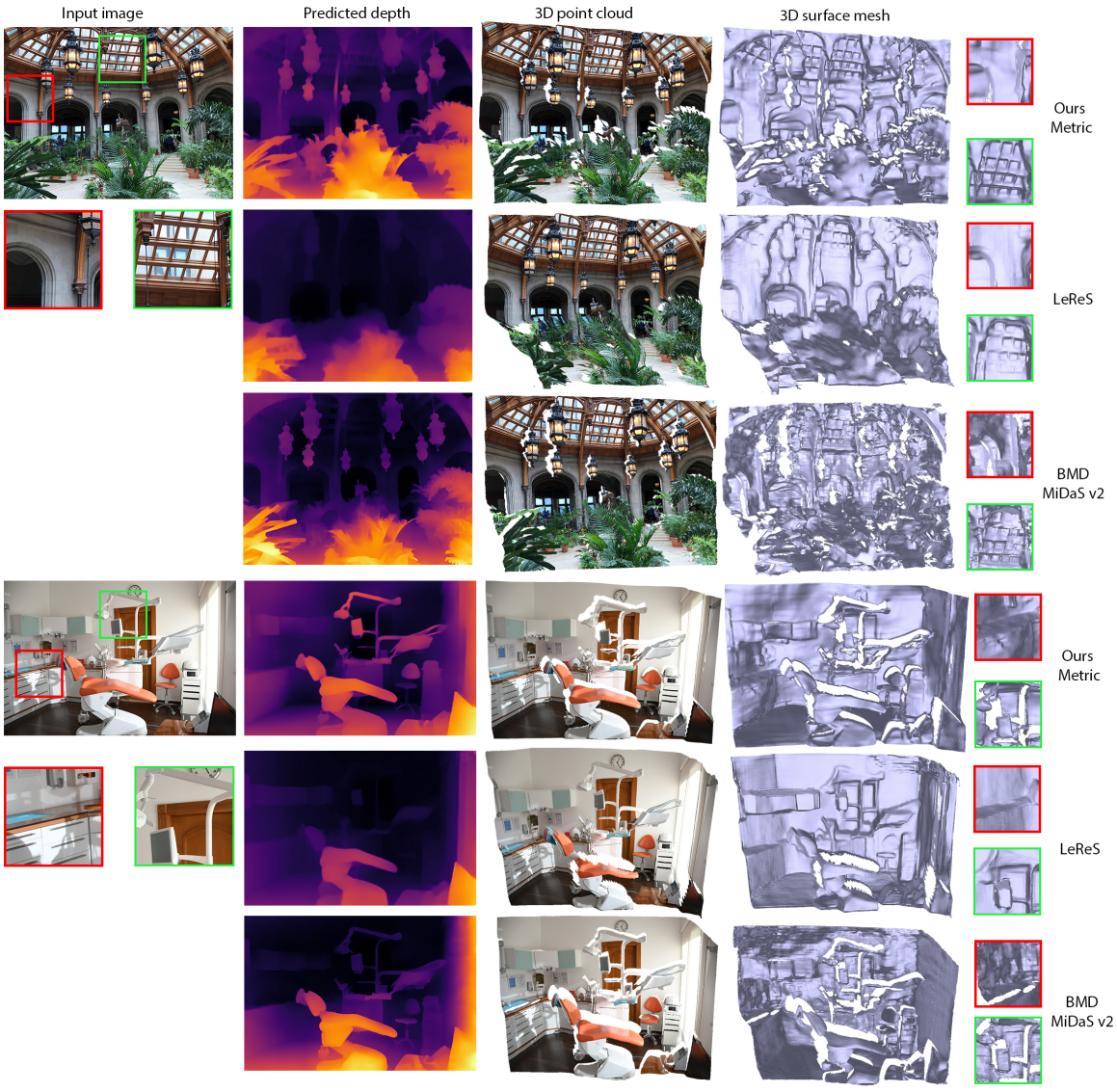

Figure 8.2: We compare the point clouds and recovered surface meshes from the high-resolution depth estimation by our metric network with LeReS [59] estimates metric depth but at low resolution with a consistent geometric structure. Midas with boosted depth produces high-resolution depth but with distorted scene structure. We highlight the geometric structure characteristics in red insets and the depth details in green insets. Our results contain both high-resolution details and consistent geometric structure.

Figure 8.3: Overview of the project point clouds from our high-resolution metric network with and without removal of the high gradient edges that create floating pixels in the 3D space.

and consistent geometric structure (such as walls and cabinets) shown in red insets. LeReS lacks high-frequency details and generates smooth depth edges, whereas BMD Midas show strong depth edges but with distorted scene structure.

Directly projecting a depth map to a 3D point cloud can create artifact 3D points around the edges that get stretched in the 3D point cloud. Furthermore, these artifacts in the point cloud can generate irregular mesh around the edges when recovering surface mesh from point clouds using NDC [5]. To overcome this issue, we create a post-processing step after depth estimation that filters out high-gradient generating edge pixels in the depth image. With this technique, there does not exist any artifact or floating pixels in the 3D point cloud around edges, leading to crisp mesh generating. We provide an overview of the qualitative results of point clouds with and without the edge pixels filtering technique in Figure 8.3. The figure shows that the depth around the dog's nose (top) has smooth edges leading to stretched 3D point clouds around the dog noise. After filtering the artifacts around the dog's nose, the result looks sharp and clean. The bottom image shows a similar observation around the helmet's outer edges.

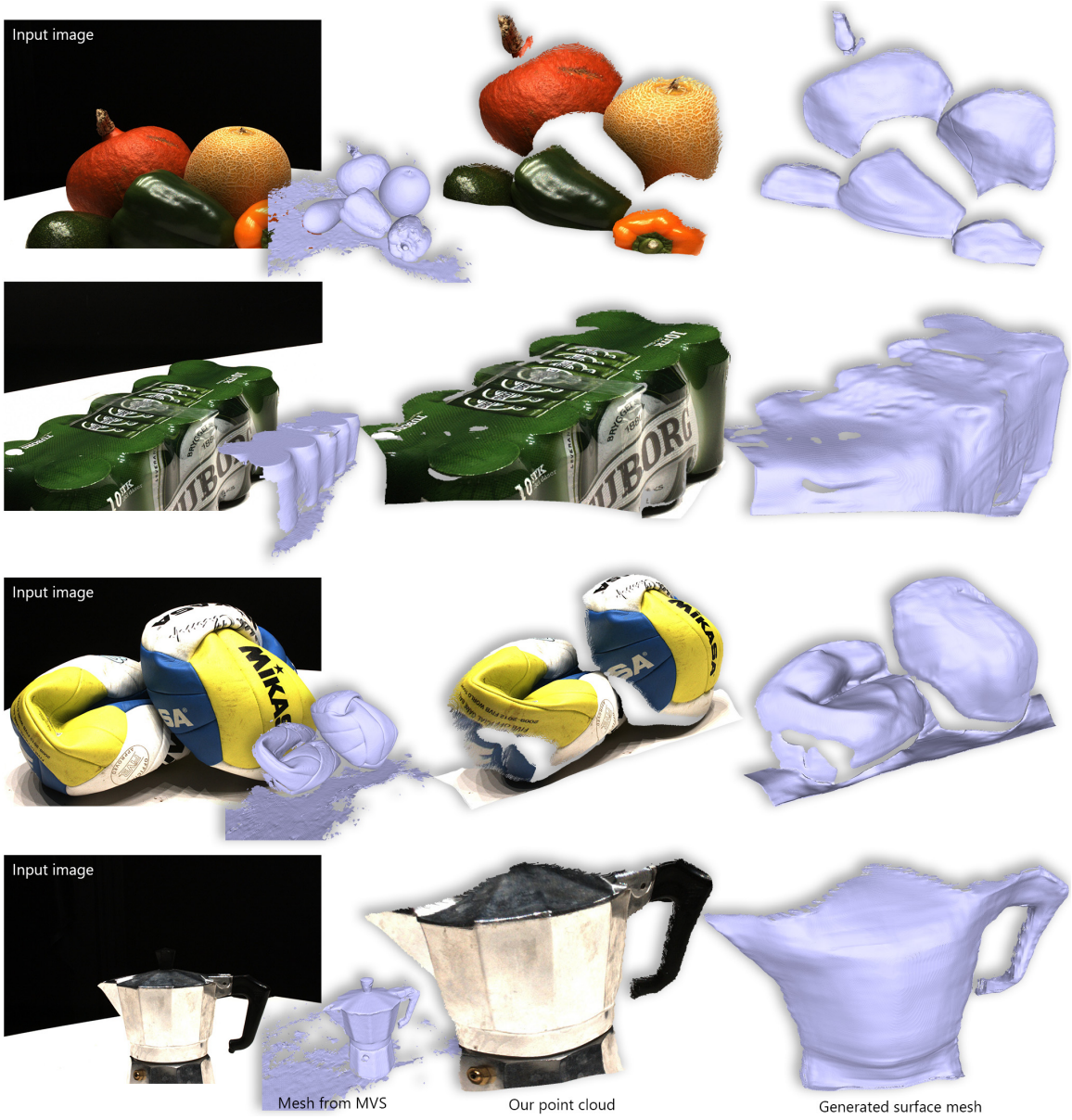

Figure 8.4: We show the surface meshes generated for close-up object images based on the 3D point clouds extracted from the depth estimations of our metric network on the DTU dataset [1].

Our high-resolution dense coherent 3D point naturally lends itself to extracting surface mesh. Specifically, we first estimate high-resolution metric depth for close-up images of objects. Then we use the meshing network from Chen *et al.* [5] to recover the surface mesh of a variety of single object images [1] from the dense 3D point cloud. In Figure 8.4, we show the qualitative results of the meshes for all the objects along with the ground-truth mesh generated by MVSTER [52], which uses a multi-view stereo setup to generate mesh from

multi-views images of the objects. For the input image in the first row, we can recover the surface mesh of all the vegetables with precise orientation. For the second row, the mesh from our point clouds captures the L-shaped structure of the pack of cans. This indicates that applying the geometric constraints to train the metric network does capture the correct surface orientation of objects in the wild. In the third row, our results capture the precise shape and orientation of the two deflated balls. Lastly, the mesh from our high-resolution metric depth-based point cloud can capture the shape of the container. Overall, despite using only a single image for each object, our monocular metric network can generate depth that captures the accurate structure and sufficient details compared to the ground-truth mesh generated from a multi-view stereo setup.

Chapter 9

Limitations

The main advantage of our ordinal depth network is estimating depth with both ordinal scene structure and high-frequency details. The depth from the ordinal network is an important aspect of depth estimation from the metric network. A vital requirement is to consider high-resolution input images to achieve good performance from our two-step framework. In this chapter, we study the effect of noise in the input image and its implications on the ordinal estimates. Additionally, we highlight another limitation of our framework concerning details in the extracted mesh compared to multi-stereo-based approaches.

9.1 Sensitivity to Image Noise

In Chapter 6, we introduce our high-resolution depth estimation network by utilizing the ordinal estimates introduced in Chapter 4 at different resolutions with the input image. Specifically, the low-resolution ordinal estimate is estimated at the receptive field size to capture the global scene structure, and the high-resolution ordinal estimate is estimated at $\mathcal{R}_{20}$ [32] to capture the high-frequency depth details. The effectiveness of the metric network

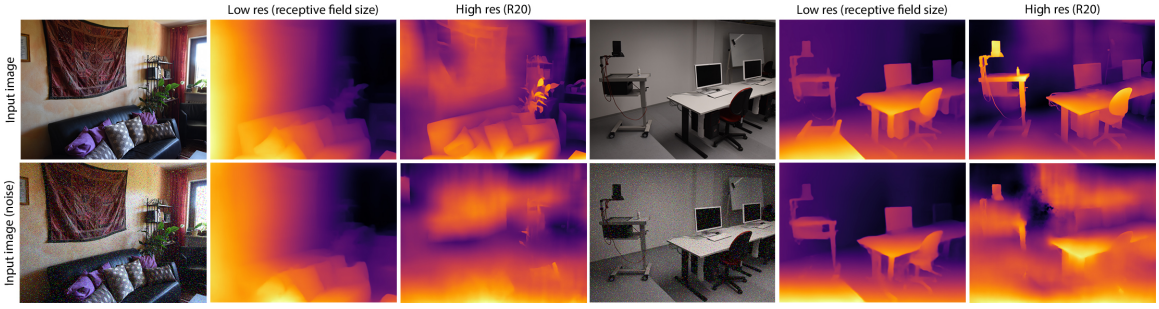

Figure 9.1: We provide an overview of a limitation of our depth estimation network on noisy input images from IBims-1 [20] dataset. Our high-resolution ordinal estimate generates low-frequency artifacts on images with highly noisy signals, as seen in the "High-res (R20)" estimate for both noisy images.

Table 9.1: We quantitatively evaluate our ordinal networks "Ours-ordinal" and "Ours-bmd" on the NYU [12] dataset with other ordinal depth networks using ordinal and edge accuracy metrics.

Methods	NYU			
	$SEE_{k3} \downarrow$	$SEE_{k5} \downarrow$	Ord. $\downarrow$	$D^3R \downarrow$
VN TPAMI	0.124	0.117	0.124	0.638
SGR	0.140	0.134	0.179	0.578
SGR-bmd	0.145	0.139	0.182	0.572
LeReS-ordinal	**0.096**	**0.089**	**0.096**	**0.436**
MDS	0.117	0.110	0.118	0.517
MDS-bmd	0.114	0.107	0.122	0.579
DPT	0.120	0.114	0.109	0.493
Ken Burns	0.106	0.100	0.082	0.459
Ours-ordinal	0.110	0.104	0.104	0.467
Ours-bmd	0.111	0.104	0.124	0.473

Table 9.2: We quantitatively compare our metric network "Ours-si" with other depth network on NYU [12] dataset using both scale-invariant depth and edge metrics.

Methods	NYU				
	RMSE $\downarrow$	Abs. $\downarrow$	$\delta_1 \uparrow$	Ord. $\downarrow$	$D^3R \downarrow$
MD	1.341	33.329	52.521	0.216	0.561
MC	0.662	17.921	70.738	0.189	0.687
VN ICCV	-	-	-	-	-
LeRes	**0.513**	**13.679**	**81.793**	**0.095**	**0.391**
Ours-si	0.594	13.706	79.841	0.134	0.524

is strongly dependent on the quality of the depth features captured in the ordinal estimates. Therefore, the quality of the ordinal depth depends on the quality of the input image. On datasets with high-resolution images with low image white noise, the estimated ordinal depth captures relevant features, particularly the high-resolution ordinal depth. However, datasets with high image noise, like NYU [12], can cause the high-resolution ordinal estimate to contain low-frequency artifacts. In Figure 9.1, we consider applying Gaussian noise to the input IBims-1 image with a similar resolution as NYU to demonstrate the effect of white noise on the quality of ordinal depth estimations. Despite this limitation of our approach, we quantitatively evaluate our ordinal depth network in Table 9.1 and our metric depth network in Table 9.2. In both tables, our networks still produce results on par with the state-of-the-art depth networks that estimate the depth at a lower resolution on NYU [12] dataset.

9.2 Limited Details in 3D Reconstructions

The goal of our approach is to estimate geometrically consistent high-resolution depth with sharp depth details. This is beneficial for generating meaningful 3D reconstructions of monocular scenes better than prior depth estimation methods. However, the depth estimation is limited to the information captured in the input image. This restricts the sharp

details captured in the extracted surface mesh compared to multi-view stereo methods [52] that leverage diverse scene information from multiple camera views as shown in Figure 8.4.

Chapter 10

Conclusion

To estimate high-resolution and detailed metric depth, we describe a framework to estimate metric depth for monocular images through ordinal depth. We decompose the problem into two steps to propagate the relevant local and global context information about the scene structure and depth details from the ordinal depth to solve metric depth estimation. The ordinal estimation can effectively provide relevant contextual information by exploiting the unique traits of sparse and dense ordinal losses in a combined setup to generate depth details and scene structure. The ordinal depth space is important in exploiting the full potential of the dense loss for ordinal estimation. The dense ordinal loss harmonizes well with the sparse relaxed ranking loss, making a more informed penalization on the depth ordering. The triplet sampling of sparse points provided sufficient information and reference for depth order, thus further improving the relaxed ranking loss. Our analysis of the different aspects of the ordinal network demonstrates that combining all the above aspects improves the quality of the ordinal depth estimation on a diverse range of images.

Estimating geometric consistent metric depth from the ordinal inputs simplifies the task of the metric network to reason about the geometry through multiple depth and surface normal-based constraints on the provided ordinal global and local contextual information. A vital component of the geometric constraint is the sparse ratio loss to fix the relative geometric information between points. In addition, the surface normal dense loss and multi-scale normal gradient loss demonstrate their effect in generating consistent geometric depth. Our analysis of the different components in the geometric losses indicates their effectiveness in improving the geometric structure of the metric depth. Further, recovering the meaningful 3D surface mesh through a dense consistent 3D point cloud from the high-resolution depth demonstrates the significance of the geometric constraints.

High-resolution metric monocular depth unlocks a variety of future directions. An exciting direction would be to explore different architectures for the metric network, as our current network has limitations when the inference resolution is higher than the training resolution. For example, a potential network to explore is an image transformer [9] that overcomes the issues related to image resolution. Furthermore, high-resolution metric depth

can be used as a geometric prior in neural rendering architectures for high-resolution 3D surface reconstruction. Likewise, the high-resolution monocular metric depth is beneficial for high-resolution video depth estimation, which might require additional temporal geometric constraints for temporal stability.

Bibliography

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *Int. J. Comput. Vis.*, 2016.

[2] Chuangrong Chen, Xiaozhi Chen, and Hui Cheng. On the over-smoothing problem of cnn based disparity estimation. In *Int. Conf. Comput. Vis.*, 2019.

[3] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *Adv. Neural Inform. Process. Syst.*, 2016.

[4] Weifeng Chen, Donglai Xiang, and Jia Deng. Surface normals in the wild. In *Int. Conf. Comput. Vis.*, 2017.

[5] Zhiqin Chen, Andrea Tagliasacchi, Thomas Funkhouser, and Hao Zhang. Neural dual contouring. *ACM Trans. Graph.*, 2022.

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021.

[7] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A large dataset of object scans. *arXiv preprint arXiv:1602.02481*, 2016.

[8] Erick Delage, Honglak Lee, and Andrew Y Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2006.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021.

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

[11] 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.

[12] 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.

[13] Alistair R Fielder and Merrick J Moseley. Does stereopsis matter in humans? 1996.

[14] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[15] Rockstar Games. Grand theft auto v. `http://www.rockstargames.com`.

[16] 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.*, 2012.

[17] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *Int. Conf. Comput. Vis.*, 2005.

[18] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. 2020.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[20] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Eur. Conf. Comput. Vis. Worksh.*, 2018.

[21] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. One shot 3d photography. *ACM Trans. Graph.*, 2020.

[22] Philipp Krähenbühl. Free supervision from video games. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[23] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Int. Conf. 3D Vision*, 2016.

[24] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

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

[26] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An open framework for photorealistic indoor scene datasets. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[27] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015.

[28] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. *AAAI*, 2021.

[29] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Eur. Conf. Comput. Vis.*, 2018.

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Int. Conf. Comput. Vis.*, 2017.

[31] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[32] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[33] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Trans. Graph.*, 2019.

[34] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 2003.

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Int. Conf. Comput. Vis.*, 2021.

[36] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[37] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Int. Conf. Comput. Vis.*, 2021.

[38] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[39] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008.

[40] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Germ. Conf. Pattern Recog.*, 2014.

[41] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[42] 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.

[43] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. Mach. Learn.*, 2019.

[45] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, 2020.

[46] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[47] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

[48] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Trans. Graph.*, 2018.

[49] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *Int. Conf. 3D Vision*, 2019.

[50] Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. Deeplens: Shallow depth of field from a single image. *ACM Trans. Graph.*, 2018.

[51] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *Int. Conf. Intel. Rob. and Syst.*, 2020.

[52] Xiaofeng Wang, Zheng Zhu, Fangbo Qin, Yun Ye, Guan Huang, Xu Chi, Yijia He, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. *arXiv preprint arXiv:2204.07346*, 2022.

[53] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[54] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[55] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[56] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[57] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Int. Conf. Comput. Vis.*, 2019.

[58] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020.

[59] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[60] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. Learning ordinal relationships for mid-level vision. In *Int. Conf. Comput. Vis.*, 2015.