Toward Explainable Monocular Depth Estimation

Seyed MohammadHossein Hosseini

A Thesis Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of

Master of Applied Science

Graduate Program in Electrical Engineering and Computer Science

York University
Toronto, Ontario

October 2026

© Seyed MohammadHossein Hosseini, 2026

Toward explainable monocular depth estimation

ABSTRACT

Modern monocular depth estimation systems achieve impressive accuracy through large-scale training, yet operate as black boxes, posing challenges for safety-critical certification and generalization to novel scenarios. This thesis quantifies the role of three principles of human depth perception: ground theory, which anchors depth on recovery of the ground surface; perceptual organization, which partitions the image into semantic regions to locally regularize estimates; and linear perspective, which exploits geometric regularities from the projection of parallel lines. By integrating these cues, we construct an explainable monocular depth estimation framework that, despite its simplicity, rivals opaque deep networks in specific contexts. We further show that fusion with our system reduces absolute relative error by up to 33.4% in indoor ground regions and by up to 10.7% overall on SYNS-Patches. These findings suggest that large-scale training does not fully encode fundamental geometric priors, and that explicit geometric modeling remains essential for robust depth inference.

Thesis supervisor: James H. Elder

Title: Professor

Acknowledgments

This research would not have been possible without the guidance and mentorship of my advisor, Prof. James Elder. His insight, encouragement, and steady support throughout this journey were instrumental in shaping this work. I also thank my colleagues and friends in the Elder Lab, and Bob in particular, for their collaboration and many thoughtful discussions. Finally, I am deeply grateful to my parents and my sister for their unwavering love, encouragement, and support throughout my academic journey.

Contents

Abstract ii

Acknowledgments iii

Table of Contents iv

List of Tables vi

List of Figures vii

1 Introduction **1**

1.1 Theoretical Foundations: The Principles of Depth 2

1.1.1 Ground Theory 2

1.1.2 Perceptual Organization and Semantics 3

1.1.3 Linear Perspective and Geometry 3

1.2 Gaps in the Current MDE Literature 4

1.3 Our Objective: An Explainable Reference System 5

2 Literature Review **8**

2.1 Evolution of Monocular Depth Estimation 8

2.2 Classical Foundations and Geometric Priors 8

2.3 The Deep Learning Era: Supervised and Self-Supervised MDE 9

2.4 Zero-Shot Metric Depth Estimation 10

2.5 Integrating Semantics and Ground Geometry 11

2.6 Datasets 12

3 Methodology **13**

3.1 Semantic Segmentation 14

3.1.1 Indoor/Outdoor Classification 14

3.2 Geometric Ground Anchoring 16

3.2.1 Metric Depth Recovery 17

3.2.2 Depth Constraint 20

3.3 Depth Propagation and Structural Regularity 21

3.3.1 Segments Abutting the Ground 21

3.3.2 Segments at the Bottom of the Image 21

3.3.3 Linear Perspective 21

3.4 Dense Depth Completion 24

4 Experiments **25**

4.1 Implementation Details 25

4.1.1 Datasets and Metrics 26

4.2 Metric Depth Estimation 28

4.3 Qualitative Evaluation 31

4.4 Fusion 31

4.5 Ablations 36

4.5.1 Ground Plane Modeling 36

4.5.2 Linear Perspective 36

4.6 Failure Modes 37

4.7 Conclusion 38

5 Discussion and Conclusion **39**

5.1 Key Findings and Implications 40

5.2 Limitations and Future Work 43

5.3 Conclusion 44

References **45**

List of Tables

2.1 Datasets for monocular depth estimation 12

4.1 Camera parameters on the evaluation splits 27

4.2 Baseline training data and camera inputs 28

4.3 Quantitative results on outdoor scenes 29

4.4 Quantitative results on indoor scenes 30

4.5 Relative error reductions from model fusion 34

4.6 Ground-plane modeling ablation 37

4.7 Linear-perspective ablation 37

List of Figures

1.1 Linear perspective as a depth cue 4

3.1 Overview of the depth-estimation pipeline 14
3.2 Semantic segmentation and ground-mask extraction 15
3.3 Camera and world coordinate frames 16
3.4 Horizon prior and likelihood distributions 19
3.5 Horizon estimation from ground and sky masks 20
3.6 Depth propagation from ground to adjacent surfaces 22
3.7 Line constraints and dense depth completion 23

4.1 Qualitative depth comparison on SYNS-Patches 32
4.2 Depth errors and reconstructed scene geometry 33
4.3 Qualitative results on KITTI and VKITTI2 34
4.4 Correlations between model depth errors 35

5.1 Failure cases in a forest scene 42

Chapter 1

Introduction

The human visual system relies on a diversity of perceptual principles and statistical regularities to recover depth from a single image. These include knowledge of intrinsic parameters, priors over extrinsic parameters (such as gaze angle and height relative to the ground), texture gradients, linear perspective, and occlusion. For all these principles, there are well-established mathematical theories, yet fully elaborating their statistical details remains an active area of research.

In contrast, in the realm of computer vision, the problem of recovering depth from a single image, termed *monocular depth estimation* (MDE), has evolved differently. This thesis focuses on monocular metric depth estimation, where depth is expressed in physical units. While early approaches recognized the importance of geometric reasoning [1], modern state-of-the-art methods rely primarily on self-supervised and supervised learning over massive datasets [2–5]. These systems are impressive, yet they operate as "black boxes": Their knowledge is implicit, distributed over millions of network weights, rather than grounded in explicit geometric principles.

This dichotomy leads us to the central problem of this thesis: *How can we bridge the gap between the explainable geometric principles of human vision and the performance of modern*

computational systems? Specifically, we investigate how explicit modeling of ground theory, perceptual organization, and linear perspective can create a robust, explainable reference system for metric depth estimation.

The long-term goal is to develop reliable metric depth estimation systems that make effective use of explicit geometric knowledge while remaining applicable across diverse scenes. The reference system developed here provides a basis for investigating which aspects of scene depth can be recovered from a compact set of perceptual and geometric principles, and where additional information is needed. This investigation can guide both the development of a more capable stand-alone depth estimator and the integration of explicit geometric constraints with learned depth models.

1.1 Theoretical Foundations: The Principles of Depth

To understand how to improve computational models, we must first look at the principles of human perception. While researchers have traditionally cataloged various "depth cues" [6], this thesis organizes them under three governing principles that bridge cognitive science and computer vision: Ground Theory, Perceptual Organization, and Linear Perspective.

1.1.1 Ground Theory

Gibson [7] argued that human space perception is anchored on the ground surface. This *ground theory* has been hugely influential in the human cognition literature and substantiated through many empirical studies (e.g., [8]). Here we explore how ground theory can be incorporated into AI systems for monocular depth estimation.

1.1.2 Perceptual Organization and Semantics

While the importance of perceptual organization in monocular depth estimation was recognized in seminal contributions to the computer vision literature (e.g., [1]), it has become implicit in modern systems.

In these early papers, relatively low-level methods such as super-pixels [9] were used to chunk the image into coherent pieces over which stable estimates could be made. Here we take a different approach based on recent work in the cognitive literature [10]. While the conventional view [11–13] has been that extraction of 3D spatial attributes of a scene occurs prior to and thus informs inference of semantic information, this recent work suggests that the opposite can in fact often be true: Some key semantic attributes of a scene are extracted early and can thus inform the inference of 3D spatial attributes. Inspired by this result, we examine the utility of state-of-the-art semantic segmentation methods to organize the image into semantically labelled disjoint chunks over which inferences can be made. Importantly, we show how semantic segmentation can provide the critical boundary conditions needed to meaningfully propagate sparse depth information to recover dense depth maps.

1.1.3 Linear Perspective and Geometry

The discovery of the laws of linear perspective is often hailed as a defining achievement of the early Renaissance, attesting to the importance of linear perspective in determining human perception of depth from a single image (Figure 1.1). In the computer vision literature, linear perspective has played an important role in online camera calibration algorithms [14–17] and in monocular 3D inference in constrained scenarios [18], but its role more generally in monocular depth estimation has not been explored. Here we study how linear perspective can be integrated with ground theory and perceptual organization to inform monocular depth perception.



Figure 1.1: Converging railway tracks illustrate linear perspective as a depth cue: parallel lines in the scene project toward a vanishing point in the image [19].

1.2 Gaps in the Current MDE Literature

The importance of machine learning in solving MDE was recognized twenty years ago, and accuracy has tracked advances in deep learning since [20, 21]. Current methods, such as Depth Anything [2] or Metric3D [4], excel at capturing fine details. However, they face significant challenges:

1. **Implicit Knowledge:** The geometric logic of these models is hidden. We do not know why a model predicts a certain depth.

2. **Generalization and Scale:** Zero-shot generalization remains difficult due to shifts in intrinsic and extrinsic camera parameters. State-of-the-art methods often struggle to maintain metric scale accuracy in unseen environments [22].

3. **Lack of Explicit Grounding:** While some recent work has investigated ground regularities [23, 24], a systematic framework for propagating ground constraints across

complex scenes using semantic boundaries remains underexplored.

4. **Human–AI Perception Gap:** There remains a large gap in our understanding of how MDE in the human brain relates to MDE in modern AI. Our objective is to investigate whether principles associated with human perception provide useful constraints for computational models.

1.3 Our Objective: An Explainable Reference System

The goal of this thesis is to take a step toward bridging these gaps. We propose a novel hybrid framework, an explainable reference system, that explicitly integrates semantic segmentation with geometric reasoning.

Here, *explainability* refers to the ability to trace depth estimates to identifiable image evidence, explicit modeling assumptions, and specified inference rules. Our system relies on a deep network for semantic segmentation, and we do not attempt to explain this network's predictions. We address explainability in the subsequent depth-inference process, conditioned on the semantic segmentation it provides.

Instead of relying solely on end-to-end learning, we aim to simplify monocular depth estimation by reducing it to precise ground plane modeling and propagation. Our approach is built on three contributions corresponding to the principles outlined above:

- **Ground Theory Implementation:** We use semantic segmentation to identify the projection of the ground surface. Using knowledge of camera parameters and an estimate of the horizon, we infer the metric depth of all pixels projecting from this surface, resolving scale ambiguity.

- **Perceptual Organization via Semantics:** We treat semantic segments as disjoint chunks over which inferences can be made. Consistent with ground theory, we fix

boundary conditions for segments abutting the ground surface. We then apply a fast multi-grid Laplace completion method [25] to propagate these sparse estimates to all pixels within each semantic segment.

- **Linear Perspective Integration:** We extract line segments and, under the Manhattan assumption, use them to infer 3D orientations. These sparse depth estimates are fused with the ground model to inform the dense depth map.

By aligning with ecological theories of human visual perception, our approach provides a principled foundation for robust, metrically accurate monocular depth estimation.

Assumptions and Required Inputs: Our reference system operates under the following input requirements and modeling assumptions:

- **Known camera parameters:** Camera intrinsics and extrinsics, including camera orientation and height above the ground, are assumed known.
- **Planar ground:** The ground surface is approximated by a single plane. For indoor scenes, this plane is assumed horizontal; for outdoor scenes, it may be tilted.
- **Vertical ground-contact surfaces:** Surfaces directly abutting the ground are approximated as vertical, allowing depth estimates to be propagated upward from their ground-contact boundaries.

The consequences of violations of these assumptions are examined in Chapter 5.

Thesis Structure: The remainder of this thesis is organized as follows: Chapter 2 reviews related work in monocular depth estimation, identifying how current "black box" methods diverge from classical geometric approaches. Chapter 3 details the proposed method, establishing the mathematical formulation for ground plane modeling and Laplace depth propagation. Chapter 4 presents experimental results, evaluating how close this explainable system

comes to state-of-the-art performance and its ability to generalize. Chapter 5 concludes with a discussion on fusing this reference system with deep learning for future hybrid models.

Chapter 2

Literature Review

2.1 Evolution of Monocular Depth Estimation

Monocular Depth Estimation (MDE) has evolved from classical methods relying on hand-crafted features and geometric priors to modern deep learning approaches. In computer vision, the problem of recovering depth from a single image is fundamentally ill-posed, as a single 2D projection can correspond to infinitely many 3D configurations. Consequently, the field's history is a progression of how researchers have modeled the priors necessary to resolve this ambiguity.

2.2 Classical Foundations and Geometric Priors

Before the dominance of deep learning, monocular depth estimation relied on interpretable geometric modeling. Early work by Hoiem et al. [1] and Delage et al. [26] followed a cue-based approach, incorporating simplifying assumptions about scene structure, such as flat ground planes and vertical surfaces.

The *Make3D* algorithm [27] represents a cornerstone of explainable monocular depth

estimation. It segments an image into superpixels and assumes each corresponds to a planar surface. Global consistency is enforced through a probabilistic graphical model that captures relationships between neighboring regions, encouraging coherent surface geometry across the scene.

Despite their interpretability and strong geometric grounding, these methods were limited by their reliance on handcrafted features and rigid planar assumptions, which restricted their ability to generalize to diverse and complex real-world environments.

2.3 The Deep Learning Era: Supervised and Self-Supervised MDE

Architectural Innovations

The establishment of larger ground-truthed datasets and deep learning models (e.g., [28]) led to rapid improvements in accuracy. These improvements were driven more by machine learning innovations than computer vision insights. These innovations include the introduction of Transformer architectures [29, 30] for improved feature representation, conditional Random Field optimization [31, 32] for improved depth map consistency, and binning strategies [33, 34] to reframe depth estimation as a classification problem, leading to improved accuracy. Diffusion-based models like Marigold [20] and GRIN [21] have also recently shown promise, using powerful generative priors to produce detailed outputs.

A smaller recent line of work has also explored internal interpretability in depth transformers. For example, the authors of [35] introduce SVD-inspired attention (SVDA) into a DPT-style architecture, making attention behavior measurable through spectral indicators such as entropy, rank, sparsity, and alignment. This remains a model-internal notion of interpretability, rather than an explicit geometric account of depth inference.

Self-Supervised Learning

A key shift in MDE has been the move toward self-supervision, which leverages geometric constraints from stereo pairs [36, 37] or monocular video sequences [38, 39] as a supervisory signal. By massively increasing the size and diversity of training datasets, self-supervision has led to dramatic improvements in accuracy and generalization capabilities. The core objective is to minimize the photometric reprojection error between a target image and a source image warped using the predicted depth and camera pose. To improve robustness, edge-aware smoothness losses [37], auto-masking of occluded or static pixels [39], and advanced network architectures [40, 41] are commonly employed.

2.4 Zero-Shot Metric Depth Estimation

A dominant current trend is the development of models that infer metric scale from image content and large-scale data priors in zero-shot settings (environments not seen during training).

Foundational works like *MiDaS* [42] established the effectiveness of massive data mixing, while *ZoeDepth* [43] introduced separate prediction heads for indoor and outdoor scales. *Depth Anything* [2] further leverages unlabeled and synthetic data in a teacher-student framework to improve robustness. *Metric3D* [4] introduces a canonical camera transformation and a joint depth–normal optimization module to enable accurate zero-shot metric depth estimation. More recently, universal models like *UniDepth* [5] and *DepthPro* [44] attempt to infer metric depth without known camera intrinsics by explicitly estimating focal length or using spherical representations.

A common thread among these state-of-the-art methods is that metric scale is treated as an implicit property to be learned from data. These "black-box" recovery mechanisms can degrade unpredictably under domain shifts because they lack an explicit, scene-independent

geometric reference, a gap this thesis aims to address.

2.5 Integrating Semantics and Ground Geometry

Surprisingly, very few modern deep learning MDE methods make explicit use of geometric principles such as ground theory. One important exception is *GEDepth* [23], which assumes knowledge of camera parameters and models the ground as planar but potentially pitched up or down along the optic axis. This formulation provides an explicit geometric constraint that helps resolve scale ambiguity, allowing the network to focus on estimating the residual depth relative to this grounded reference.

Other researchers have explored integrating semantic context to refine depth estimation. Jung et al. [45] introduced a cross-task multi-embedding attention module to improve object boundary accuracy by aligning depth and semantic features. Shao et al. [32] modeled scenes as piecewise planar regions using surface normal and geometric constraints. However, systematic methods for propagating ground-based geometric constraints across complex, semantically structured scenes remain underexplored. This thesis builds on these classical geometric principles to develop a more explainable and metrically accurate framework.

2.6 Datasets

The development of MDE relies on diverse datasets (see Table 2.1). This work primarily utilizes the SYNS-Patches dataset [46]. Unlike the standard KITTI [47] or NYU [48] benchmarks, SYNS-Patches provides high-diversity scenes (natural, residential, industrial) with precise LiDAR ground truth and well-calibrated camera parameters.

Table 2.1: Common datasets used for monocular depth estimation, in decreasing order of size (number of images).

Dataset	**Scenes**	**Source**	**Size**
HM3D [49]	Indoor	Real-world	∼2,000K
Mapillary PSD [50]	Outdoor	Real-world	750K
DrivingStereo [51]	Outdoor	Real-world	∼181K
Cityscapes [52]	Outdoor	Real-world	∼170K
DIML [53]	Outdoor	Real-world	∼122K
DDAD [40]	Outdoor	Real-world	∼80K
Hypersim [54]	Indoor	Synthetic	54K
Virtual KITTI [55]	Outdoor	Synthetic	37K
nuScenes [56]	Outdoor	Real-world	10K
SYNS [57]	Indoor/Outdoor	Real-world	1.6K
DIODE [58]	Indoor/Outdoor	Real-world	771
NYU [48]	Indoor	Real-world	654
KITTI [47]	Outdoor	Real-world	652
ETH3D [59]	Outdoor	Real-world	431
iBims-1 [60]	Indoor	Real-world	100

Chapter 3

Methodology

In this chapter, we present a framework for estimating metric depth from a single image by anchoring scene scale to the geometry of the ground plane. Unlike end-to-end regression approaches that operate as black boxes, our method decomposes the scene into a set of interpretable geometric reference systems. This design is inspired by human perception, where the ground provides a fundamental reference for interpreting three-dimensional structure.

Figure 3.1 illustrates the proposed pipeline. (1) First, a state-of-the-art semantic segmentation network identifies ground pixels. (2) Their depth is then recovered either by assuming a planar ground and estimating its geometry from the horizon and camera parameters, or by learning expected depth as a function of image height within the ground region. (3a) Next, objects in contact with the ground inherit depth at their points of contact, establishing additional anchors. (3b) In parallel, linear perspective cues are used to estimate the orientation of line segments in 3D. (4) Finally, a dense depth map is obtained through Laplace completion, while preserving semantic boundaries.

Each component is described in the following sections.

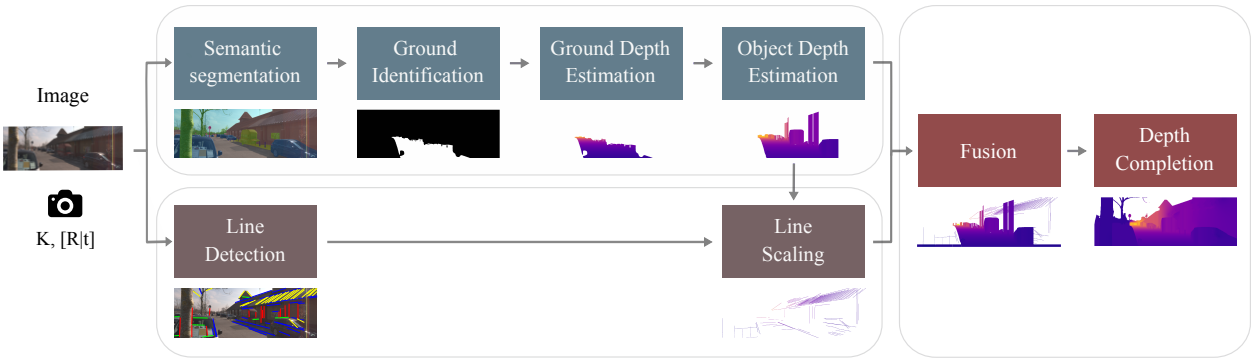


Figure 3.1: Our explainable reference system for monocular metric depth estimation. The process includes semantic segmentation to extract the ground mask, geometric ground depth calculation, depth propagation to adjacent objects, 3D line segment integration, and Laplace depth completion.

3.1 Semantic Segmentation

We generate a semantic segmentation map using the InternImage model [61], pre-trained on the ADE20K dataset [62]. Its diversity across natural and urban environments makes it well suited for our task. We define a subset of semantic classes as ground: *floor, road, runway, sidewalk, track, path, sand, land, grass, field*. From the segmentation output, we construct a binary ground mask by taking the union of pixels belonging to these classes (Figure 3.2).

3.1.1 Indoor/Outdoor Classification

Ground depth estimation differs significantly between indoor and outdoor scenes. We approximate the ground for indoor environments as horizontal, simplifying depth calculation. For indoor scenes, we assume a horizontal ground plane and express it in camera coordinates using the known camera orientation; no horizon estimation is performed. This approximation is appropriate for level floors but does not represent ramps, floor regions at different heights, or other departures from a single horizontal plane. Outdoor terrains, however, are more variable, and we model these as potentially non-horizontal planes. To handle this difference, we employ a scene classifier based on the semantic segmentation map. We compute



(a) Input RGB image from our SYNS Patches dataset.



(b) Semantic segmentation map (InternImage [61])



(c) Extracted ground plane region.

Figure 3.2: Semantic segmentation and ground-mask extraction for a SYNS-Patches image. (a) RGB input. (b) InternImage semantic labels. (c) Binary mask formed from the classes treated as ground. The mask identifies the pixels used for ground-depth estimation.

a feature vector $\mathbf{x} = [x_1, \ldots, x_k]$, where x_i is the proportion of pixels assigned to the i-th semantic class. We use ground-truthed data to determine maximum likelihood estimates of the class-conditional multinomial distributions $p(\mathbf{x}|C = i)$ and $p(\mathbf{x}|C = o)$ for indoor and outdoor scenes, respectively, assuming conditional independence between pixels. We then

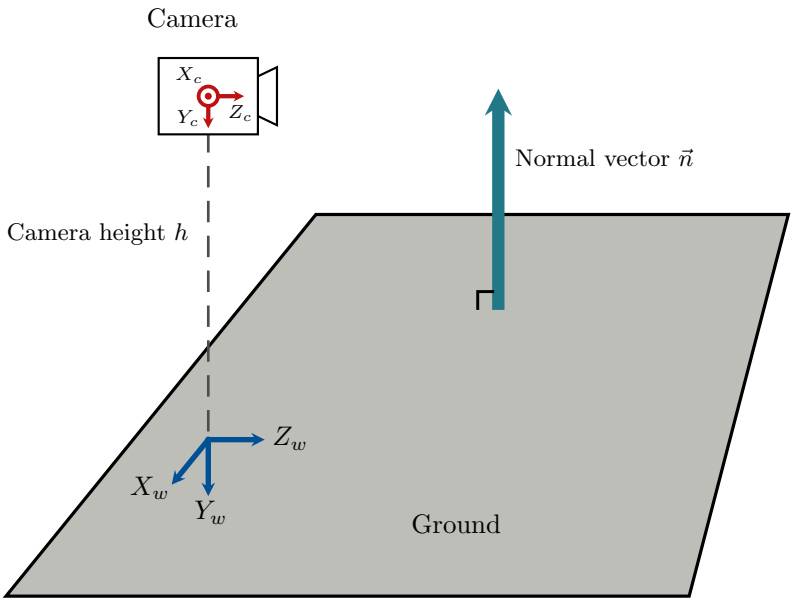


Figure 3.3: Viewing geometry. The camera coordinate frame (X_c, Y_c, Z_c) (red) is centered at the camera with Z_c aligned with the optical axis, while the world coordinate frame (X_w, Y_w, Z_w) (blue) is centered directly below the camera, on the ground plane. Here h denotes the vertical camera height relative to the ground, and $\vec{n}$ is the ground-plane unit normal.

classify according to maximum likelihood.

$$\log p(\mathbf{x}|C) \propto \sum_{i=1}^{k} (Nx_i) \log \left(p\left(x_i|C\right) \right). \tag{3.1}$$

3.2 Geometric Ground Anchoring

Assuming a calibrated camera with known intrinsics $\mathbf{K}$ and extrinsics $[\mathbf{R} \mid \mathbf{t}]$, a world point $\bar{\mathbf{x}}_w = [x_w, y_w, z_w, 1]^\top$ projects to image coordinates $\mathbf{u} = [u, v, 1]$ as:

$$\lambda \mathbf{u} = \mathbf{K}[\mathbf{R} \mid \mathbf{t}]\bar{\mathbf{x}}_w. \tag{3.2}$$

Figure 3.3 illustrates the geometry. The camera coordinate frame (X_c, Y_c, Z_c) is attached

to the optical center, with Z_c pointing along the optical axis. The world coordinate frame (X_w, Y_w, Z_w) is centered on the ground directly below the camera, with the vertical axis aligned with gravity. The ground plane normal is denoted $\vec{n}$, while h denotes the vertical camera height above the ground. Together with the world-to-camera rotation $\mathbf{R}$ and translation $\mathbf{t}$, these quantities determine the mapping between camera and world coordinates used throughout this chapter.

3.2.1 Metric Depth Recovery

We model the ground as a plane defined by a surface normal $\mathbf{n} \in \mathbb{R}^3$ and a scalar offset d. The plane equation is:

$$\mathbf{n}^\top \mathbf{x}_w = d, \tag{3.3}$$

where $\mathbf{x}_w$ denotes the corresponding point in the scene.

Given a pixel $\mathbf{u} = [u, v, 1]^\top$ in homogeneous coordinates, we seek to recover the metric world depth z_w of the corresponding point in the scene. A ray originating from the camera center and passing through $\mathbf{u}$ can be expressed in camera coordinates as:

$$\mathbf{x}_c = z_c \mathbf{K}^{-1} \mathbf{u}, \tag{3.4}$$

where z_c is the unknown depth along the optical axis.

To find the intersection of this ray with the ground plane, we transform the ray into world coordinates. Using the relationship $\mathbf{x}_c = \mathbf{R}\mathbf{x}_w + \mathbf{t}$, we obtain:

$$\mathbf{x}_w = \mathbf{R}^\top (\mathbf{x}_c - \mathbf{t}) = z_c \mathbf{R}^\top \mathbf{K}^{-1} \mathbf{u} - \mathbf{R}^\top \mathbf{t}. \tag{3.5}$$

For compactness, we define $\mathbf{A} = (\mathbf{K}\mathbf{R})^{-1} = \mathbf{R}^\top \mathbf{K}^{-1}, \quad \mathbf{b} = -\mathbf{R}^\top \mathbf{t}$, where $\mathbf{b}$ corresponds to the camera center in world coordinates.

Substituting Eq. (3.5) into the plane equation Eq. (3.3) and solving for z_c gives:

$$\mathbf{n}^\top(z_c\mathbf{A}\mathbf{u} + \mathbf{b}) = d \quad \Longrightarrow \quad z_c = \frac{d - \mathbf{n}^\top\mathbf{b}}{\mathbf{n}^\top\mathbf{A}\mathbf{u}}. \tag{3.6}$$

The metric world depth z_w corresponds to the third component of $\mathbf{x}_w$. Substituting z_c into Eq. (3.5) yields:

$$z_w = z_c(\mathbf{A}\mathbf{u})_3 + \mathbf{b}_3, \tag{3.7}$$

where $(\cdot)_3$ denotes the third component.

In our setup, the camera height h is known. This assumption is practical for calibrated installations in which camera height above the ground can be measured and remains stable. It is more restrictive for handheld cameras. Geometrically, the quantity $d - \mathbf{n}^\top\mathbf{b}$ represents the perpendicular distance from the camera center to the plane, which equals the camera height h. Substituting this into Eq. (3.6), we obtain the expression:

$$z_w = \frac{h\,(\mathbf{A}\mathbf{u})_3}{\mathbf{n}^\top\mathbf{A}\mathbf{u}} + \mathbf{b}_3. \tag{3.8}$$

If the camera parameters and ground orientation are known, Eq. (3.8) provides a closed-form solution for the depth z_w of every pixel classified as ground, serving as the metric anchor for subsequent depth propagation. Indoors, the ground is assumed horizontal; outdoors, its orientation is recovered from the horizon line, as described below.

Estimating the Horizon Line

Since the camera height is known, the horizon line uniquely determines the ground plane. To estimate the horizon we employ a Bayesian inference framework based on the regions of the image identified as ground and sky. In particular, we initially estimate the vertical position of the horizon y_h in each image column. From training data, we derive Gaussian

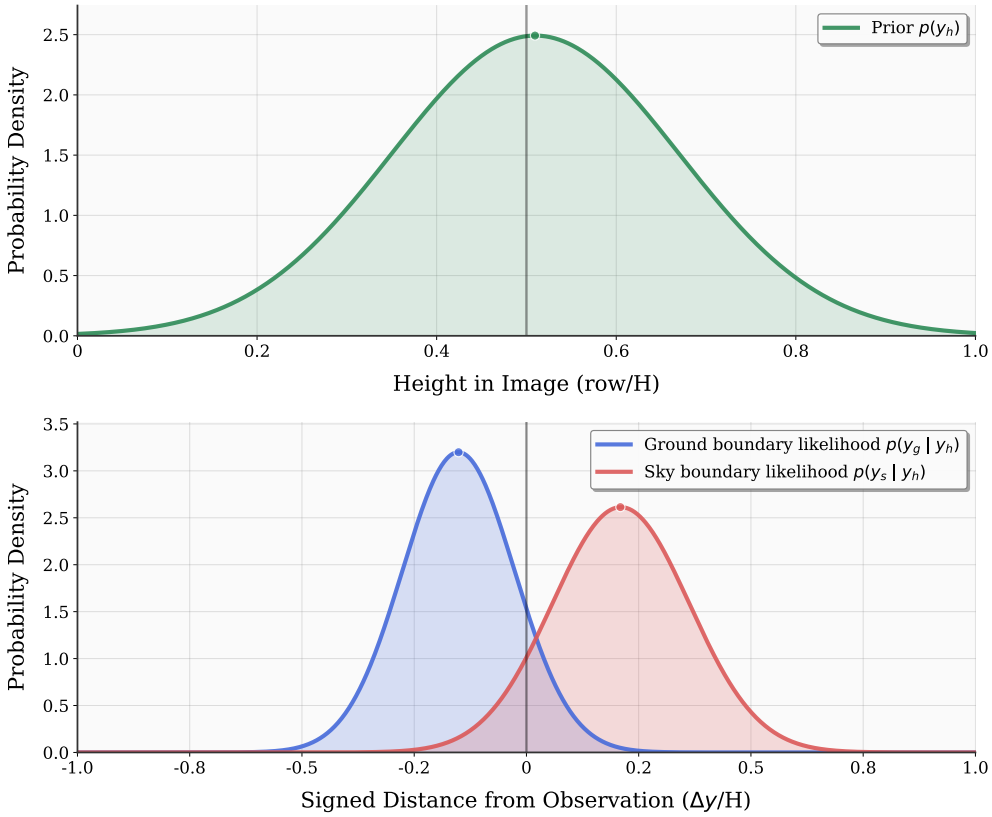


Figure 3.4: Distributions used for horizon estimation. Top: Gaussian prior over the horizon's vertical image position. Bottom: Gaussian likelihoods based on the ground and sky boundaries. Positions and offsets are normalized by image height.

approximations for the prior distribution $p\,(y_h)$ and the likelihood distributions $p\,(y_g|y_h)$ and $p\,(y_s|y_h)$, where y_g and y_s are the vertical positions of the highest ground pixel and lowest sky pixel in the column, respectively. Assuming shift invariance, these conditional distributions reduce to $p\,(y_g - y_h)$ and $p\,(y_s - y_h)$ (Fig. 3.4). Observing y_g and y_s shifts these likelihood functions and the posterior is determined by multiplying these shifted likelihoods with the prior.

This yields a Gaussian posterior for the horizon height y_{hi} in each column i, characterized by a mean μ_i and variance σ_i^2. We determine the horizon line as the maximum likelihood line passing through all columns, based on a weighted linear least-squares fit to the local estimates, weighing the residual in each column by its posterior precision σ_i^{-2} (Figure 3.5).

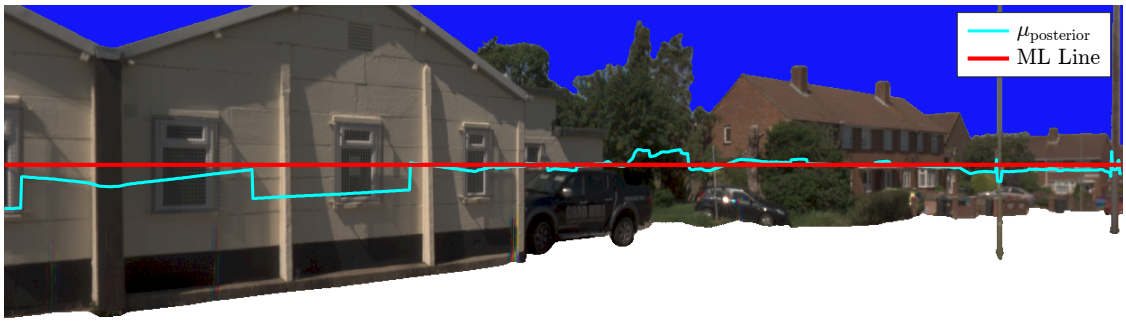


Figure 3.5: Horizon estimation from semantic masks. White and blue indicate ground and sky regions, respectively. The cyan curve shows the posterior mean of the horizon position in each column; the red line is the fitted maximum-likelihood horizon.

Determining the Ground Plane Normal from the Horizon Line

From the estimated 2D horizon line, $a_h u + b_h v + c_h = 0$, we recover the 3D orientation of the ground plane. Identifying the 3D ray directions $\mathbf{v}_1$ and $\mathbf{v}_2$ for two distinct points on the horizon line, the ground-plane unit normal is simply $\mathbf{n} = \frac{\mathbf{v}_1 \times \mathbf{v}_2}{\|\mathbf{v}_1 \times \mathbf{v}_2\|}$ and metric depth for all ground pixels can be determined from Eqn. (3.8).

3.2.2 Depth Constraint

We note that ground truth is typically based upon LiDAR, with a maximum range of $\sim$ 150m. Our ground plane depth estimates, on the other hand, extend to ∞ as we approach the horizon. Given imperfections in the semantic segmentation and uncertainty in the horizon estimation, this can lead to catastrophic depth errors. To avoid this problem, we learn from training data a point estimate of depth $z(y)$ as a function of height y in the image, and revert to this point estimate for the entire ground surface if the geometric estimate for any ground pixel exceeds 150m.

3.3 Depth Propagation and Structural Regularity

3.3.1 Segments Abutting the Ground

As per ground theory, pixels within non-ground segments adjacent to ground pixels inherit the depth of those ground pixels (Figure 3.6b). In our current reference system we make the assumption that these surfaces are vertical and propagate these depths through each of these columns within the abutting segment.

$$z_{obj}(u, v) \leftarrow z_{\texttt{ground}}(u, v_g). \tag{3.9}$$

where z represents the depth map, (u, v) are the pixel coordinates of a point on the non-ground object and (u, v_g) denotes the coordinates of the abutting ground points. The resulting depth map is illustrated in Figure 3.6c.

3.3.2 Segments at the Bottom of the Image

Objects extending to the bottom edge of the image lack visible ground contact, preventing propagation of ground depths. To resolve this, we learn class-specific estimates of the expected depth when observed at the lower image boundary and assign this as a boundary condition to these boundary pixels.

3.3.3 Linear Perspective

Lines detected in the image are likely to project from lines embedded in the 3D scene. Moreover, these 3D lines very often occur in orthogonal groups aligned with a Manhattan frame [14]. To impose structural regularity on the scene geometry, we incorporate 3D line segments as explicit constraints. We use state-of-the-art methods to detect these segments [63] and estimate their Manhattan frame [17]. If the Manhattan frame is determined to be reli-



(a) Estimated ground depth



(b) Non-ground segments abutting the ground.

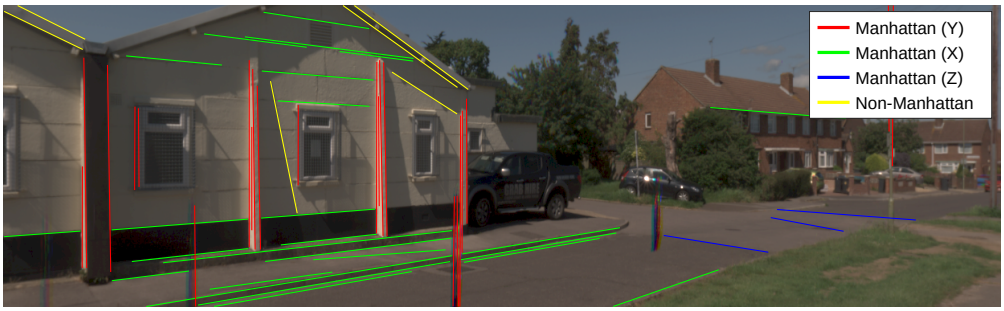


(c) Propagated depth for abutting regions.

Figure 3.6: Depth propagation from the ground to adjacent surfaces. (a) Estimated ground depth. (b) Non-ground segments that abut the ground. (c) Depth propagated along image columns from ground-contact boundaries under the vertical-surface assumption. Colors in (a) and (c) indicate depth in meters.

able, each line segment is assigned a 3D Manhattan direction or classified as non-Manhattan (Figure 3.7a).

Reconstruction of Manhattan lines. When the estimated Manhattan frame is reliable, each segment classified as Manhattan is assigned one of the frame's three mutually



(a) Detected Manhattan and non-Manhattan lines.



(b) Depth map with fused 3D line segments.



(c) Final dense depth map after completion.

Figure 3.7: Line constraints and dense depth completion. (a) Detected segments, colored by Manhattan direction or classified as non-Manhattan. (b) Depth estimates after integrating reconstructed 3D lines with the existing depth map. (c) Dense depth after Laplace completion within semantic regions.

orthogonal 3D directions. Together, the line segment and its assigned direction determine the corresponding 3D line up to scale. We resolve this remaining ambiguity using existing depth estimates within a distance of $\sqrt{2}$ pixels from the line segment in the image, thereby anchoring the reconstruction metrically.

Reconstruction without a Manhattan direction. For non-Manhattan segments, or when the estimated Manhattan frame is unreliable, we retain the assumption that each line segment projects from a straight 3D line, but do not impose a Manhattan orientation. The reconstructed line must lie within its *interpretation plane*, defined by the camera center and the viewing rays through the line segment. Existing depth estimates near the segment are then used to recover both its orientation within this plane and its metric scale. In this case, the depth support must constrain the orientation as well as the scale.

Integration with depth completion. Each reconstructed 3D line provides depth estimates along its image projection. As illustrated in Figure 3.7b, these estimates can extend into regions where depth remains undefined after the preceding stages. They provide fixed constraints for dense depth completion (Section 3.4).

3.4 Dense Depth Completion

Ground theory, perceptual organization and linear perspective generate incomplete, sparse depth estimates over non-ground regions of the image. To generate dense depth from these sparse estimates, we adapt a method originally developed for luminance completion [25]. In particular, we solve Laplace's equation, $\nabla^2 z = 0$, for all unknown pixels, treating our known depth estimates as fixed boundary conditions and semantic boundaries as insulators. This diffusion process smoothly fills all remaining regions, respecting object contours. For segments with no estimated depth pixels, we first assign values to boundary pixels directly abutting completed regions, and then apply Laplace completion to fill the new regions. This process is repeated until all non-sky semantic segments are assigned valid depth values. Figure 3.7c shows an example dense depth map resulting from this completion process.

Chapter 4

Experiments

This chapter details the implementation, evaluation, and analysis of our explainable depth estimation framework. We compare our geometrically-grounded approach against state-of-the-art black box deep learning models under the input conditions summarized in Table 4.2. The comparison does not use identical camera information for all methods. We also evaluate whether our geometric estimates provide complementary information through fusion.

4.1 Implementation Details

Our framework combines Python-based semantic segmentation using InternImage with a MATLAB implementation of horizon estimation, 3D line projection, and Laplace diffusion for dense depth completion in C, leveraging efficient sparse matrix operations.

Aside from the pre-trained semantic segmentation model, our framework remains predominantly training-free. The few parameters requiring calibration are restricted to the following lightweight components:

1. **Indoor/Outdoor Classifier** (Eq. 3.1): A Multinomial Naive Bayes classifier with two probability vectors, $\mathbf{p}_{\text{indoor}}$ and $\mathbf{p}_{\text{outdoor}}$.

2. **Horizon Estimation** (Sec. 3.2.1): Three univariate Gaussian distributions used to model the probability of the horizon's y-coordinate in each image column: $\mathcal{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$ for the prior location of the horizon, $\mathcal{N}(\mu_{\text{gnd}}, \sigma^2_{\text{gnd}})$ and $\mathcal{N}(\mu_{\text{sky}}, \sigma^2_{\text{sky}})$ for the likelihood of horizon y-coordinate given the ground and sky masks, respectively.

3. **Bottom Object Depth** (Sec. 3.3.2): A kernel regression model $D(s)$ conditioned on semantic class s, which predicts the depth of non-ground objects in contact with the image's bottom edge.

These models are trained using the SYNS-Patches dataset [46].

4.1.1 Datasets and Metrics

Training. We do not train any networks, but learn several parameters governing our explainable method. For this we employ the SYNS-Patches validation dataset [46, 64]. SYNS-Patches is derived from the SYNS dataset [57], which consists of co-registered HDR imagery and LiDAR spherical scans collected from 92 distinct scenes spanning agricultural fields, natural landscapes, residential neighborhoods, industrial zones, and indoor environments. SYNS-Patches [46, 64] is derived by extracting 18 planar projections from each spherical SYNS scene, yielding approximately 1.6K images at a resolution of 376×1242, matching the aspect ratio of KITTI. This work relies heavily on SYNS-Patches due to its scene diversity, dense metric depth, and well-calibrated camera parameters.

Test. We quantitatively evaluate our system on three benchmarks: The SYNS-Patches test split [46] (400 indoor and outdoor scenes), the Cityscapes validation split [52] (500 outdoor urban images), and an 800-image ScanNet test subset following the evaluation protocol of [20]. For all three datasets, ground-truth camera height, pitch, and roll are available in the provided metadata; their ranges across the evaluation splits are summarized in Table 4.1. While our explainable reference method is given these camera parameters, most of

Table 4.1: Camera height and orientation ranges on the quantitative evaluation splits. N is the number of evaluated images. Heights are in meters; pitch and roll are in degrees. Brackets indicate ranges, and single values indicate fixed parameters.

Dataset	N	Height (m)	Pitch (°)	Roll (°)
SYNS-Patches	400	1.65	0.00	0.00
Cityscapes	500	[1.18, 1.22]	[2.18, 2.86]	0.00
ScanNet	800	[0.74, 1.86]	[−73.95, 41.42]	[−18.68, 15.87]

the comparison deep network models are not given these parameters but have to implicitly or explicitly infer them. For SYNS-Patches and Cityscapes this may not be a major issue since the camera parameters are close to normative and most networks were trained under similar conditions. However, the camera parameters vary much more broadly in ScanNet. For this dataset the difference in input between our method and comparison methods is important and should be kept in mind when interpreting results.

We also show qualitative outdoor examples on KITTI [47] and VKITTI2 [55].

Evaluation Metrics. To quantitatively assess metric depth estimation performance, we employ standard error and accuracy metrics. Let z_i^* denote the ground-truth depth and z_i the predicted depth for pixel i, and let N denote the total number of valid pixels (i.e., pixels with ground-truth depth within the evaluation range of the dataset). We report the following:

$$\text{AbsRel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|z_i^* - z_i|}{z_i^*}, \tag{4.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i^* - z_i)^2}, \tag{4.2}$$

Additionally, we report F_A [65], a geometric accuracy metric defined on back-projected 3D point clouds rather than directly on depth values. For each image, the predicted and

Table 4.2: Training data and camera inputs for each model in our evaluation. Checkmarks indicate that ground-truth intrinsics $\mathbf{K}$ or extrinsics $(\mathbf{R}, \mathbf{t})$ are supplied at inference time; crosses indicate that they are not supplied.

Method	Training / Pretraining Data	Given $\mathbf{K}$	Given $\mathbf{R}, \mathbf{t}$
AdaBins [33]	NYUv2 (indoor) and KITTI (outdoor), separately trained supervised models	✗	✗
NewCRFs [31]	Supervised training on NYUv2, KITTI, and Matterport3D	✗	✗
ZoeDepth [43]	MiDaS-M12 pretraining with metric fine-tuning on NYUv2/KITTI	✗	✗
DepthAnything [66]	1.5M labeled images (e.g., HRWSI, MegaDepth, TartanAir) plus 62M unlabeled images	✓	✗
ZeroDepth [22]	Metric training on Parallel Domain, TartanAir, Waymo, and OmniData	✓	✗
UniK3D [3]	26-source cross-camera training mixture (e.g., Waymo, Matterport3D, ScanNet++)	✓	✗
DepthPro [44]	Mixed real and synthetic data (e.g., Hypersim, TartanAir, VKITTI2)	✗	✗
Metric3Dv2 [4]	18-dataset geometric training mixture (e.g., DDAD, Waymo, ScanNet, Hypersim)	✓	✗
UniDepth [5]	9-dataset, 3M-image real-world training mixture (e.g., Waymo, ScanNet, Taskonomy)	✓	✗
Ours	Lightweight statistical modeling on SYNS-Patches	✓	✓

ground-truth depth maps are back-projected using the camera intrinsics to obtain two 3D point sets. For a 3D distance tolerance τ, precision is defined as the fraction of predicted points whose nearest neighbor in the ground-truth point cloud lies within τ, while recall is the fraction of ground-truth points whose nearest neighbor in the predicted point cloud lies within τ. The corresponding F1 score is then computed as the harmonic mean of precision and recall. The metric F_A is the normalized area under the resulting F1-versus-threshold curve over $\tau \in [0, d_{\max}/20]$, where $d_{\max}$ denotes the maximum evaluation depth for the dataset. This corresponds to tolerances up to 5% of the maximum depth range.

Unlike per-pixel metrics such as AbsRel and RMSE, F_A evaluates geometric consistency after back-projection, offering a complementary, geometry-level view of reconstruction quality.

4.2 Metric Depth Estimation

Baseline methods and fairness of comparison. Table 4.2 summarizes the training data for each baseline, along with the camera information provided to the model at inference time in our evaluation setup. Several baselines are supplied with ground-truth camera intrinsics $\mathbf{K}$, although only Metric3Dv2 and ZeroDepth explicitly require intrinsics as part of their

Table 4.3: Quantitative comparison on **outdoor** scenes. *Ground* evaluation is restricted to ground regions only, while *Full* evaluation considers the entire scene. AbsRel and F_A are reported as percentages, and RMSE is reported in meters. Best and second-best performance is bold-faced and underlined, respectively.

Method	**Ground Region**						**Full Scene**					
	SYNS-Patches			Cityscapes			SYNS-Patches			Cityscapes		
	AbsRel↓	RMSE↓	F_A↑	AbsRel↓	RMSE↓	F_A↑	AbsRel↓	RMSE↓	F_A↑	AbsRel↓	RMSE↓	F_A↑
AdaBins [33]	27.0	7.02	75.7	20.2	2.74	91.2	37.2	13.41	68.3	27.6	14.10	77.9
NewCRFs [31]	26.7	7.66	73.5	22.4	5.98	85.9	37.8	13.68	66.1	43.0	21.10	63.2
ZoeDepth [43]	20.7	4.91	84.06	21.5	3.58	88.0	27.4	9.36	51.9	25.1	10.93	74.3
DepthAnything [66]	18.6	<u>4.12</u>	87.5	13.1	<u>2.38</u>	94.4	25.8	<u>7.48</u>	80.4	15.2	7.39	82.7
ZeroDepth [22]	30.0	7.21	72.2	19.8	3.06	90.3	34.4	11.00	77.0	20.1	7.44	81.5
UniK3D [3]	33.7	6.66	68.7	15.7	2.56	93.4	39.7	10.74	64.4	18.8	7.53	83.4
DepthPro [44]	32.3	6.37	73.1	12.9	3.11	94.0	36.5	10.03	71.0	16.6	7.97	82.6
Metric3Dv2 [4]	<u>18.2</u>	4.43	**88.2**	**7.6**	2.48	**96.2**	<u>22.0</u>	7.61	**81.9**	**8.5**	**5.33**	**90.9**
UniDepth [5]	**17.1**	**3.90**	<u>88.1</u>	<u>10.6</u>	**2.17**	<u>96.1</u>	**21.0**	**6.81**	<u>80.9</u>	<u>12.1</u>	<u>5.87</u>	<u>88.6</u>
Ours	21.2	5.07	85.7	13.4	5.08	92.2	39.8	11.35	73.1	29.6	14.80	72.4

released inference interface. UniDepth, UniK3D, and DepthPro are designed to estimate camera intrinsics if not provided, whereas our method is the only approach that additionally uses extrinsic parameters $(\mathbf{R}, \mathbf{t})$. In addition, some baselines are domain-specialized rather than fully unified: AdaBins, NewCRFs, and DepthAnything V1 use separate indoor and outdoor checkpoints.

Quantitative evaluation. Tables 4.3 and 4.4 present a quantitative comparison of the proposed explainable reference system against state-of-the-art black box approaches on outdoor (SYNS-Patches and Cityscapes) and indoor (SYNS-Patches and ScanNet) test sets, respectively. Performance is reported separately for the ground region and for the full image.

Outdoor Scenes (Table 4.3) For outdoor ground surfaces, our method remains reasonably competitive. On SYNS-Patches, it ranks fifth in both absolute relative error and RMSE, and fourth in F_A. On Cityscapes, it again ranks fifth in absolute relative error, but falls to ninth in RMSE and sixth in F_A. As expected, performance deteriorates in the full-scene setting, since the method derives much of its strength from ground-based geometric cues. On

Table 4.4: Quantitative comparison on **indoor** scenes. *Ground* evaluation is restricted to ground regions only, while *Full* evaluation considers the entire scene. AbsRel and F_A are reported as percentages, and RMSE is reported in meters. Best and second-best performance is bold-faced and underlined, respectively.

Method	**Ground Region**						**Full Scene**					
	SYNS-Patches			ScanNet			SYNS-Patches			ScanNet		
	AbsRel↓	RMSE↓	F_A↑	AbsRel↓	RMSE↓	F_A↑	AbsRel↓	RMSE↓	F_A↑	AbsRel↓	RMSE↓	F_A↑
AdaBins [33]	29.6	1.91	61.4	33.0	0.66	39.7	44.4	3.28	76.6	42.5	0.74	40.2
NewCRFs [31]	18.8	2.02	74.4	40.5	0.97	17.8	46.3	2.92	76.4	37.2	0.94	35.2
ZoeDepth [43]	63.0	5.35	19.1	12.3	0.28	69.5	57.8	4.73	51.0	15.8	0.31	<u>69.9</u>
DepthAnything [66]	32.7	4.70	70.0	<u>11.3</u>	<u>0.26</u>	<u>71.5</u>	50.4	4.35	54.9	**12.5**	**0.27**	**73.5**
ZeroDepth [22]	29.7	1.47	64.3	50.2	1.12	8.8	27.6	1.49	74.7	44.6	1.02	22.7
UniK3D [3]	24.3	2.20	71.6	40.5	0.84	15.2	21.1	1.93	81.3	34.6	0.69	37.0
DepthPro [44]	12.0	1.04	87.0	12.9	0.29	66.3	17.5	1.32	86.6	<u>13.8</u>	<u>0.29</u>	69.2
Metric3Dv2 [4]	<u>10.8</u>	**0.38**	<u>92.0</u>	49.6	1.03	4.9	**10.7**	**0.94**	**91.2**	49.1	0.96	18.8
UniDepth [5]	13.5	<u>0.58</u>	87.3	37.3	0.78	19.4	<u>12.7</u>	<u>0.96</u>	<u>89.1</u>	33.9	0.66	39.3
Ours	**9.4**	1.05	**93.2**	**4.6**	**0.18**	**90.9**	42.1	3.70	74.4	38.1	0.66	51.3

SYNS-Patches full-scene evaluation, it ranks last in absolute relative error, eighth in RMSE, and fifth in F_A. On Cityscapes full-scene evaluation, it ranks ninth in all three metrics.

Indoor Scenes (Table 4.4) On the indoor ground-region benchmarks, our system performs strongly, ranking first in absolute relative error and F_A on both SYNS-Patches and ScanNet. This is somewhat surprising, given the simplicity of our system. It may be partially explained by the fact that we have given the intrinsic and extrinsic camera parameters to our reference system. The black box systems, on the other hand, have not been trained on SYNS-Patches and must generalize across varying camera geometries or adapt to them implicitly. Furthermore, floor surfaces in indoor environments are often flat and planar, which makes our ground-plane prior a stronger physical model than in irregular outdoor terrain.

However, in the full-scene setting the performance drops as the system must propagate depth beyond these reliable surfaces. On SYNS, our method ranks sixth in AbsRel and eighth in both RMSE and F_A. On ScanNet, it ranks seventh in AbsRel and fourth in both RMSE and F_A.

In summary, the quantitative analysis shows that our simple explainable reference system

does not generally match the performance of the strongest learned baselines, which typically contain hundreds of millions of parameters and are trained on massive multi-domain datasets. Nevertheless, the fact that it is more competitive for the ground regions than non-ground regions does support the strategy of semantically separating these regions in order to apply principles of ground theory.

4.3 Qualitative Evaluation

Figure 4.1 presents a qualitative comparison between our method and two state-of-the-art black-box approaches on SYNS-Patches. Although the learning-based baselines often recover finer local depth variations within semantic regions, our method accurately preserves the large-scale ground geometry and overall scene layout in many ground-dominated scenes.

Figure 4.2 compares absolute relative error maps and bird's-eye-view point cloud reconstructions on SYNS-Patches. In the highlighted regions in bird's-eye view, our method produces coherent planar structure at (a) and sharp wall and floor boundaries at (b). By contrast, the competing methods exhibit more boundary blur and larger localized error in these regions.

Finally, Figure 4.3 shows qualitative results on KITTI and VKITTI2. Despite differences in camera parameters and scene characteristics, the method consistently recovers coherent geometry, supporting the robustness and broad applicability of the proposed geometric framework.

4.4 Fusion

Figure 4.4 shows Pearson correlations between the pixelwise depth estimation errors made by pairs of models. Interestingly, we see that overall, errors committed by our explainable

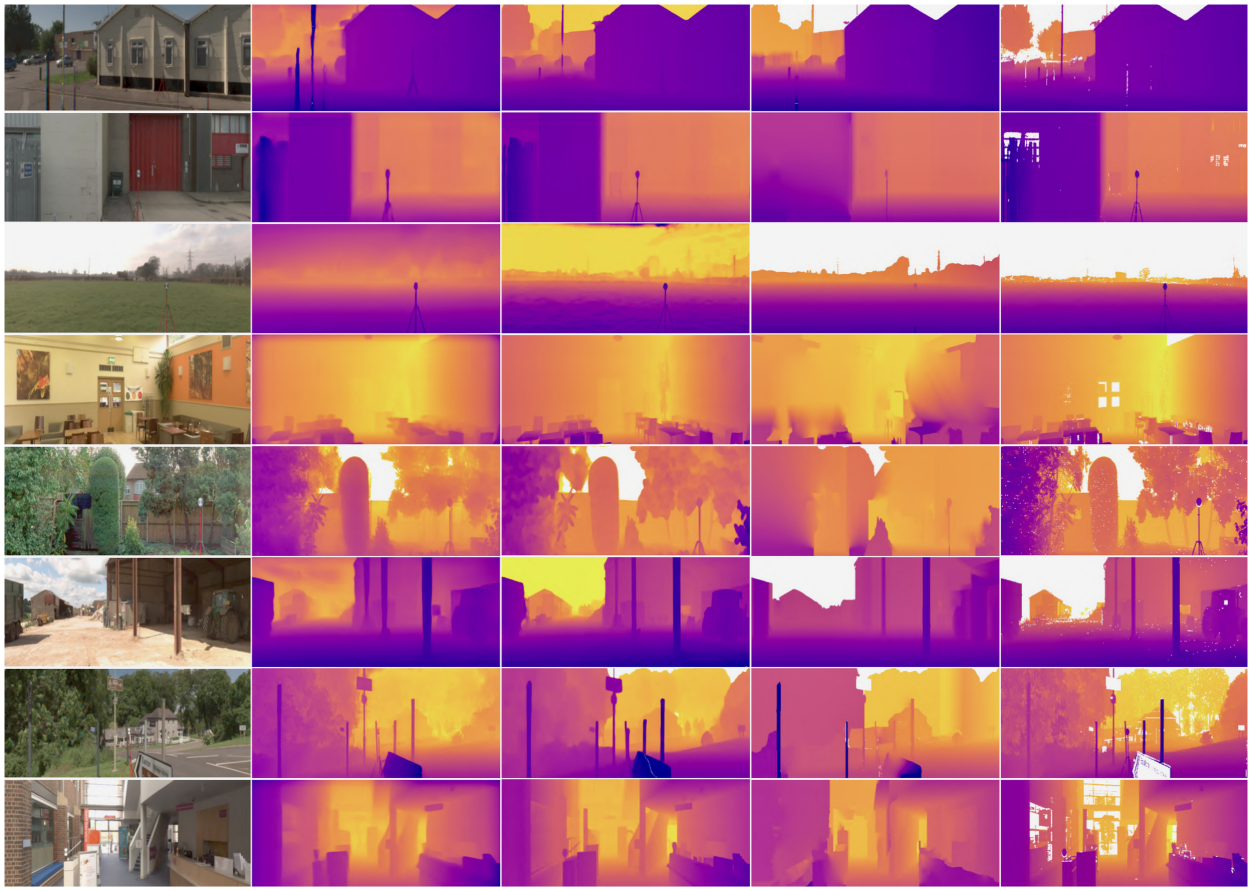


Figure 4.1: Qualitative depth comparison on SYNS-Patches. Columns show the RGB input, Depth Anything, Metric3Dv2, our prediction, and ground-truth depth. Each row shows a different scene. The comparison illustrates differences in ground geometry, object boundaries, and within-region depth variation.

reference model tend to correlate less with errors committed by black box models ($\rho = 0.30$ on average), relative to correlations between black box models ($\rho = 0.45$ on average). (Note that the rightmost column is more blue than other columns.) This suggests that these black box models may derive some benefit from fusing with our explainable model.

To test this, we define a simple weighted sum ensemble model:

$$\hat{z} = w\, z_{\mathrm{e}} + (1 - w)\, z_{\mathrm{bb}}, \tag{4.3}$$

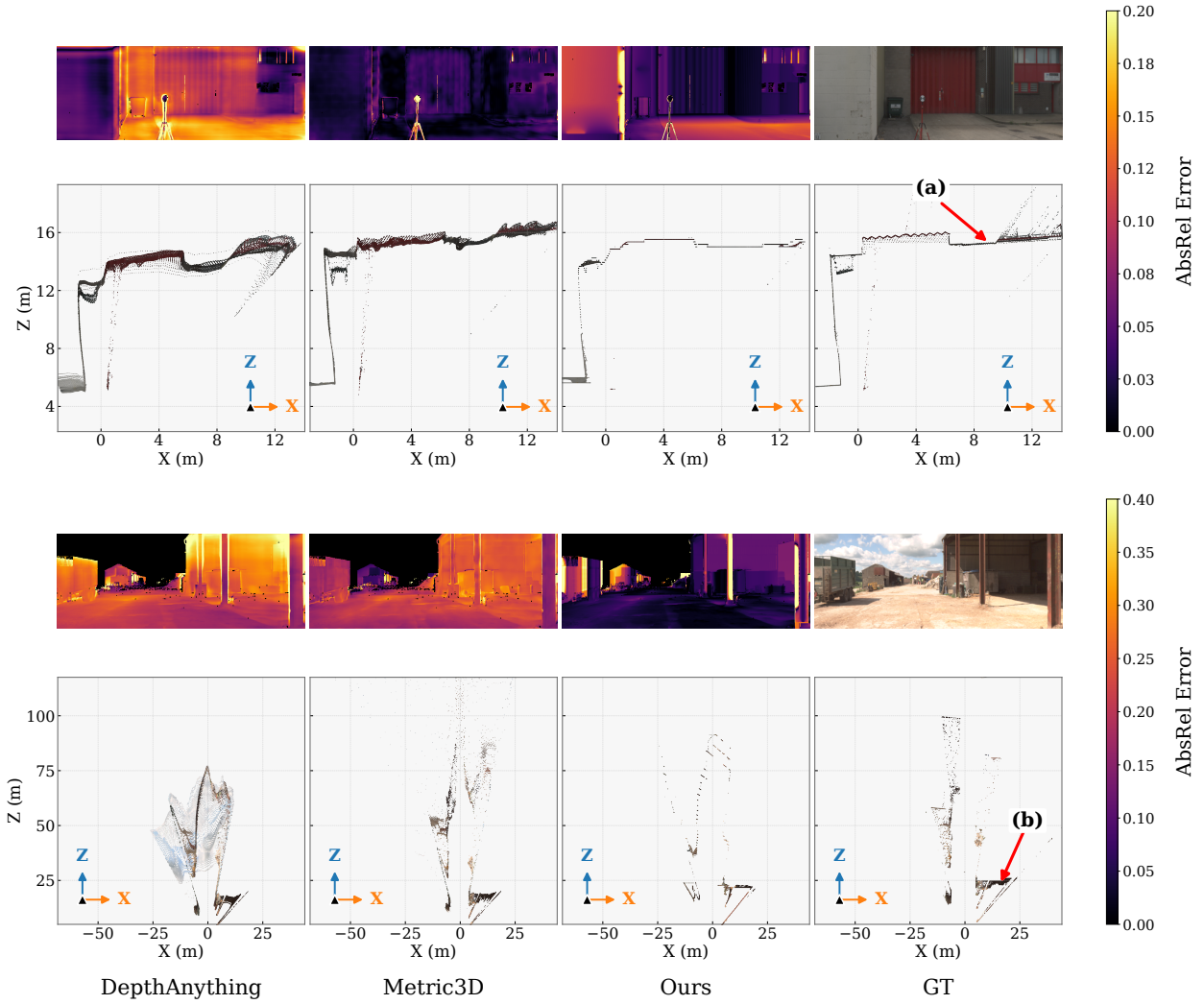


Figure 4.2: Qualitative comparison on two SYNS-Patches images. The top rows show the RGB inputs and absolute relative error maps, while the bottom rows show the corresponding bird's-eye-view point cloud reconstructions of non-ground regions, plotted in metric coordinates with shared axes. The axes indicate the camera coordinate frame (X right, Z forward). Highlighted regions illustrate **(a)** wall and door structures and **(b)** the wall–ground junction.

where z_e and z_{bb} are depth estimates from explainable and black box models, respectively, and the blending weight $w \in [0, 1]$ is optimized to minimize the absolute relative error on the SYNS-Patches training set, separately for each pair of models.

Table 4.5 shows the results of this experiment. We find that fusing our explainable model with black box models yields significant performance gains, with an average overall

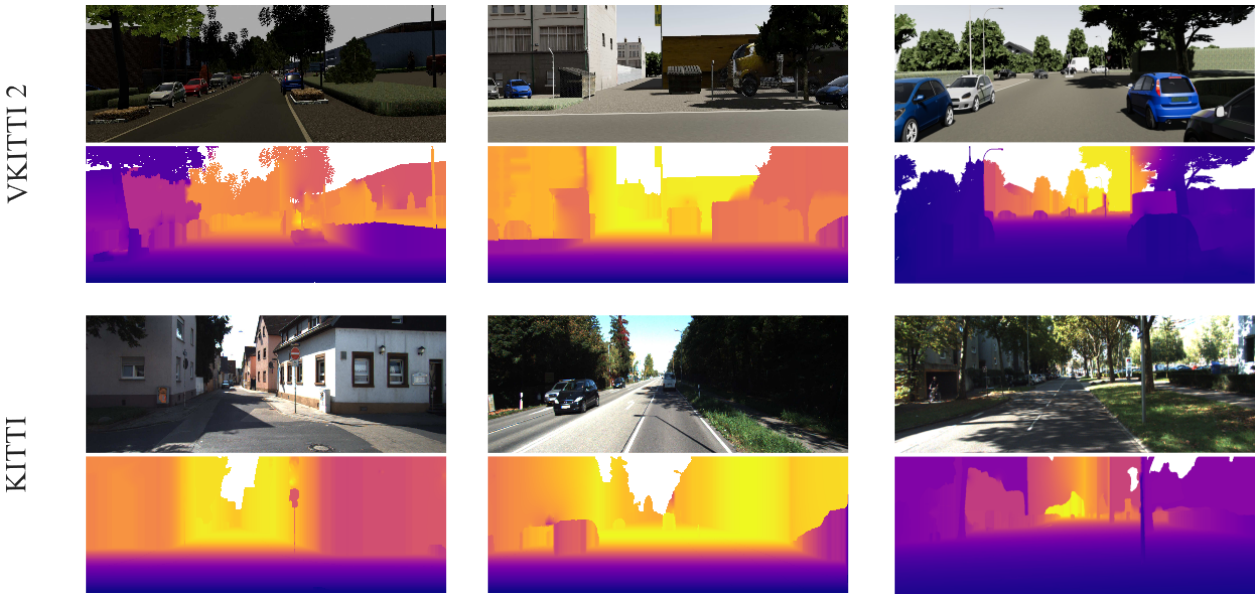


Figure 4.3: Qualitative results on VKITTI2 (upper group) and KITTI (lower group). Within each group, RGB images appear above their predicted depth maps. The examples show the geometry recovered by our pipeline across the two driving datasets.

Table 4.5: Relative percentage reductions in absolute relative error from fusion with our explainable model on the SYNS-Patches test set. Positive values indicate improvement over the original baseline; negative values indicate deterioration. The weight w is the contribution of our model. Full, Gnd, and Non-Gnd denote full-scene, ground-region, and non-ground-region evaluations.

Target Method	w	**Indoor** (%)			**Outdoor** (%)			Overall (%)
		Full	Gnd	Non-Gnd	Full	Gnd	Non-Gnd	
AdaBins [33]	0.42	16.4	33.4	16.2	9.2	25.3	5.7	10.7
NewCRFs [31]	0.45	22.7	24.0	22.7	7.5	23.5	4.0	10.7
ZoeDepth [43]	0.26	17.0	23.9	16.9	−1.4	5.6	−3.0	4.3
DepthAnything [66]	0.19	11.6	17.3	11.4	−0.4	2.7	−1.0	3.1
ZeroDepth [22]	0.29	−2.4	23.5	−2.8	10.9	22.8	7.5	8.9
UniK3D [3]	0.34	−6.6	25.0	−7.2	11.4	24.3	7.5	9.4
DepthPro [44]	0.33	−28.2	16.7	−28.7	14.3	29.1	10.5	10.4
Metric3Dv2 [4]	0.06	−3.6	2.7	−3.8	1.6	4.5	0.6	1.1
UniDepth [5]	0.04	0.6	1.7	0.6	0.6	2.1	0.2	0.6
Mean	0.26	3.1	18.7	2.8	6.0	15.5	3.6	6.6

improvement of 6.6% in absolute relative error and a maximum overall reduction of 10.7% on SYNS-Patches. This suggests that many deep networks do not fully encode the three simple principles explored here, and that incorporating such priors can mitigate some failure modes.

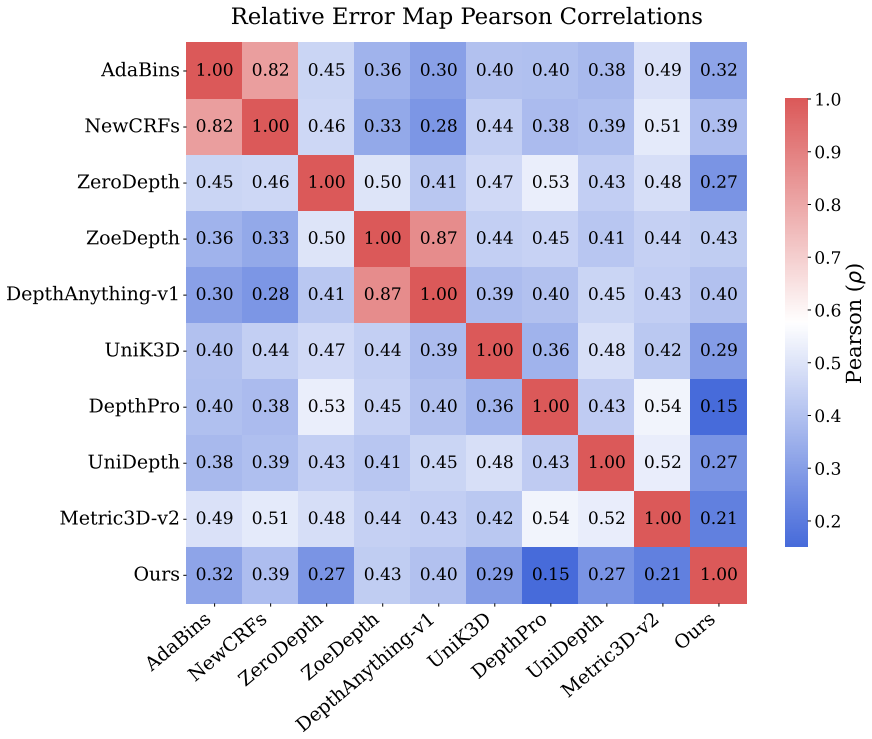


Figure 4.4: Pearson correlations between signed relative depth-error maps on SYNS-Patches. Each cell compares a pair of models; warmer colors indicate higher correlations. The final row and column compare our explainable model with each learned baseline.

The largest improvements occur in the ground region, averaging 18.7% indoors and 15.5% outdoors, with reductions in absolute relative error of up to 33.4% in indoor ground regions. For methods already strong in non-ground regions (e.g., DepthPro indoors), blending with our estimates can introduce error in those areas, though ground-region improvements remain consistent. Further, we see that benefits of fusion are small for the most powerful models (Metric3Dv2 and UniDepth) suggesting that these networks have come close to embodying these principles.

4.5 Ablations

We conduct ablation experiments on the SYNS-Patches test set to validate key attributes of our explainable model.

4.5.1 Ground Plane Modeling

We compare several approaches to recover ground depth (Table 4.6). The most flexible geometric model (Inclined & Rolled) significantly outperforms the naive horizontal assumption, reducing absolute relative error from 28.6% to 23.0%. The best AbsRel in Table 4.6 is achieved by combining geometric estimation with the learned profile, reaching 21.2%. However, a paired t-test across 301 scenes comparing the full ground-plane pipeline against the profile-only variant shows that this difference is not statistically significant ($p = 0.24$). This suggests that, on this benchmark, the geometric component does not yield a statistically reliable improvement over the learned profile.

The limited gain over the learned profile stems primarily from errors in horizon estimation rather than a ceiling on geometric modeling. When supplied with the ground-truth horizon, the geometric model substantially outperforms the profile-only variant (10.2% vs. 21.5% AbsRel error). This indicates that the benefits of geometric reasoning are largely constrained by the accuracy of the horizon estimate.

4.5.2 Linear Perspective

Table 4.7 evaluates performance on non-ground regions with and without line segment integration. Incorporating line segments improves all three metrics, reducing AbsRel from 53.3% to 42.8% and RMSE from 16.51 m to 14.37 m, while increasing F_A from 69.7% to 77.6%. These results confirm the substantial importance of linear perspective cues for non-ground monocular depth estimation.

Table 4.6: Comparison of ground-plane models and their combinations with a learned depth profile on outdoor scenes from the SYNS-Patches test set. AbsRel and F_A are percentages; RMSE is in meters. Arrows indicate the preferred direction, and the best value in each column is bold.

Ground Plane Modeling	AbsRel(%) ↓	RMSE (m) ↓	F_A(%) ↑
Profile (learned)	21.5	**4.98**	**86.0**
Horizontal	28.6	7.62	83.0
Inclined	23.1	6.02	84.1
Inclined & Rolled	23.0	5.89	84.6
Horizontal + Profile	24.7	5.05	85.4
Inclined + Profile	22.0	5.23	85.0
Inclined & Rolled + Profile	**21.2**	5.06	85.7

Table 4.7: Effect of line-segment integration on non-ground depth estimation for indoor and outdoor scenes from the SYNS-Patches test set. AbsRel and F_A are percentages; RMSE is in meters. Arrows indicate the preferred direction, and the best value in each column is bold.

Method	AbsRel(%) ↓	RMSE (m) ↓	F_A(%) ↑
With Linear Perspective	**42.8**	**14.37**	**77.6**
Without Linear Perspective	53.3	16.51	69.7

4.6 Failure Modes

To partition sources of error, we conducted an oracle experiment on the SYNS-Patches, using ground-truth depth values for all ground pixels and then completing the rest of our pipeline based upon these correct ground truth depth values. Replacing estimated ground depths with ground-truth values reduces full-scene absolute relative error by 12.9 percentage points, from 39.8% to 26.9%, demonstrating the influence of ground-depth accuracy on the subsequent inference stages. The remaining absolute relative error of 26.9% could reflect imperfections in the semantic segmentation, deviations from our approximation of surfaces abutting the ground as vertical, and limitations of Laplace completion in large or geometrically complex regions. The gap between ground-region performance (9.4%) and full-scene performance (42.1%) on SYNS-Patches indoor scenes highlights that the ground anchor itself is relatively accurate for indoor scenes - the main bottleneck lies in propagating this estimate to non-ground regions with complex geometry.

We believe that deviations in terrain planarity and horizon detection inaccuracies are primary sources of the ground-depth errors whose correction yields this 12.9-percentage-point reduction in full-scene absolute relative error.

4.7 Conclusion

The experiments show that our reference system is particularly competitive in indoor ground regions, although it does not generally match the strongest learned baselines over full scenes. Fusion improves several baselines, indicating that the geometric estimates provide complementary information. The ablation and oracle experiments demonstrate the influence of line constraints and ground-depth accuracy on subsequent inference. These results should be interpreted in light of the supplied camera parameters and the assumptions governing ground geometry and depth propagation.

Chapter 5

Discussion and Conclusion

In this thesis, we assessed the extent to which three fundamental principles—*Ground Theory*, *perceptual organization*, and *linear perspective*—can account for monocular metric depth perception. To investigate this, we developed a simple, explainable geometric framework that leverages these principles to reconstruct a dense metric depth map from a single image.

Semantic segmentation forms the foundation of the system by identifying key scene components such as the ground, sky, and discrete objects. This enables the imposition of geometric constraints that reflect the physical structure of the environment. In particular, the ground plane provides a reference that helps resolve the inherent scale ambiguity of monocular depth estimation and supports the inference of object depths throughout the scene.

While a number of state-of-the-art, high-capacity black-box networks trained on large and diverse datasets outperform this simple explainable system in standard pixel-wise metrics, it is perhaps surprising that the proposed explainable system remains competitive in several respects despite its simplicity. This finding suggests that these fundamental geometric and perceptual principles capture meaningful aspects of monocular depth perception and can provide a viable foundation for metric reconstruction.

5.1 Key Findings and Implications

Importance of the Ground. Our system's performance in ground regions, particularly in indoor environments (Tables 4.3, 4.4), highlights the foundational role of the ground as a metric reference frame. By semantically isolating this plane, we reduce a high-dimensional depth estimation task to a more manageable problem.

Linear Perspective as a Structural Scaffold. A significant finding is that integrating linear perspective cues via 3D line segments reduces absolute relative error by 10.5 percentage points, from 53.3% to 42.8% (see Table 4.7), in non-ground regions. This confirms that structural regularities provide a reliable geometric scaffold that deep learning models may overlook in favor of stochastic texture and context signals.

Complementarity with Learned Models. A compelling validation of our approach comes from the fusion results. A straightforward linear combination of our low-capacity explainable model with high-capacity black-box networks reduces absolute relative error by up to 33.4% in indoor ground regions and by up to 10.7% overall on SYNS-Patches (Table 4.5). This suggests that many deep networks do not fully encode the three simple principles explored here. Notably, the benefits of fusion are largest for weaker models and diminish for the strongest (Metric3Dv2, UniDepth), suggesting that these networks have come closer to internalizing these geometric priors.

Predictable Failure Modes and Interpretability. Unlike opaque neural networks, our framework exhibits traceable and structurally consistent failure modes. Because it is grounded in projective geometry and explicit ecological priors, errors can be directly attributed to violations of its underlying assumptions, such as ground planarity, the Manhattan-world model, or the approximation of ground-adjacent objects as vertical

surfaces.

Figure 5.1 illustrates predictable failure modes in a forest scene where several modeling assumptions are violated. The uneven ground surface, covered with grass, debris, and undergrowth, is approximated as a single plane, leading to errors in ground depth estimation. In addition, suspended branches and nearby vegetation violate the assumption that non-ground segments are objects resting on the ground directly beneath them. Consequently, they inherit depth from the distant ground and are overestimated. The scene also lacks clear Manhattan structure, which limits the availability of reliable 3D guiding lines.

Value of Explainability. The explicit formulation supports both scientific understanding and failure diagnosis. By inspecting intermediate estimates and modeling assumptions, we can trace how depth predictions arise and identify potential sources of error. For example, ground-depth errors can be examined in relation to the planar approximation, while errors on suspended branches can be examined through the assumed support relationships. This supports targeted improvements and controlled experiments assessing the contribution of individual perceptual principles.

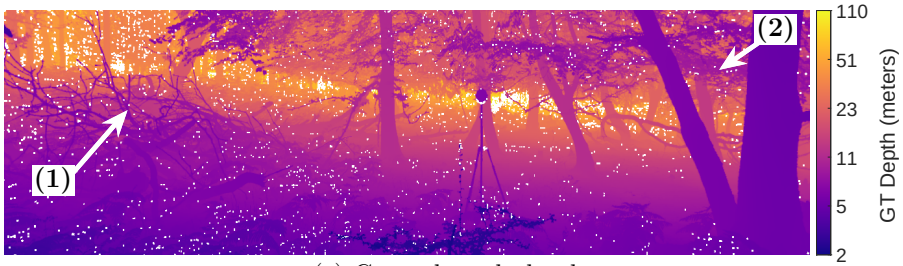
Explainability also helps us anticipate failure modes and respond appropriately. Explicit assumptions reveal conditions under which particular geometric cues may be unreliable, providing a basis for deciding when to rely on a cue, reduce its influence, or seek alternative depth information. For example, the Manhattan-frame reliability score [17] can inform the selection or weighting of line-based constraints. More broadly, this transparency can guide adaptive cue weighting in hybrid systems and support safer operation by making the model's limitations explicit and informing how its predictions are used.



(a) Input RGB.



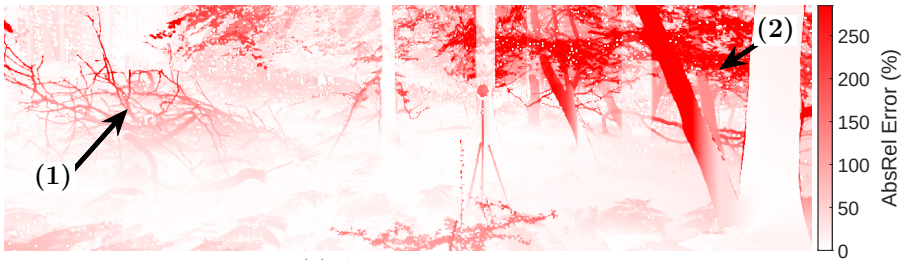
(b) Ground plane mask.



(c) Ground truth depth.



(d) Estimated metric depth.



(e) Absolute relative error.

Figure 5.1: Failure cases in a forest scene. Panels show the RGB image, ground-plane mask, ground-truth depth, estimated depth, and absolute relative error. Errors concentrate in the debris-covered, non-planar ground region (1) and in suspended branches and foliage that inherit distant ground depth (2).

5.2 Limitations and Future Work

There are several limitations to our current framework that present opportunities for future research:

- **Online Calibration:** While we assumed known camera parameters, integrating online estimation systems (e.g., [17]) could reduce reliance on supplied calibration. Our metric inference would additionally require camera height to be supplied or estimated. Online calibration was not evaluated in this thesis.

- **Non-Planar Ground:** The primary failure point for ground estimation occurs in heavily forested or rugged terrains where the planar assumption breaks down. Incorporating *shape-from-texture* modules and non-planar ground representations could extend the system's applicability to varied ground surfaces.

- **Intra-Segment Geometry:** Our approach is currently limited in modeling the 3D shape of surfaces within individual semantic segments. A promising future direction is the deployment of "subroutine" algorithms, such as SAM3D [67], on non-ground segments independently, conditioned on the boundary depth anchors provided by our system.

Long-Term Goals. The proposed method supports two complementary directions. First, we aim to develop a more self-contained system by estimating camera intrinsics and orientation online, investigating camera-height estimation using known scene dimensions or learned scale priors, and incorporating more general ground representations. Second, building on our fusion results, we aim to integrate the geometric reference system with learned depth models. A network conditioned on both the RGB image and the depth map generated by our pipeline could refine scene details and compensate for violations of the geometric assumptions.

5.3 Conclusion

In conclusion, our pipeline provides an explainable framework for understanding the roles of ground theory, perceptual organization, and linear perspective in monocular depth estimation. By aligning the computational method with ecological theories of human perception, we open a potential avenue for a more comprehensive investigation of ways in which human and AI monocular depth estimation may be aligned, and ways in which they diverge. This approach is complementary to the standard data-driven approaches in the AI literature, and may be important in the quest for more explainable, reliable, and generalizable depth estimation systems.

References

[1] D. Hoiem, A. A. Efros, and M. Hebert. "Geometric context from a single image". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Vol. 1. 2005, pp. 654–661.

[2] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. "Depth Anything v2". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 21875–21911.

[3] L. Piccinelli, C. Sakaridis, M. Segu, Y.-H. Yang, S. Li, W. Abbeloos, and L. Van Gool. "UniK3D: Universal camera monocular 3D estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025, pp. 1028–1039.

[4] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen. "Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2024.

[5] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu. "UniDepth: Universal monocular metric depth estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 10106–10116.

[6] E. B. Goldstein. *Sensation and Perception*. 8th ed. Belmont, CA: Wadsworth, Cengage Learning, 2010.

[7] J. Gibson. *The Perception of the Visual World*. Boston, MA: Houghton Mifflin, 1950.

[8] H. Sedgwick. "JJ Gibson's "Ground Theory of Space Perception"". In: *i-Perception* 12.3 (2021).

[9] X. Ren and J. Malik. "Learning a classification model for segmentation". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2003, pp. 10–17.

[10] M. D. Anderson, J. H. Elder, E. W. Graf, and W. J. Adams. "The time-course of real-world scene perception: Spatial and semantic processing". In: *iScience* 25.12 (2022).

[11] A. Oliva. "Gist of the scene". In: *Neurobiology of Attention*. Ed. by L. Itti, G. Rees, and J. K. Tsotsos. Burlington, MA: Elsevier, 2005, pp. 251–256.

[12] A. Oliva and A. Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope". In: *International Journal of Computer Vision* 42.3 (2001), pp. 145–175.

[13] A. Oliva and A. Torralba. "Building the gist of a scene: The role of global image features in recognition". In: *Progress in Brain Research*. Vol. 155. Elsevier, 2006, pp. 23–36.

[14] J. M. Coughlan and A. L. Yuille. "Manhattan World: Orientation and outlier detection by Bayesian inference". In: *Neural Computation* 15.5 (2003), pp. 1063–1088.

[15] P. Denis, J. H. Elder, and F. Estrada. "Efficient edge-based methods for estimating Manhattan frames in urban imagery". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 5303. 2008, pp. 197–210.

[16] R. Tal and J. Elder. "An accurate method for line detection and Manhattan frame estimation". In: *Proceedings of the Asian Conference on Computer Vision Workshops (ACCV Workshops)*. Vol. II. 2012, pp. 580–593.

[17] Y. Qian and J. Elder. “A reliable online method for joint estimation of focal length and camera rotation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2022.

[18] Y. Qian, S. Ramalingam, and J. H. Elder. “LS3D: Single-view gestalt 3D surface reconstruction from Manhattan line segments”. In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 2018, pp. 399–416.

[19] S. Bower. *Perspective Rule #1: Converging Lines*. https://drawingperspectives.com/2020/03/25/1-perspective-rule-1-converging-lines/. 2020.

[20] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler. “Repurposing diffusion-based image generators for monocular depth estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[21] V. Guizilini, P. Tokmakov, A. Dave, and R. Ambrus. “GRIN: Zero-shot metric depth with pixel-level diffusion”. In: *Proceedings of the International Conference on 3D Vision (3DV)*. 2025, pp. 112–122.

[22] V. Guizilini, I. Vasiljevic, D. Chen, R. Ambrus, and A. Gaidon. “Towards zero-shot scale-aware monocular depth estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 9233–9243.

[23] X. Yang, Z. Ma, Z. Ji, and Z. Ren. “GEDepth: Ground embedding for monocular depth estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023.

[24] A. Cecille, S. Duffner, F. Davoine, T. Neveu, and R. Agier. “GroCo: Ground constraint for metric self-supervised monocular depth”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2024, pp. 54–69.

[25] J. H. Elder. “Are edges incomplete?” In: *International Journal of Computer Vision* 34.2 (1999), pp. 97–122.

[26] E. Delage, H. Lee, and A. Y. Ng. “A Dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 2006, pp. 2418–2428.

[27] A. Saxena, M. Sun, and A. Y. Ng. “Make3D: Learning 3D scene structure from a single still image”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.5 (2008), pp. 824–840.

[28] D. Eigen and R. Fergus. “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2650–2658.

[29] Z. Cheng, Y. Zhang, and C. Tang. “Swin-depth: Using transformers and multi-scale fusion for monocular-based depth estimation”. In: *IEEE Sensors Journal* 21.23 (2021), pp. 26912–26920.

[30] R. Ranftl, A. Bochkovskiy, and V. Koltun. “Vision transformers for dense prediction”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 12179–12188.

[31] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan. “Neural window fully-connected CRFs for monocular depth estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 3916–3925.

[32] S. Shao, Z. Pei, W. Chen, X. Wu, and Z. Li. “NDDepth: Normal-distance assisted monocular depth estimation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 7931–7940.

[33] S. F. Bhat, I. Alhashim, and P. Wonka. "AdaBins: Depth estimation using adaptive bins". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 4009–4018.

[34] S. F. Bhat, I. Alhashim, and P. Wonka. "LocalBins: Improving depth estimation by learning local distributions". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2022, pp. 480–496.

[35] V. Arampatzakis, G. Pavlidis, N. Mitianoudis, and N. Papamarkos. "Interpretable vision transformers in monocular depth estimation via SVDA". In: *arXiv preprint arXiv:2602.11005* (2026).

[36] R. Garg, V. K. Bg, G. Carneiro, and I. Reid. "Unsupervised CNN for single view depth estimation: Geometry to the rescue". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2016, pp. 740–756.

[37] C. Godard, O. Mac Aodha, and G. J. Brostow. "Unsupervised monocular depth estimation with left-right consistency". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 270–279.

[38] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. "Unsupervised learning of depth and ego-motion from video". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1851–1858.

[39] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. "Digging into self-supervised monocular depth estimation". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019, pp. 3828–3838.

[40] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon. "3D Packing for Self-Supervised Monocular Depth Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[41] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia. "MonoViT: Self-supervised monocular depth estimation with a vision transformer". In: *Proceedings of the International Conference on 3D Vision (3DV)*. 2022, pp. 668–678.

[42] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.3 (2020), pp. 1623–1637.

[43] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. "ZoeDepth: Zero-shot transfer by combining relative and metric depth". In: *arXiv preprint arXiv:2302.12288* (2023).

[44] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun. "Depth Pro: Sharp monocular metric depth in less than a second". In: *arXiv preprint arXiv:2410.02073* (2024).

[45] H. Jung, E. Park, and S. Yoo. "Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 12642–12652.

[46] J. Spencer, C. S. Qian, C. Russell, S. Hadfield, E. Graf, W. Adams, A. J. Schofield, J. H. Elder, R. Bowden, H. Cong, et al. "The monocular depth estimation challenge". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 623–632.

[47] D. Eigen, C. Puhrsch, and R. Fergus. "Depth map prediction from a single image using a multi-scale deep network". In: *Advances in Neural Information Processing Systems* 27 (2014).

[48] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. “Indoor segmentation and support inference from RGBD images”. In: *European Conference on Computer Vision (ECCV)*. 2012, pp. 746–760.

[49] S. K. Ramakrishnan et al. “HM3D: 1000 Large-Scale 3D Environments for Embodied AI”. In: *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*. 2021.

[50] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder. “The mapillary vistas dataset for semantic understanding of street scenes”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 4990–4999.

[51] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia. “DrivingStereo: A Large-scale dataset for stereo matching in autonomous driving scenarios”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 899–908.

[52] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. “The Cityscapes dataset for semantic urban scene understanding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[53] D. Kim, D. W. Min, S.-H. Ham, and K. Sohn. “Deep monocular depth estimation via integration of global and local predictions”. In: *IEEE Transactions on Image Processing* 27.8 (2018), pp. 4131–4144.

[54] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind. “Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 10912–10922.

[55] Y. Cabon, N. Murray, and M. Humenberger. "Virtual KITTI 2". In: *arXiv preprint arXiv:2001.10773* (2020).

[56] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. "nuScenes: A Multimodal dataset for autonomous driving". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11621–11631.

[57] W. J. Adams, J. H. Elder, E. W. Graf, J. Leyland, A. J. Lugtigheid, and A. Muryy. "The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude". In: *Scientific Reports* 6.1 (2016), p. 35805.

[58] I. Vasiljevic et al. "DIODE: A dense indoor and outdoor depth dataset". In: *arXiv preprint arXiv:1908.00463* (2019).

[59] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. "A multi-view stereo benchmark with high-resolution images and multi-camera videos". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3260–3269.

[60] T. Koch, C. Wittich, F. Männel, R. A. Schubert, and R. Stiefelhagen. "Evaluation of CNN-based single-image depth estimation methods". In: *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*. 2018, pp. 331–348.

[61] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, et al. "InternImage: Exploring large-scale vision foundation models with deformable convolutions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 14408–14419.

[62] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. "Semantic understanding of scenes through the ADE20K dataset". In: *International Journal of Computer Vision* 127.3 (2019), pp. 302–321.

[63] E. J. Almazán, R. Tal, Y. Qian, and J. H. Elder. "MCMLSD: A dynamic programming approach to line segment detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2031–2039.

[64] J. Spencer, C. Russell, S. Hadfield, and R. Bowden. "Deconstructing self-supervised monocular reconstruction: The design decisions that matter". In: *arXiv preprint arXiv:2208.01489* (2022).

[65] E. P. Örnek, S. Mudgal, J. Wald, Y. Wang, N. Navab, and F. Tombari. "From 2D to 3D: Re-thinking benchmarking of monocular depth prediction". In: *arXiv preprint arXiv:2203.08122* (2022).

[66] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. "Depth Anything: Unleashing the power of large-scale unlabeled data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[67] X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li, et al. "SAM 3D: 3Dfy anything in images". In: *arXiv preprint arXiv:2511.16624* (2025).