Consistent Monocular Depth Estimation with Contact Region Boundary-Aware Refinement

Yinuo Wang[1], Qing Miao[1](✉), and Wangmeng Zuo[1]

Harbin Institute of Technology, Harbin, China
25B903188@stu.hit.edu.cn, csqmiao@imu.edu.cn, cswmzuo@gmail.com

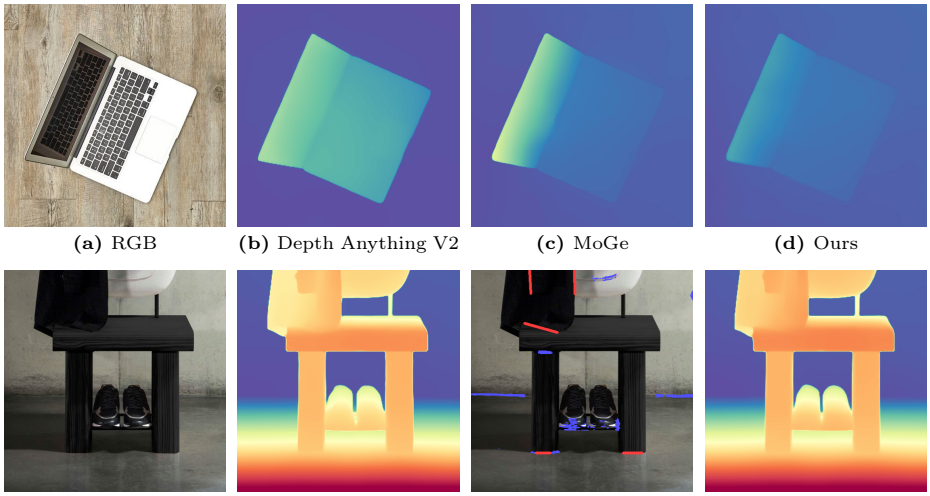
Abstract. While contemporary monocular depth estimation (MDE) methods achieve remarkable overall acacy, they consistently produce erroneous depth discontinuities at object contact regions, particularly between objects and supporting surfaces. In this paper, we address this critical limitation by presenting a boundary-aware monocular depth estimation framework that enforces depth continuity at contact areas through the principled exploitation of contact boundaries as explicit structural priors. Specifically, we propose a boundary detection and filtering module that explicitly identifies object contact regions, yielding a novel boundary-aware representation that enables depth-consistent learning at contact areas. Furthermore, we introduce a boundary-aware feature fusion strategy that seamlessly incorporates contact boundary priors into the depth decoding process, effectively rectifying the persistent discontinuities that elude existing approaches. Our framework further supports interactive refinement, allowing users to manually specify missing contact boundaries for controllable depth correction. Extensive experiments on three unseen benchmarks with dense object interactions demonstrate the effectiveness of our approach, consistently outperforming baselines with particularly pronounced gains in contact regions. Code is available at https://github.com/abai969/contact-depth-refinement.git

Keywords: Monocular depth estimation · Depth consistency · Boundary-aware and Fusion Mechanism

1 Introduction

Monocular depth estimation (MDE) infers per-pixel depth from a single RGB image and underpins a wide range of applications, including autonomous driving and robotic perception [4,36,39,49]. Depth foundation models trained under pixel-wise supervision [37,38] have recently achieved substantial gains in prediction accuracy and cross-domain generalization. Nevertheless, pixel-wise supervision treats each pixel independently, thereby neglecting inter-pixel geometric dependencies. This limitation has motivated recent approaches that reformulate MDE within 3D point-cloud representations [23,34], imposing stronger geometric consistency constraints and yielding more structurally coherent depth maps.

Despite the strengthened geometric consistency, depth discontinuities at object contact regions remain an open challenge for both per-pixel regression and



(a) RGB **(b)** Depth Anything V2 **(c)** MoGe **(d)** Ours

(e) RGB **(f)** Initial prediction **(g)** Refined boundary **(h)** Refined prediction

Fig. 1: Depth predictions and boundary refinement examples. Top row (a-d): existing methods (b, c) exhibit depth discontinuities at object contacts, while ours (d) preserves depth continuity. Bottom row (e-h): interactive refinement with a user-specified contact boundary (g) corrects depth inconsistency in the initial prediction (f).

3D point-cloud-based approaches. As illustrated in the top row of Fig. 1, Depth Anything V2 [38] (Fig. 1b) produces a sharp depth discontinuity at the contact boundary between a laptop and the floor, despite their physical adjacency. Although MoGe [34] (Fig. 1c), which derives depth from dense 3D point representations, alleviates this artifact in large contact areas, it still exhibits noticeable discontinuities at finer-grained contacts, such as between the laptop screen and keyboard. This limitation stems from the absence of explicit contact-structure awareness during depth decoding, as neither paradigm provides direct guidance on where depth continuity should be preserved across physically adjacent surfaces.

In this paper, we propose a boundary-aware monocular depth estimation framework that enforces depth continuity at contact regions by equipping the depth decoder with explicit contact-structure priors. To explicitly model where depth continuity should be preserved, we introduce a contact boundary detection and filtering module that distills boundary representations from raw edge maps. Furthermore, we propose a boundary-aware feature fusion strategy that integrates contact priors into the depth decoder, enabling the network to produce depth-consistent predictions at contact regions. By incorporating specific training constraints on contact boundaries, the framework effectively leverages boundary cues during both the decoding and optimization phases, resulting in consistent and fine-grained depth predictions at object contact regions. As shown in Fig. 1d, our method yields smoother depth transitions at contact regions compared to existing approaches. Moreover, the framework supports interactive

refinement by allowing users to manually specify missing contact boundaries, thereby improving local depth consistency in regions where automatic extraction fails, as demonstrated in the bottom row of Fig. 1. More implementation details are provided in the supplementary material.

Our main contributions are summarized as follows:

- We propose a boundary-aware monocular depth estimation framework that explicitly models object contact boundaries as structural priors to enforce depth consistency in these regions.
- We introduce a contact boundary detection and filtering module and a boundary-aware feature fusion strategy that jointly suppress erroneous depth discontinuities, while further supporting interactive refinement through user-specified contact boundaries.
- Extensive experiments on three unseen benchmarks featuring dense contact relationships demonstrate consistent improvements over strong baselines while requiring only limited training data.

2 Related Work

Monocular Depth Estimation. Over the years, research on monocular depth estimation has transitioned from manually designed algorithms to learning-based frameworks, resulting in remarkable improvements in depth estimation accuracy. [4, 6, 7, 10, 18, 27, 36, 39, 49]. Despite these advances, many existing approaches still aim to recover depth in absolute physical units [1, 2, 4, 13, 22, 23]. Predicting depth in real-world units requires reliable metric supervision, which is typically obtained through active sensing devices such as RGB-D cameras, LiDAR, or calibrated stereo systems. Due to the limited availability and high acquisition cost of such measurements, existing datasets primarily cover structured environments, particularly indoor and driving scenes. However, the limited scalability of metric supervision has motivated alternative formulations. To alleviate the reliance on metric annotations, subsequent works relax the requirement of absolute scale and instead estimate depth up to a global scale and shift [8, 19, 28, 37, 38, 42]. Such formulations enable training with more diverse and loosely annotated data, thereby improving generalization across domains. In parallel, generative approaches model depth prediction as a conditional generation task, leveraging large-scale pretrained models to enhance structural consistency [5, 9, 15, 21]. Recent advances in monocular depth estimation range from direct pixel-wise regression to geometry-aware representations. Without predicting depth maps directly, more recent methods learn structured 3D representations, such as affine-invariant point maps, to improve geometric consistency [22, 23, 34, 35, 43, 44]. However, these methods still lack explicit modeling of object contact relations, which often leads to unstable depth transitions in contact regions.

Edge Cues in Depth Estimation. Edge information has long served as an effective geometric constraint for improving depth prediction [7,24,26,33,40,41,45]. Early self-supervised methods introduce edge-aware smoothness regularization,

where image gradients modulate depth smoothness to prevent over-smoothing across object boundaries [7, 41]. Similarly, edge-preserving methods explicitly introduce image gradient cues during depth upsampling or multi-scale refinement to better maintain sharp depth boundaries [16]. More recent approaches [28, 37, 38] further enforce structural consistency through gradient-based supervision, implicitly leveraging edge information by encouraging alignment between predicted and ground-truth depth gradients. Although these techniques improve boundary sharpness and reduce spurious blurring, they generally treat edges as generic indicators of depth discontinuity. In contrast, object contact regions exhibit more subtle geometric behavior: depth may remain continuous despite strong image edges [22], or conversely require constrained transitions even when visual cues are weak. Existing edge-aware formulations do not explicitly model such contact relations, leaving depth estimation unstable in physically interacting regions. Our method instead treats object contact cues within edge areas as structural priors to reduce depth inconsistencies at contact boundaries.

3 Methodology

Our framework builds upon MoGe [34], an affine-invariant monocular geometry estimator that predicts dense 3D point maps, and augments it with explicit contact-structure awareness to enforce depth consistency at object contact regions. The key idea is to first identify where depth continuity should be preserved, and then leverage this information to guide depth decoding. To this end, we introduce a contact boundary detection and filtering module that distills contact-aware representations from raw edge maps (§3.2), and a boundary-aware feature fusion strategy that integrates these representations into the depth decoder to suppress erroneous discontinuities at contact regions (§3.3). An overview of the method is shown in Fig. 2.

3.1 Preliminary

In the general pipeline of monocular depth estimation, e.g., MoGe, given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, visual features $\mathbf{F} = \mathcal{B}_\phi(\mathbf{I})$ are extracted using a DINOv2 backbone $\mathcal{B}_\phi(\cdot)$, preserving the spatial structure of the image. Subsequently, a multi-scale convolutional decoder $\mathcal{D}_\theta(\cdot)$ progressively upsamples $\mathbf{F}$ to the original resolution, producing a dense, affine-invariant 3D point map $\hat{\mathbf{P}} = \mathcal{D}_\theta(\mathbf{F})$. Each pixel (u, v) in the input image is associated with a 3D point $\hat{\mathbf{P}}_{u,v} \in \mathbb{R}^3$, where the z-coordinate encodes the affine-invariant depth.

Since the predicted point map is affine-invariant, *i.e.*, defined up to an unknown global scale and translation, the model is trained with a set of alignment-based geometric losses. Let $\hat{\mathbf{p}}_i$ denote the 3D point predicted at the pixel i, and $\mathbf{p}_i$ its corresponding ground-truth point. MoGe employs a geometric loss $\mathcal{L}_G$ to enforce accurate global recovery of depth and 3D structure.

$$\mathcal{L}_G = \sum_{i \in M} \frac{1}{z_i} \left\| s\hat{\mathbf{p}}_i + \mathbf{t} - \mathbf{p}_i \right\|_1, \tag{1}$$

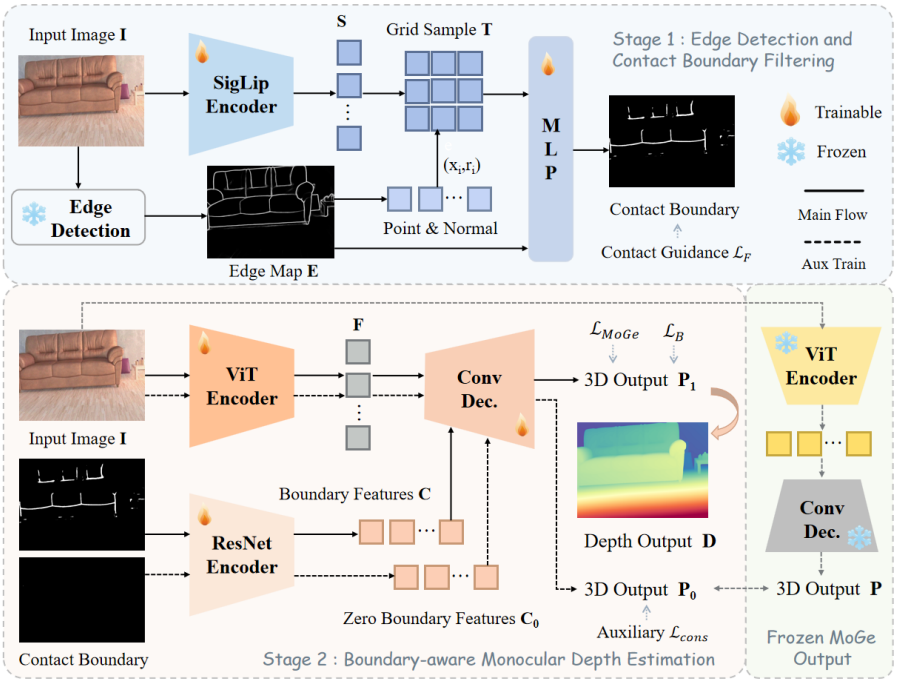


Fig. 2: Overview of the proposed framework. In the first stage, contact boundaries are extracted from raw edge maps through detection and filtering. In the second stage, the extracted contact boundaries are integrated as structural priors into the depth decoder to enforce depth consistency at contact regions. Dashed components are active only during training.

where M denotes the valid regions, z_i is the ground-truth depth at pixel i, and s and $\mathbf{t}$ represent the global scale and translation vector, respectively, which are optimized using the Robust and Efficient (ROE) parallel search algorithm [34].

MoGe further extends the global alignment loss $\mathcal{L}_G$ (Eq. (1)) to local neighborhoods at multiple scales to enhance local geometry. Specifically, for each anchor point, neighboring points within spheres of varying radii are considered, and the same affine alignment constraints and ROE-based optimization are applied. The resulting multi-scale local alignment loss $\mathcal{L}_{S(\alpha)}$ is computed at three neighborhood scales $\alpha \in \{1/4, 1/16, 1/64\}$ in a coarse-to-fine manner. To improve surface quality and handle regions with undefined geometry, the model additionally incorporates two auxiliary losses. A normal loss $\mathcal{L}_N$ enforces consistency between surface normals computed from the predicted point map and the ground truth, and a mask loss $\mathcal{L}_V$ supervises the prediction of valid geometric regions. The overall MoGe loss is defined as:

$$\mathcal{L}_{MoGe} = \mathcal{L}_G + \sum_{\alpha} \mathcal{L}_{S(\alpha)} + \mathcal{L}_N + \mathcal{L}_V. \tag{2}$$

Although MoGe achieves accurate relative depth prediction, it still exhibits discontinuities and noticeable gaps at object contact regions. To address this limitation, we extract and filter object contact boundaries and integrate them as structural priors into the depth decoder, enabling depth-consistent predictions at these regions.

3.2 Contact Boundary Detection and Filtering

To enforce depth continuity at contact regions, we devise a hierarchical detection-and-filtering scheme that progressively localizes and refines object contact boundaries in images, comprising an edge detection module and a subsequent contact boundary filtering module.

Edge Detection Module. Edge detection methods yield dense boundary maps from a single input image [11,20,48], capturing structural transitions across multiple scales and thereby providing a rich reservoir of candidate contact boundaries. Critically, contact boundaries may manifest not only at the interface between distinct objects but also within the internal structure of a single object, demanding a detector capable of resolving both inter-object and intra-object boundaries simultaneously. To this end, we adopt MuGE [47], a multi-granularity edge detector that exposes a tunable parameter $a \in [0, 1]$ governing the level of structural detail, affording a principled trade-off between boundary completeness and fine-grained complexity.

MuGE follows an encoder–decoder architecture wherein a VGG-based [31] encoder extracts hierarchical multi-scale features and a lightweight convolutional decoder progressively upsamples them into a dense edge map. We formalize MuGE as an edge detector $\mathcal{U}(\cdot)$ and set the tunable parameter $a \in [0, 1]$ to 0.5, striking a balance between capturing inter-object boundaries and internal object edges. This yields the edge map $\mathbf{E} = \mathcal{U}(\mathbf{I})$, $\mathbf{E} \in \mathbb{R}^{H \times W}$. However, as a general-purpose edge map, $\mathbf{E}$ encompasses all structural boundaries present in the image, of which only a subset corresponds to genuine inter-object contact boundaries, while the remaining edges are irrelevant to physical object contact. Relying solely on $\mathbf{E}$ is therefore inadequate for our objective, which demands precise identification of contact boundaries to the exclusion of all others. This observation motivates a dedicated filtering stage designed to selectively retain only those edges attributable to true inter-object contact.

Contact Boundary Filtering Module. Determining whether an edge corresponds to a contact boundary fundamentally requires understanding the semantic relationship between the regions on its two sides, such as physical contact or support. Such relationships can be explicitly described in natural language—for example, "a cup placed on a table"—where the notion of contact between objects is clearly conveyed. Such relational interactions are naturally expressed in natural language, where phrases describing spatial or support relations (e.g., an object resting on another surface) explicitly convey object contact. Motivated by this alignment between visual contact cues and linguistic relational descriptions, we employ a pretrained vision–language image encoder to extract features that better capture cross-region relational information. Such

models [3, 25, 32, 46] are jointly trained on large-scale image–text pairs to align visual representations with the semantic space of language, enabling visual features to not only capture appearance and structural cues but also implicitly encode semantics related to object attributes and their interactions. Even without explicit textual input at inference time, the image encoder retains this semantic modeling capability, allowing the resulting feature maps to more effectively represent boundary-relevant properties and thus provide more reliable representations for contact boundary prediction. Such semantically enriched visual features are therefore particularly well suited for identifying contact boundaries.

Hence, we employ the SigLIP [32,46] visual encoder as the feature extraction backbone, which produces dense and locally consistent features that capture both appearance and semantic cues. Let $\mathbf{S} = \mathcal{G}_\phi(\mathbf{I})$ denote the resulting feature map. To determine whether each detected edge pixel corresponds to a contact boundary, we extract feature descriptors from $\mathbf{S}$ at locations indicated by the edge map $\mathbf{E}$. Each point is represented as $\mathbf{q}_i = (\mathbf{x}_i, \mathbf{r}_i)$, where $\mathbf{x}_i \in \mathbb{R}^2$ denotes the pixel coordinate and $\mathbf{r}_i \in \mathbb{R}^2$ denotes the corresponding local unit normal direction. For each $\mathbf{q}_i$, we sample features from $\mathbf{S}$ along its normal on both sides of the boundary, denoted as $\mathbf{s}_i^+ = \mathbf{S}(\mathbf{x}_i + \delta\mathbf{r}_i)$ and $\mathbf{s}_i^- = \mathbf{S}(\mathbf{x}_i - \delta\mathbf{r}_i)$, and compute their difference $\Delta\mathbf{s}_i = \mathbf{s}_i^+ - \mathbf{s}_i^-$. This cross-boundary difference captures local contrast and aids in distinguishing true contact edges from other edges. Each point-wise edge feature is then formed by concatenating the original and sampled features:

$$\mathbf{t}_i = \psi\big(\mathbf{s}_i, \mathbf{s}_i^+, \mathbf{s}_i^-, \Delta\mathbf{s}_i\big), \tag{3}$$

where $\psi(\cdot)$ represents a fixed concatenation. The set of all edge features, $T = \{\mathbf{t}_i\}_{i=1}^N$, is processed by a lightweight MLP classifier $f_\theta(\cdot)$ to predict contact probabilities for the candidate points. The predicted contact probabilities $\hat{b}_i$ for each boundary point are then assigned to their respective boundary location to form the final contact boundary map $\hat{\mathbf{B}} \in \mathbb{R}^{H\times W}$, with non-boundary pixels retaining their original edge responses from $\mathbf{E}$.

To train the contact boundary filtering module, we freeze the edge detection module and optimize the filtering parameters to minimize pixel-level discrepancies between predictions and ground truth, where contact boundaries are automatically generated via edge detection and depth consistency verification on both sides, without manual annotation. Let $b_i \in \mathbf{B}$ denote the ground-truth at pixel i. Given the set of valid edge pixels M, the BCE loss is defined as:

$$\mathcal{L}_{\text{bce}} = \frac{1}{|M|} \sum_{i\in M} w_i \cdot \text{BCE}(\hat{b}_i, b_i), \tag{4}$$

where w_i is a positional weight that assigns larger emphasis to positive edge pixels. While the BCE loss provides local supervision, it does not explicitly enforce consistency along continuous boundary structures. To this end, we introduce an edge variance loss that regularizes predictions at the curve level. Specifically, from the ground-truth map $\mathbf{B}$, we extract a set of connected curves $L = \{l_k\}_{k=1}^{|L|}$, where each l_k is a contiguous sequence of positive pixels forming a locally shortest connected contact region on the curve. For each curve l_k, let $\hat{l}_k$ denote the

predicted probabilities at the pixels within this curve. The edge variance loss is then formulated as:

$$\mathcal{L}_{\mathrm{e}} = \frac{1}{|L|} \sum_{k=1}^{|L|} \mathrm{Var}(\hat{l}_k), \tag{5}$$

where $\mathrm{Var}(\cdot)$ denotes empirical variance of predicted probabilities along the curve l_k. Finally, the total contact boundary filtering module loss is expressed as:

$$\mathcal{L}_F = \mathcal{L}_{\mathrm{bce}} + \lambda_e \, \mathcal{L}_{\mathrm{e}}. \tag{6}$$

3.3 Boundary-aware Monocular Depth Estimation

In this section, the extracted contact boundary map $\hat{\mathbf{B}}$ is integrated into MoGe to enforce depth continuity across contact boundaries, enabling boundary-aware monocular depth estimation.

Boundary Feature Fusion. Through the edge detection and contact filtering modules, we obtain a contact probability map $\hat{\mathbf{B}}$ that indicates object contact regions in the image. However, $\hat{\mathbf{B}}$ remains a pixel-level confidence map without hierarchical structure or contextual representation capability. Its representation differs substantially from the multi-scale feature space used in the depth estimation pipeline, making it unsuitable for direct integration. To bridge this gap, we introduce a boundary encoding module that lifts the contact probability map into a high-dimensional feature space, producing structurally enriched boundary-aware representations. We adopt a ResNet-based encoder [12] for this purpose. Formally, the encoding process is defined as $\mathbf{C} = \mathcal{R}_\phi(\hat{\mathbf{B}})$, where $\mathbf{C}$ denotes the encoded boundary feature map.

The encoded feature $\mathbf{C}$ enhances the structural expressiveness of boundary information and reconstructs contact relationships across scales. To allow boundary cues to directly influence depth recovery, we inject $\mathbf{C}$ into the decoding stage in a multi-scale manner. Specifically, we align $\mathbf{C}$ with the image feature maps $\mathbf{F}$ obtained in Sec. 3.1 and fuse them via the operator $\psi(\cdot)$, yielding $\tilde{\mathbf{F}} = \psi(\mathbf{F}, \mathbf{C})$. The decoder then operates on the fused representation to produce the boundary-aware affine-invariant point cloud $\hat{\mathbf{P}}' = \mathcal{D}_\theta(\tilde{\mathbf{F}})$.

Boundary-aware Depth Supervision. We further introduce a boundary-aware loss on the boundary-enhanced affine-invariant point cloud $\hat{\mathbf{P}}'$ to encourage depth continuity at object contact regions. Following the alignment strategy in Eq. (1) of MoGe, we solve for the optimal scale s^* and translation $\mathbf{t}^*$ that best align the prediction with the ground truth. Since our goal is to regularize depth continuity at contact regions, we extract the third coordinate of the aligned predicted and ground-truth points, denoted as $\hat{d}_i$ and d_i, respectively. Let M_b denote the set of pixels located on the predicted contact boundaries $\hat{\mathbf{B}}$. We enforce depth alignment within M_b via $\mathcal{L}_{\mathrm{bd}}$, defined as:

$$\mathcal{L}_{\mathrm{bd}} = \frac{1}{|M_b|} \sum_{i \in M_b} \left| \hat{d}_i - d_i \right|. \tag{7}$$

While $\mathcal{L}_{\text{bd}}$ enforces absolute depth alignment within contact regions, it does not constrain the smoothness or consistency of depth variations. Point-wise depth supervision alone cannot guarantee consistent depth gradients across boundaries. To regularize the local boundary geometry, we model depth variations using finite-difference spatial gradients and introduce a first-order gradient consistency constraint. For each pixel i, the depth gradients for the ground truth and prediction are respectively defined as $\nabla d_i = (\nabla_x d_i, \nabla_y d_i)^\top$ and $\nabla \hat{d}_i = (\nabla_x \hat{d}_i, \nabla_y \hat{d}_i)^\top$. We define the normalized ground truth gradient direction as:

$$\mathbf{n}_i = \frac{\nabla d_i}{\|\nabla d_i\|_2}. \tag{8}$$

To enforce gradient consistency while being robust to scale variations, we measure the gradient residual projected onto this normalized direction:

$$\mathcal{L}_{\text{grad}} = \frac{1}{|M_b|} \sum_{i \in M_b} \left| \mathbf{n}_i^\top \left(\nabla \hat{d}_i - \nabla d_i \right) \right|. \tag{9}$$

In addition, to suppress spurious oscillations within boundary regions, we impose a smoothness regularization along the same gradient normal direction:

$$\mathcal{L}_{\text{m}} = \frac{1}{|M_b|} \sum_{i \in M_b} \left| \mathbf{n}_i^\top \nabla \hat{d}_i \right|. \tag{10}$$

Collectively, the depth regression, gradient consistency, and smoothness terms provide complementary zero-order and first-order constraints, encouraging accurate and geometrically consistent depth estimation at object boundaries.

However, excessively relying on boundary cues may lead the boundary-guided decoder to deviate from the original prediction when such cues are absent or unreliable. To mitigate this issue, we introduce an auxiliary consistency constraint to regularize the decoder behavior. Specifically, when the boundary map is replaced with an all-zero map, i.e., $\hat{\mathbf{B}} = \mathbf{0}$, the decoder produces an auxiliary point prediction for each pixel, denoted as $\hat{\mathbf{p}}_i^{\text{aux}}$ at pixel i. We enforce consistency between this auxiliary prediction and the entire frozen MoGe output $\hat{\mathbf{p}}_i$ over the full pixel set M:

$$\mathcal{L}_{\text{cons}} = \frac{1}{|M|} \sum_{i \in M} \left\| \hat{\mathbf{p}}_i^{\text{aux}} - \hat{\mathbf{p}}_i \right\|_1. \tag{11}$$

In summary, the boundary-aware depth loss is defined as a weighted sum of the four terms:

$$\mathcal{L}_B = \lambda_b \mathcal{L}_{\text{bd}} + \lambda_g \mathcal{L}_{\text{grad}} + \lambda_m \mathcal{L}_{\text{m}} + \lambda_{aux} \mathcal{L}_{\text{cons}}. \tag{12}$$

Finally, the overall training objective integrates the original MoGe losses with the proposed boundary-aware components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{MoGe} + \mathcal{L}_F + \mathcal{L}_B, \tag{13}$$

where $\mathcal{L}_{MoGe}$ includes the original global and local point losses (§3.1), $\mathcal{L}_F$ supervises the contact boundary predictions (§3.2), $\mathcal{L}_B$ denotes the boundary-aware depth loss, which enhances prediction accuracy in boundary regions while preserving global depth consistency (§3.3).

4 Experiments

4.1 Implementation Details

Data. For training, we randomly select 4,000 images from chosen scenes in the Hypersim [29] dataset, a widely used synthetic indoor dataset that provides high-quality ground-truth depth maps with natural depth transitions around object contact boundaries, facilitating the learning of fine-grained depth details. For evaluation, we use three indoor benchmarks that were not seen during training: NYUv2 [30], iBims-1 [17], and HAMMER [14]. The NYUv2 dataset is a standard benchmark for indoor depth estimation commonly used in previous work. The iBims-1 dataset provides high-resolution depth maps, enabling precise evaluation of fine geometric structures. The HAMMER dataset contains challenging indoor scenes with missing textures, reflective surfaces, and transparent materials, which are known to be difficult for MDE. All datasets are processed following the protocol of MoGe [34], including image resizing and alignment strategies, to maintain consistency with previous comparisons.

Metrics. We evaluate depth estimation accuracy using absolute relative error $\text{Rel}^d = |d - \hat{d}|/d$ and the percentage of inliers δ_1^d with $\max(d/\hat{d}, \hat{d}/d) < 1.25$, across three depth representations: scale-invariant depth, affine-invariant depth, and affine-invariant disparity. To further examine geometric consistency, we employ the same benchmarks as used for depth evaluation to assess point estimation accuracy. Specifically, we report metrics on scale-invariant, affine-invariant, and locally affine-invariant point maps to comprehensively evaluate the geometric accuracy of the predicted 3D point map. The local affine-invariant setting is evaluated within regions defined by dataset annotations or object segmentation masks, enabling assessment of fine-scale geometric consistency. The point map accuracy is quantified by the relative point error $\text{Rel}^p = \|\hat{\mathbf{p}} - \mathbf{p}\|_2 / \|\mathbf{p}\|_2$ and δ_1^p with $\|\hat{\mathbf{p}} - \mathbf{p}\|_2 / \min(\|\mathbf{p}\|_2, \|\hat{\mathbf{p}}\|_2) < 0.25$. All detailed alignment and evaluation settings follow MoGe [34].

Implementation. The initialization and training strategies for the model components are as follows. The edge detection module is initialized with pretrained weights from MuGe [47]. The feature extraction backbone of the contact boundary filtering module is initialized from a pretrained SigLIP2-SO400M/16-512 visual encoder [32]. The depth prediction module uses a ViT encoder and a CNN decoder, initialized with MoGe-Large weights [34], while the newly introduced boundary encoder, designed to fuse boundary information, is initialized from ResNet-18 [12] pretrained weights and its other newly added layers are zero-initialized. Training proceeds in two stages. During training of the first-stage module, the backbone of the contact boundary filtering module is frozen, while its MLP classifier is trained for 5 epochs with a batch size of 2 and a learning rate of 1×10^{-4}. After the initial training of the first-stage module, the entire model is jointly fine-tuned for 12 epochs with a batch size of 8. The contact boundary filtering backbone and the depth prediction encoder use a learning rate of 1×10^{-6}, while all newly introduced components, including the MLP classifier, use 1×10^{-5}. The learning rate is halved every 200 steps. All experiments are

Table 1: Comparison of relative depth estimation on three zero-shot benchmarks. Lower Rel^d and higher δ_1^d, both expressed in percentage, indicate better performance. In the table, **bold** values denote the best results, while values with an underline denote the second-best results. Gray entries correspond to models trained on the respective benchmarks and are excluded from the ranking.

Method	Scale-invariant						Affine-invariant						Affine-invariant disparity					
	NYUv2		iBims-1		HAMMER		NYUv2		iBims-1		HAMMER		NYUv2		iBims-1		HAMMER	
	Rel^d	δ_1^d	Rel^d	δ_1^d	Rel^d	δ_1^d	Rel^d	δ_1^d	Rel^d	δ_1^d	Rel^d	δ_1^d	Rel^d	δ_1^d	Rel^d	δ_1^d	Rel^d	δ_1^d
SharpNet	–	–	–	–	–	–	8.01	93.6	13.1	82.1	18.8	64.3	–	–	–	–	–	–
ZoeDepth	5.62	96.3	7.45	93.2	7.42	94.7	4.76	97.3	5.85	95.7	6.65	95.7	5.21	97.7	6.19	96.9	6.84	96.4
DUSt3R	4.40	97.1	4.98	95.8	–	–	3.73	97.8	3.94	96.6	–	–	4.24	98.1	4.49	96.6	–	–
DepthFM	–	–	–	–	–	–	5.50	96.3	6.12	95.1	15.4	73.6	–	–	–	–	–	–
DA V1	4.77	97.5	5.53	95.8	6.88	96.4	3.82	98.3	4.23	97.3	5.77	97.3	4.20	98.4	4.18	97.6	5.56	98.0
DA V2	5.03	97.3	4.32	**97.9**	5.92	97.7	4.16	97.9	3.44	**98.3**	4.73	98.9	4.11	98.3	3.47	**98.5**	4.97	99.1
Metric3D V2	4.69	97.4	4.23	97.7	4.19	**99.1**	3.94	97.6	3.28	**98.3**	3.02	**99.0**	13.4	81.5	8.55	92.3	**3.17**	**99.2**
UniDepth	3.86	**98.4**	4.79	97.4	4.19	98.4	3.40	**98.6**	3.76	98.2	3.55	98.9	3.78	**98.7**	4.06	98.1	3.64	99.1
MoGe	3.44	**98.4**	3.46	97.0	3.77	98.1	2.92	**98.6**	2.74	97.9	3.00	98.3	3.38	98.6	3.23	98.0	3.30	98.5
Ours	**3.43**	**98.4**	**3.35**	97.7	**3.74**	98.1	**2.90**	**98.6**	**2.67**	98.1	**2.98**	98.3	**3.36**	98.6	**3.12**	98.2	3.27	98.5

Table 2: Comparison of relative point estimation on three zero-shot benchmarks. Lower Rel^p and higher δ_1^p indicate better performance. Local point map accuracy is evaluated on affine-invariant point maps within local object regions, and entries marked with "-" indicate unavailable results.

Method	Scale-invariant						Affine-invariant						Local point					
	NYUv2		iBims-1		HAMMER		NYUv2		iBims-1		HAMMER		NYUv2		iBims-1		HAMMER	
	Rel^p	δ_1^p	Rel^p	δ_1^p	Rel^p	δ_1^p	Rel^p	δ_1^p	Rel^p	δ_1^p	Rel^p	δ_1^p	Rel^p	δ_1^p	Rel^p	δ_1^p	Rel^p	δ_1^p
DUSt3R	5.53	97.1	6.18	95.4	–	–	4.45	97.4	5.04	96.0	–	-	–	–	5.44	95.9	–	–
UniDepth	5.33	**98.4**	5.29	97.4	**4.83**	**98.5**	3.93	**98.4**	4.65	**98.0**	4.15	**98.7**	–	–	5.92	96.0	–	–
MoGe	4.86	**98.4**	4.63	97.1	6.45	98.1	3.68	98.3	3.61	97.3	3.88	98.1	–	–	4.16	96.8	–	–
Ours	**4.84**	**98.4**	**4.49**	**97.7**	6.29	98.1	**3.65**	98.3	**3.52**	**98.0**	**3.83**	98.1	–	–	**4.03**	**97.1**	–	–

performed on a single NVIDIA A100-80G GPU. We set $\lambda_e = 100$ in Eq. (6) to control the connectivity of filtered boundaries. The boundary-related weights λ_b, λ_g and λ_m in Eq. (12) are set to 0.4, 0.3, and 0.3, enforcing depth consistency within boundary regions. The auxiliary weight $\lambda_{aux} = 500$ maintains the initial output when the boundaries are unreliable.

4.2 Quantitative Evaluations

Following the protocol described in Sec. 4.1, we evaluate the zero-shot performance of our method and compare it with several representative MDE methods [2,13,23,34,37,38], focusing specifically on relative depth estimation. The results on three indoor benchmarks, which feature diverse object contact regions, are presented in Tables 1 and 2. Since image-level metrics often obscure geometric inaccuracies in localized, critical contact regions, we also report affine-invariant

Table 3: Comparison of contact region depth estimation on the NYUv2 and iBims-1 datasets.

Method	NYUv2		iBims-1	
	Rel^d	δ_1^d	Rel^d	δ_1^d
Sharpnet	2.29	98.5	2.36	98.3
MoGe	1.59	99.0	1.71	98.7
Ours	**1.31**	**99.2**	**1.25**	**98.8**

metrics restricted to the vicinity of object contacts, as shown in Table 3. This contact-centric evaluation offers a more granular assessment of local depth fidelity, which is essential for capturing complex boundary interactions.

Relative depth estimation. Table 1 presents a comparison of depth estimation under different alignment settings, providing a view of method behavior across evaluation protocols. On the iBims-1 benchmark, our method consistently improves over MoGe across all depth representations. For scale-invariant depth, the Rel^d error decreases from **3.46** to **3.35**, while δ_1^d increases from **97.0%** to **97.7%**. As shown in Table 3, contact-region performance is further improved, with Rel^d decreasing from **1.71** to **1.25**, which indicates the effectiveness of boundary-aware modeling in preserving fine-grained depth structures. We further evaluate on NYUv2 and HAMMER under diverse conditions, consistently outperforming the baseline despite limited boundary annotations in NYUv2 and challenging visual conditions in HAMMER.

Relative point estimation. Table 2 presents a comparison of point map estimation under different alignment settings. Compared with depth evaluation, point-based metrics show larger gains, especially under affine-invariant alignment, consistent with improvements in depth metrics and indicating improved geometric consistency. On the iBims-1 dataset, our method reduces the affine-invariant point error Rel^p from **3.61** (MoGe) to **3.52**, and improves δ_1^p from **97.3%** to **98.0%**. Local point metrics show more pronounced gains, with relative point error decreasing from **4.16** to **4.03**, reflecting improved consistency in fine-scale structures. Similar trends are observed on NYUv2 and HAMMER, where our method consistently achieves lower Rel^p or comparable δ_1^p across all settings, confirming robustness.

Overall, these results demonstrate that our framework, which integrates contact boundary information within the depth prediction process, enhances depth continuity and yields more coherent 3D point predictions, effectively leveraging the boundaries as structural priors.

4.3 Qualitative Comparison

To better highlight improvements in depth predictions at object contact boundaries, we provide qualitative comparisons in Fig. 3. In this figure, several representative methods are compared side by side, with the fifth column showing automatically extracted contact boundaries that guide our framework and enable visual assessment of their effect. Across scenes, depth discontinuities often occur in contact regions where geometric cues are limited. For example, in the first image, thin structures such as folded tabletop corners cause abrupt depth changes in other methods, while ours remains more continuous under contact boundary guidance. Similar effects are observed under challenging conditions, as shown in the third-row example. Visual degradation such as blur or distortion makes depth estimation less reliable, especially in distant regions, while boundary cues help preserve depth consistency. Overall, although global scene layout remains similar across methods, contact boundary cues reduce depth discontinuities in fine-scale regions.

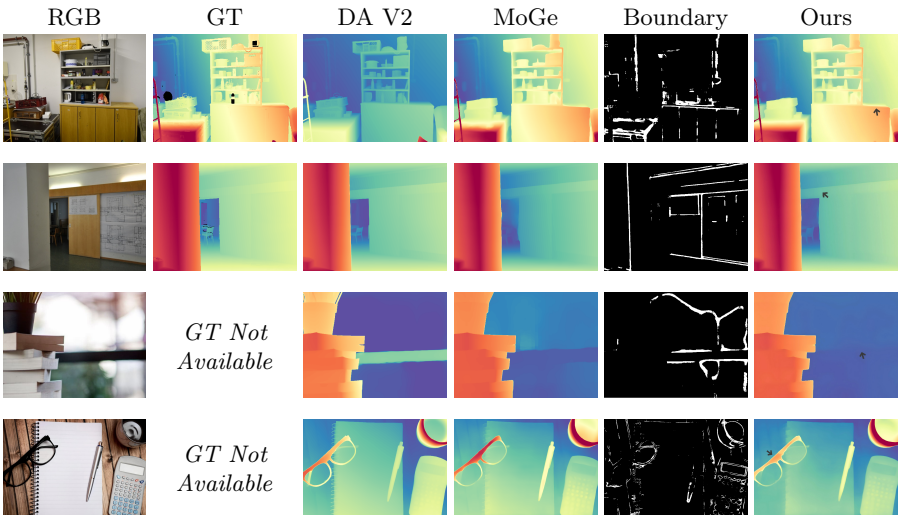


Fig. 3: Qualitative comparison of disparity maps. The top two examples are from our test set, and the bottom two are in-the-wild images collected from the internet. From left to right: input RGB image, ground truth (if available), disparity predictions of DepthAnything V2 and MoGe, and our predicted contact boundaries and disparity maps. **Black arrows** in the last column highlight regions where contact boundary guidance leads to clear improvements in depth estimation.

4.4 Ablation Study

We conduct a systematic ablation study on the iBims-1 dataset to examine how different components of our framework contribute to depth prediction and geometric consistency. Specifically, we progressively evaluate the effects of the edge detection module, the contact boundary filtering module, and the auxiliary consistency loss, $\mathcal{L}_{\text{cons}}$, to quantify their individual contributions. The results are summarized in Table 4.

The first row shows the performance of the baseline model MoGe. In the second row, we introduce the edge detection module and use detected edges as contact boundaries, bringing consistent improvements in depth estimation since edge cues partially overlap true contact boundaries. For example, the relative error under scale-invariant depth evaluation decreases from **3.46** to **3.42**, while δ_1^d increases from **97.0** to **98.2**. Similar trends hold under affine-invariant metrics, suggesting that contact boundaries provide useful guidance. However, since all detected edges are fused without filtering, noisy and misaligned boundaries introduce global structural inconsistencies. Although depth metrics improve moderately, point-based metrics fluctuate significantly and even degrade. In particular, under scale-invariant point evaluation, the relative error increases sharply from 4.63 to 9.02, indicating severe geometric distortion; similar trends are observed in the affine-invariant setting. These results reveal that naively fusing all contact cues can harm global 3D consistency despite apparent gains in depth accuracy.

Table 4: Ablation study under different depth and point evaluation protocols on iBims-1, analyzing edge detection, contact boundary filtering, and the auxiliary consistency loss ($\mathcal{L}_{\text{cons}}$). Best results are highlighted in **bold**.

Detection	Filtering	$\mathcal{L}_{\text{cons}}$	Scale-invariant Depth Map		Affine-invariant Depth Map		Affine-invariant Disparity Map		Scale-invariant Point Map		Affine-invariant Point Map		Local Point Map	
			$\text{Rel}^d \downarrow$	$\delta_1^d \uparrow$	$\text{Rel}^d \downarrow$	$\delta_1^d \uparrow$	$\text{Rel}^d \downarrow$	$\delta_1^d \uparrow$	$\text{Rel}^p \downarrow$	$\delta_1^p \uparrow$	$\text{Rel}^p \downarrow$	$\delta_1^p \uparrow$	$\text{Rel}^p \downarrow$	$\delta_1^p \uparrow$
✗	✗	✗	3.46	97.0	2.74	97.9	3.23	98.0	4.63	97.1	3.61	97.3	4.16	96.8
✓	✗	✗	3.42	**98.2**	2.70	**98.6**	3.20	**98.2**	9.02	97.1	5.45	96.3	4.05	**97.5**
✓	✗	✓	3.36	97.6	2.68	98.1	3.14	**98.2**	4.70	97.6	3.58	97.8	**4.03**	97.0
✓	✓	✓	**3.35**	97.7	**2.67**	98.1	**3.12**	**98.2**	**4.49**	**97.7**	**3.52**	**98.0**	**4.03**	97.1

During training, the auxiliary consistency loss, $\mathcal{L}_{\text{cons}}$, is introduced to regularize the model toward the original MoGe prediction when contact input is absent. As shown in the third row, point-based metric fluctuations are substantially reduced. Under the scale-invariant point evaluation, the relative error drops from **9.02** to **4.70**, indicating a clear recovery of geometric consistency. Meanwhile, depth performance is preserved and slightly improved, with the scale-invariant depth relative error decreasing from **3.42** to **3.36**. Despite the stabilizing effect on point-based metrics, they remain inferior to the base method, highlighting the need for further filtering.

Finally, an additional filtering module is incorporated to remove unreliable contact regions before fusion. Compared with the third row, both depth and point-based metrics show further improvements. For example, the scale-invariant point map relative error decreases from **4.70** to **4.49**, while depth metrics also exhibit consistent gains across all evaluation settings. These results confirm that removing unreliable contact cues significantly enhances both geometric consistency and depth accuracy, demonstrating the importance of selectively retaining reliable contact regions for robust depth estimation.

5 Conclusion

We have presented a boundary-aware monocular depth estimation framework that explicitly leverages contact boundaries as structural priors to enforce depth continuity at object contact regions. By introducing a contact boundary detection and filtering module to distill contact-aware representations from dense edge maps, and integrating them into the depth decoder via a boundary-aware feature fusion strategy, the proposed framework effectively rectifies the spurious depth discontinuities that persist in both per-pixel regression and 3D point-cloud-based approaches. The framework further accommodates interactive refinement, enabling users to manually delineate missing contact boundaries for targeted depth correction. Comprehensive experiments across three previously unseen benchmarks with dense object interactions validate the efficacy of our approach, which consistently surpasses competing baselines with particularly pronounced improvements at contact regions while requiring only limited training data.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2022YFA1004100, and the Natural Science Foundation of Inner Mongolia under Grant No.2025QN06002.

References

1. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4009–4018 (2021)
2. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
3. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2818–2829 (2023)
4. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
5. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: European Conference on Computer Vision. pp. 241–258. Springer (2024)
6. Garg, R., Bg, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European conference on computer vision. pp. 740–756. Springer (2016)
7. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
8. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
9. Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast generative monocular depth estimation with flow matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3203–3211 (2025)
10. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2485–2494 (2020)
11. He, J., Zhang, S., Yang, M., Shan, Y., Huang, T.: Bi-directional cascade network for perceptual edge detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3828–3837 (2019)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

14. Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., et al.: Is my depth ground-truth good enough? hammer–highly accurate multi-modal dataset for dense 3d scene regression. arXiv preprint arXiv:2205.04565 (2022)
15. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
16. Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In: Proceedings of the European conference on computer vision (ECCV). pp. 573–590 (2018)
17. Koch, T., Liebel, L., Körner, M., Fraundorfer, F.: Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. Computer Vision and Image Understanding **191**, 102877 (2020)
18. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)
19. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018)
20. Liu, Y., Cheng, M.M., Hu, X., Wang, K., Bai, X.: Richer convolutional features for edge detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3000–3009 (2017)
21. Patni, S., Agarwal, A., Arora, C.: Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28285–28295 (2024)
22. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. arXiv preprint arXiv:2502.20110 (2025)
23. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10106–10116 (2024)
24. Qi, X., Liu, Z., Liao, R., Torr, P.H., Urtasun, R., Jia, J.: Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(2), 969–984 (2020)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
26. Ramamonjisoa, M., Lepetit, V.: Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
27. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
28. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer.

IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
29. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021)
30. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
33. Wang, P., Shen, X., Russell, B., Cohen, S., Price, B., Yuille, A.L.: Surge: Surface regularized geometry estimation from a single image. Advances in Neural Information Processing Systems **29** (2016)
34. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5261–5271 (2025)
35. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
36. Wofk, D., Ma, F., Yang, T.J., Karaman, S., Sze, V.: Fastdepth: Fast monocular depth estimation on embedded systems. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 6101–6108. IEEE (2019)
37. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
38. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
39. Yang, N., Stumberg, L.v., Wang, R., Cremers, D.: D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1281–1292 (2020)
40. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: Lego: Learning edge with geometry all at once by watching videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 225–234 (2018)
41. Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R.: Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
42. Yin, W., Wang, X., Shen, C., Liu, Y., Tian, Z., Xu, S., Sun, C., Renyin, D.: Diversedepth: Affine-invariant depth prediction using diverse data. arXiv preprint arXiv:2002.00569 (2020)
43. Yin, W., Zhang, J., Wang, O., Niklaus, S., Chen, S., Liu, Y., Shen, C.: Towards accurate reconstruction of 3d scene shape from a single monocular image. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**, 6480–6494 (2022)
44. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 204–213 (2021)

45. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1983–1992 (2018)
46. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
47. Zhou, C., Huang, Y., Pu, M., Guan, Q., Deng, R., Ling, H.: Muge: Multiple granularity edge detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25952–25962 (2024)
48. Zhou, C., Huang, Y., Pu, M., Guan, Q., Huang, L., Ling, H.: The treasure beneath multiple annotations: An uncertainty-aware edge detector. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 15507–15517 (2023)
49. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017)