HYDEN: A HYBRID DUAL-PATH ENCODER FOR MONOCULAR GEOMETRY OF HIGH-RESOLUTION IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a **hy**brid **d**ual-path vision **en**coder (Hyden) for high-resolution monocular depth, point map and surface normal estimation, surpassing state-of-the-art accuracy with a fraction of the inference cost. The architecture pairs a low-resolution Vision Transformer branch for global context with a full-resolution CNN branch for fine details, fusing features via a lightweight MLP before decoding. By exploiting the linear scaling of CNNs and constraining transformer computation to a fixed resolution, the model delivers fast inference even on multi-megapixel inputs. To overcome the scarcity of high-quality high-resolution supervision, we introduce a self-distillation framework that generates pseudo-labels from existing models at both lower resolution full images and high-resolution crops—global labels preserve geometric accuracy, while local labels capture sharper details. To demonstrate the flexibility of our approach, we integrate Hyden and our self-distillation method into DepthAnything-v2 for depth estimation and MoGe2 for surface normal and metric point map prediction, achieving state-of-the-art results on high-resolution benchmarks with the lowest inference latency among competing methods.

1 INTRODUCTION

Monocular depth, pointmap, and surface normal estimation are core to 3D perception in driving, robotics, and mixed reality. Models like MiDaS Ranftl et al. (2020) and DepthAnything Yang et al. (2024) show strong results from a single RGB image, but most are trained at low resolution, causing degraded predictions on megapixel inputs Wang et al. (2025).

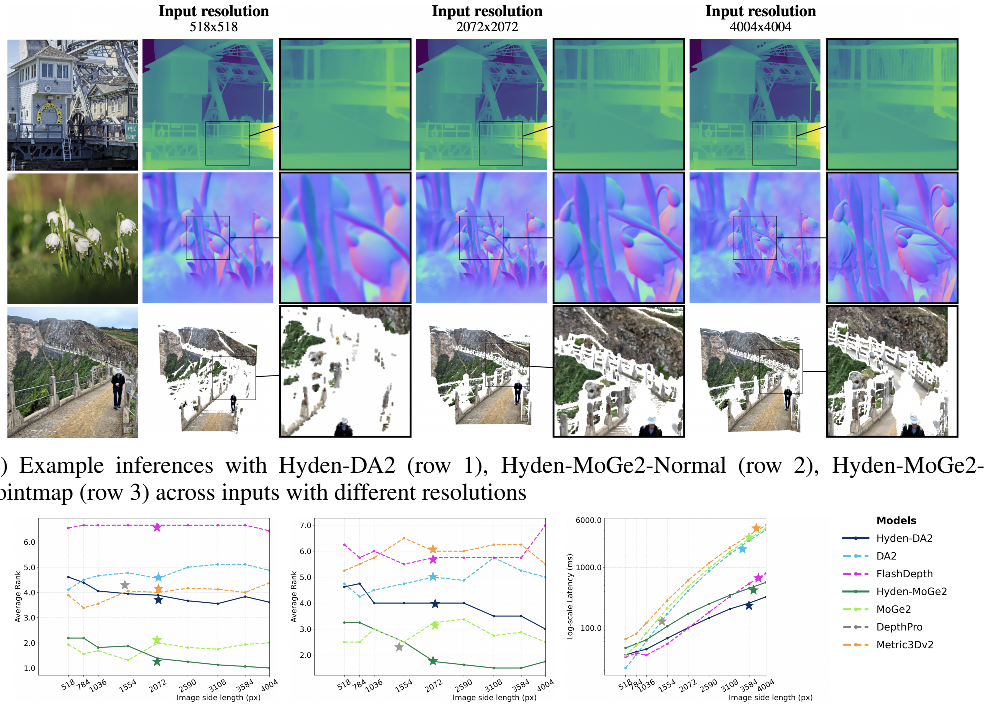
To close this gap, recent work partitions images into tiles and blends ViT features (Depth-Pro Bochkovskii et al. (2024), PatchFusion Li et al. (2024b)) or designs multi-branch ViTs (FlashDepth Chou et al. (2025)). Yet, ViT inference scales quadratically with resolution. Supervision is also problematic: real high-res supervisions are often noisy or sparse, while synthetic labels are perfect but introduce domain gaps, making model generalization difficult.

To address these issues we present Hyden—a Hybrid Dual-path Encoder coupling a full-resolution CNN with a low-resolution ViT. CNN features preserve local detail, while upscaled ViT tokens provide global context, fused through lightweight layers before task-specific decoding. This design substantially lowers inference latency while maintaining sharp predictions.

For supervision, Hyden uses self-distillation: unlabeled high-resolution images are pseudo-labeled by a frozen teacher at 518×518 for both full images and high-res crops. The original ViT branch is kept frozen, and only the CNN branch, fusion layer, and decoder are optimized using both a global loss (on the downsampled full image) and a local crop loss (on masked crop regions).

Our contributions is summarized as follows:

- We introduce Hyden, the first encoder that combines a fixed-resolution ViT for global context and a full-resolution CNN for fine detail, significantly reducing inference cost while preserving high-resolution accuracy.



(a) Example inferences with Hyden-DA2 (row 1), Hyden-MoGe2-Normal (row 2), Hyden-MoGe2-Pointmap (row 3) across inputs with different resolutions

(b) Geometry Accuracy (↓) (c) Geometry Sharpness (↓) (d) Latency (log-scale) (↓)

Figure 1: **Performance comparison across inference resolutions for geometric foundation models:** (a) Example inferences with relative depth, surface normal and point map prediction models illustrate the tradeoff between latency and sharpness across resolutions. (b) shows average ranking across datasets for relative depth accuracy, (c) reports average ranking for geometry sharpness across depth models, and (d) plots inference latency (**log-scale**) measured on an NVIDIA A100 GPU with FP16 precision. **Lower is better for all plots.** Compared to base models, both Hyden-DA2 and Hyden-MoGe2 achieve improved accuracy at high-resolution inference and deliver significant inference speedups. Hyden-MoGe2 achieves the best geometry accuracy and sharpness compared to other state-of-the-art models and consumes significantly lower inference latency. (*DepthPro is evaluated at a fixed resolution due to model constraint.)

- We propose a self-distillation framework that uses global pseudo-labels to preserve accurate geometry across resolutions and local pseudo-labels to capture sharper fine details in high-resolution predictions.
- By integrating the Hyden encoder into two leading models: DepthAnything-v2 Yang et al. (2024) and MoGe2 Wang et al. (2025)—our approach establishes new state-of-the-art performance for high-resolution depth, point map, and surface-normal prediction, while achieving average 3x lower inference latency at 2K and nearly 10× lower at 4K resolution compared to the original models (see Figure 1).

2 Related work

2.1 Zero-shot Monocular Geometry Estimation

Traditional monocular models Bhat et al. (2021); Eigen et al. (2014); Li et al. (2022); Eigen & Fergus (2015); Saxena et al. (2008) were trained on single datasets for specific domains (e.g., indoor or street-view) and generalized poorly due to limited diversity and fixed camera setups.

Relative depth To improve generalization, MegaDepth Li & Snavely (2018) and DiverseDepth Yin et al. (2020) scaled supervision with Internet-scale data. MiDaS Ranftl et al. (2020) introduced scale- and shift-invariant losses, later extended with transformers Ranftl et al. (2021);

Birkl et al. (2023). DepthAnything Yang et al. (2024) distilled pseudo labels for 62M images, while generative priors adapted diffusion models Ke et al. (2024); Rombach et al. (2022) or joint attention Fu et al. (2024). These methods generalize broadly but remain limited by scale/shift ambiguity.

Metric depth Scarce metric annotations hinder absolute scale. ZoeDepth Bhat et al. (2023) fine-tunes metric heads on relative models. Metric3D Hu et al. (2024) and DepthPro Bochkovskii et al. (2024) resolve cross-camera ambiguity via canonical transformations. UniDepth Piccinelli et al. (2025) learns implicit camera models, while MoGe2 Wang et al. (2025) predicts scale-invariant pointmaps with scale recovery.

Metric point map Another direction is predicting 3D pointmaps. Many works Yin et al. (2021); Piccinelli et al. (2025); Hu et al. (2024); Bochkovskii et al. (2024) decouple depth and camera recovery, e.g., LeRes Yin et al. (2021) regresses depth and intrinsics, UniDepth Piccinelli et al. (2024) uses camera embeddings. DUSt3R Wang et al. (2024) predicts stereo pointmaps end-to-end, and MoGe2 Wang et al. (2025) combines scale-invariant maps with scale factors.

Despite progress, most models are trained at low resolution ($\leq 518 \times 518$), losing detail when downsampled and incurring high cost at full scale.

2.2 Zero-shot Surface Normal Estimation

Normals avoid metric ambiguity and capture local shape for localization Behley & Stachniss (2018), mapping Wang et al. (2019), and reconstruction Yu et al. (2022); Wang et al. (2022). Early work derived them from RGB-D scans Silberman et al. (2012); Eigen & Fergus (2015); Qi et al. (2020) and denoised via consistency Qi et al. (2018), adaptive constraints Long et al. (2024; 2021), or uncertainty Bae et al. (2021). OmniData Eftekhar et al. (2021) scaled to 1.3B frames, while Normal-in-the-Wild Chen et al. (2017) expanded to outdoor scenes. DSINE Bae & Davison (2024) introduced a normals-specific architecture, and recent transformer-based approaches Hu et al. (2024); Wang et al. (2025) unify depth, normals, and pointmaps. Yet most remain constrained to low-resolution training, reducing sharpness.

2.3 High-Resolution Depth and Surface Normal Estimation

To recover fine details, SMD-Net Tosi et al. (2021) and Poisson-fusion Dai et al. (2023); Li et al. (2024a) sharpen boundaries, while patch pipelines—BoostingDepth Miangoleh et al. (2021), PatchFusion Li et al. (2024b), PatchRefiner Li et al. (2024c)—boost local detail but introduce artifacts and latency. PRO Kwon & Kim (2025) cuts computation but lags end-to-end models Bochkovskii et al. (2024); Chou et al. (2025). DepthPro Bochkovskii et al. (2024) improves patch efficiency, and FlashDepth Chou et al. (2025) uses dual-branch ViTs, though both rely on refinements or synthetic pretraining.

In contrast, our Hyden framework integrates a full-resolution CNN with a low-resolution ViT and self-distillation, leveraging CNNs' linear scaling to produce sharp, efficient, and generalizable predictions for depth and normals at megapixel scales.

3 Approach

3.1 Hybrid Dual-Path Vision Encoder

Our architecture employs a hybrid dual-path encoder that combines a low-resolution Vision Transformer (ViT) with a full-resolution CNN to balance global context and fine-detail preservation (Figure 2). The ViT branch processes a uniformly downsampled input (up to 518×518) to capture long-range dependencies at constant cost, leveraging any pretrained backbone (e.g., DepthAnything Yang et al. (2024)).

In parallel, the CNN branch directly processes the full-resolution image, efficiently extracting high-frequency features such as edges and textures. We adopt a ResNet-like encoder with hierarchical downsampling stages.

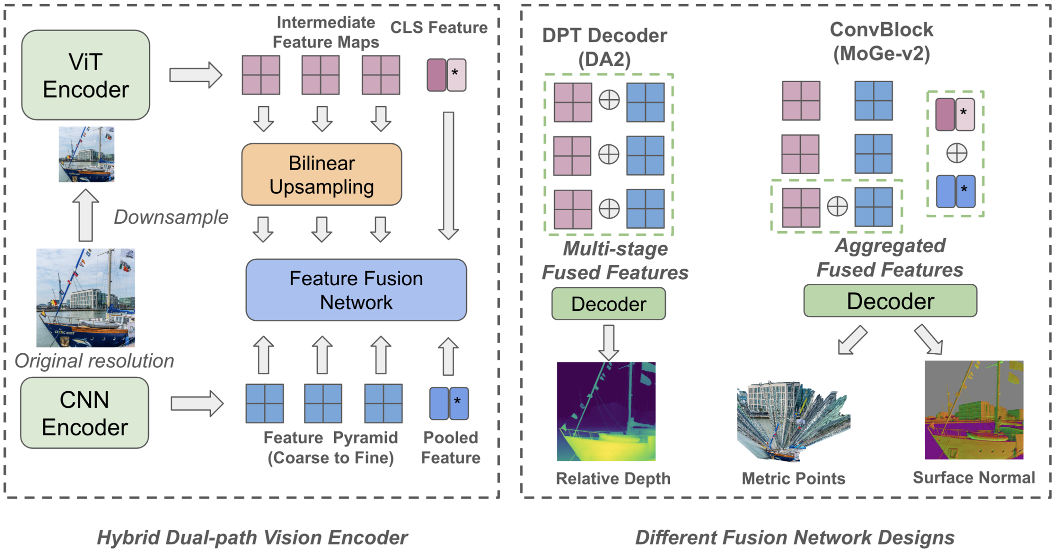


Figure 2: **Network Architecture:** The ViT encoder takes in down-sampled images while the CNN encoder takes in images with original resolution. To recover the high resolution features, the target ViT feature maps are upsampled with bilinear interpolation. CNN and ViT feature maps will be **concatenated** and fed into a feature fusion network. The fused features are used for down-stream tasks. Depending on the decoder architecture, the fusion logic needs to be slightly modified. For example, DA2 (DepthAnything-v2) uses all intermediate features from ViT and we fuse the corresponding CNN layers for each ViT features. MoGe2 uses an aggregated ViT feature map and similarly we aggregate multi-scale CNN feature maps with upsampling and concatenation, and only fuse the aggregated CNN and ViT features. For global-level feature, we apply average pooling to the CNN maps and concatenate the result with the CLS token for downstream tasks.

The two streams are fused by upsampling ViT features to the CNN resolution, concatenating them, and applying a lightweight two-layer convolution. This enables joint reasoning over global transformer context and local detail with minimal overhead.

A key advantage is scalability: ViT cost remains fixed while CNN scales linearly with resolution, enabling efficient inference on multi-megapixel images—unlike pure-ViT models with quadratic cost. The encoder is modular and task-agnostic, and we integrate it into DepthAnythingV2 Yang et al. (2024) and MoGe-V2 Wang et al. (2025) with minimal modifications to the fusion logic, preserving resolution robustness and efficiency (see supplemental materials).

3.2 Self-distillation Training

Motivation. High-resolution supervision is difficult to obtain in practice: real datasets rarely provide dense, clean depth or surface-normal labels at megapixel scales due to hardware and annotation constraints, while synthetic datasets introduce a domain gap relative to real imagery. To build a general training pipeline that upgrades an existing depth, point map or surface normal model to our hybrid dual-path encoder—and scales gracefully to high-resolution inputs—we introduce a *self-distillation* framework.

Overview. From a set of unlabeled high-resolution images $\{I\}$, we generate pseudo labels with a target model $\mathcal{T}$ (e.g., a strong zero-shot predictor). We extract (i) *global* labels from the downsampled full image (518 × 518), and (ii) *local* labels from 518 × 518 crops, which recover sharper details. Local labels may vary in scale and shift, so we align them to the global prediction before training (Figure 3).

We then replace the target model's encoder with our Hyden encoder, adapt the decoder, and *freeze* the ViT branch—training only the CNN branch, fusion module, and decoder. Since the ViT sees only the downsampled view, it requires no additional fine-tuning.

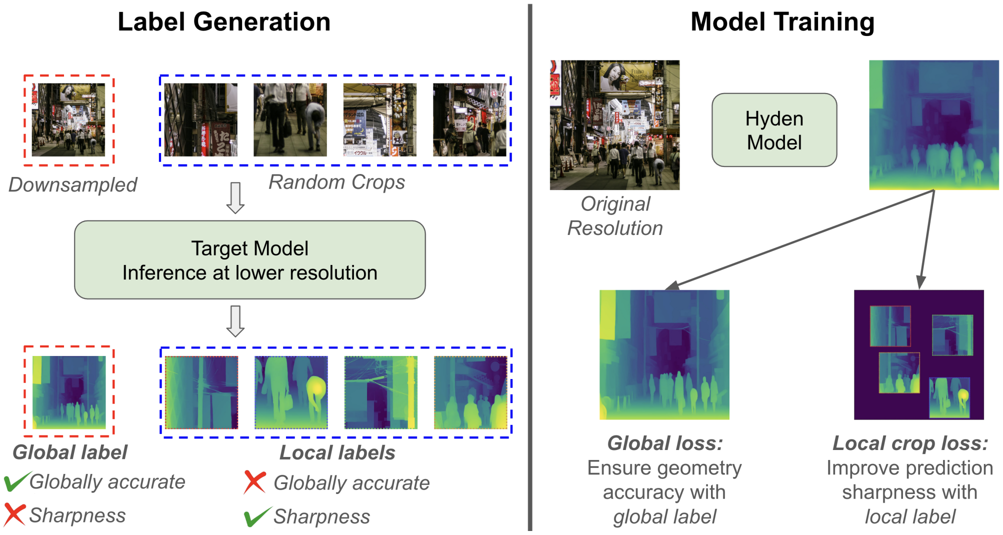


Figure 3: **Self-distillation:** (1) Label generation: Our pipeline samples multiple low-resolution views of the input, using downsampling for global context and random cropping for local details. The target model produces predictions on each view, which are mapped back to the original resolution via up-sampling or indexing. This yields pseudo labels that preserve geometric accuracy and local sharpness. (2) Model training: Using hyden encoders, models are trained on high-resolution inputs at their native resolution. A global loss is applied on downsampled predictions with global labels to retain geometry, while a local loss on full-resolution predictions enhances sharpness.

Notation. Let $I \in \mathbb{R}^{H\times W\times 3}$ be a high-resolution image and S denote the fixed low resolution ($S{=}518$). Denote by $\downarrow_S(\cdot)$ uniform downsampling to $S\times S$, and by $\mathsf{crop}_k(\cdot)$ the k-th high-resolution crop operator with spatial support $\Omega_k \subseteq \{1,\dots,H\}\times\{1,\dots,W\}$; its resized version to $S\times S$ is $\mathsf{rcrop}_k(\cdot)$. The teacher $\mathcal{T}$ produces global pseudo labels $\mathbf{y}_{\mathrm{g}}^{\mathrm{T}} = \mathcal{T}(\downarrow_S(I))$ and local pseudo labels $\mathbf{y}_k^{\mathrm{T}} = \mathcal{T}(\mathsf{rcrop}_k(I))$. Our student (hybrid) network $\mathcal{F}_\theta$ outputs a dense prediction $\mathbf{y} = \mathcal{F}_\theta(I)$ at the native resolution.

3.2.1 Task-specific losses

We use the original task objectives of the target models for both global and local supervision.

Relative depth (scale/shift-invariant). Given predicted depth d and teacher depth $\tilde{d}$ on a pixel set $\mathcal{M}$, we align the prediction by scale and shift

$$a^\star, b^\star = \arg\min_{a,b} \frac{1}{|\mathcal{M}|} \sum_{p\in\mathcal{M}} \left(a\, d_p + b - \tilde{d}_p\right)^2,$$

and compute a robust alignment loss (e.g., ℓ_1):

$$\ell_{\text{depth}}(d, \tilde{d}; \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{p\in\mathcal{M}} \left|a^\star d_p + b^\star - \tilde{d}_p\right| \tag{1}$$

Surface normals (angular/cosine). For predicted unit normals n_p and teacher normals $\tilde{n}_p$,

$$\ell_{\text{normal}}(n, \tilde{n}; \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{p\in\mathcal{M}} \left(1 - \langle n_p, \tilde{n}_p\rangle\right) \tag{2}$$

Affine-invariant point map loss. Let $\hat{\mathbf{P}}_p^{\text{aff}} \in \mathbb{R}^3$ be the predicted affine-invariant point map and $\tilde{\mathbf{P}}_p^{\text{aff}}$ the teacher/GT affine-invariant point map for pixel $p \in \mathcal{M}$. The affine-invariant point map loss is:

$$\ell_{\text{point}}^{\text{aff}}(\hat{\mathbf{P}}^{\text{aff}}, \tilde{\mathbf{P}}^{\text{aff}}; \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{p\in\mathcal{M}} \left\|\hat{\mathbf{P}}_p^{\text{aff}} - \tilde{\mathbf{P}}_p^{\text{aff}}\right\|_1 \tag{3}$$

Scale prediction loss. Given the predicted global scale $\hat{s} > 0$, we supervise it against the optimal alignment scale $s^\star$ between $\hat{\mathbf{P}}^{\text{aff}}$ and the metric GT points $\mathbf{P}_p$:

$$\mathcal{L}_{\text{scale}} = \|\log \hat{s} - \text{stopgrad}(\log s^\star)\|_2^2 \tag{4}$$

Please refer to DA2 Yang et al. (2024) and MoGe2 Wang et al. (2025) for more details in task-specific loss formulation. We write ℓ_{task} to denote either ℓ_{depth} or ℓ_{normal} depending on the task.

3.2.2 Global loss

We downsample the student prediction to $S \times S$ and compare it with the global pseudo label:

$$\mathcal{L}_{\text{global}} = \ell_{\text{task}}\Big(\downarrow_S (\mathbf{y}),\ \mathbf{y}_{\text{g}}^{\text{T}};\ \mathcal{M}_{\text{g}} \Big) \tag{5}$$

where $\mathcal{M}_{\text{g}}$ is the valid-pixel mask at the global resolution (e.g., invalid depth/normal entries removed).

3.2.3 Local crop loss

For each crop k, we compare the high-resolution student prediction with the teacher labels projected back to the crop region. Let $\uparrow_{\Omega_k} (\cdot)$ denote injecting local labels at $S \times S$ to the high-resolution support Ω_k, and M_k be the binary mask of Ω_k:

$$\mathcal{L}_{\text{local}} = \frac{1}{K} \sum_{k=1}^{K} \ell_{\text{task}}\Big(\mathbf{y},\ \uparrow_{\Omega_k} (\mathbf{y}_k^{\text{T}});\ \mathcal{M}_k\Big),$$
$$\text{where} \quad \mathcal{M}_k = M_k \cap \text{valid}\big(\uparrow_{\Omega_k} (\mathbf{y}_k^{\text{T}})\big)\,. \tag{6}$$

This formulation applies the task loss *directly at the native image resolution* but only within each crop's support. The loss is averaged across all crops.

3.2.4 Total objective

The final objective combines global and local terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{g}}\, \mathcal{L}_{\text{global}} + \lambda_\ell\, \mathcal{L}_{\text{local}},$$

with weights $\lambda_{\text{g}}, \lambda_\ell = 1$. In all experiments we freeze the ViT branch and optimize the CNN encoder, fusion layer, and task decoder end-to-end using $\mathcal{L}_{\text{total}}$.

3.3 Implementation Details

We adapt our Hyden architecture to DA2 and MoGe2, denoted as Hyden-DA2 and Hyden-MoGe2. The additional CNN encoder introduces 10M parameters, incurring only a minor computational overhead. During self distillation, we randomly resize the input images from 518 to 2072 resolution for better geometry consistency (see Section 4.2). We use 4 local crops for all our self-distillation experiments. For the unlabeled high-resolution images $\{I\}$, we sampled 50 million images from a publicly available repository of crawled web data and we resized all the images to 2072x2072 resolution. We train our models for 300k iterations with batch size 192 on 64 NVIDIA H100 GPUs. We use an initial learning rate of 1e-5 for CNN encoder and 1e-6 for feature fusion and decoder module. We use adamW Kinga et al. (2015) optimizer and use polynomial learning rate scheduler.

4 Results

4.1 Baseline and Evaluation Metrics

We benchmark Hyden models against DepthAnythingV2 Yang et al. (2024) for relative depth and MoGe2 Wang et al. (2025) for metric-scale pointmaps and normals. We also compare with high-resolution depth methods DepthPro Bochkovskii et al. (2024) and FlashDepth Chou et al. (2025), and with normals-focused DSINE Bae & Davison (2024) and Metric3Dv2 Hu et al. (2024).

Evaluation Metrics For relative depth and metric pointmaps, we evaluate on 9 datasets: NYUv2 Nathan Silberman & Fergus (2012), KITTI Geiger et al. (2013), ETH3D Schops et al. (2017), iBims-1 Koch et al. (2018; 2020), DDAD Guizilini et al. (2020), DIODE Vasiljevic et al. (2019), HAMMER Jung et al. (2023), Booster Ramirez et al. (2022), and Middlebury Scharstein et al. (2014). These span indoor, street-view, and object domains, with ETH3D, Booster, and Middlebury providing 2K+ ground truth for high-resolution evaluation. We report the average relative error for point maps and depth:

$$\mathrm{Rel}_p = \frac{\|\hat{\mathbf{p}} - \mathbf{p}\|_2}{\|\mathbf{p}\|_2}, \quad \mathrm{Rel}_d = \frac{|\hat{z} - z|}{z},$$

along with the percentage of inliers

$$\delta_1^p: \ \frac{\|\hat{\mathbf{p}} - \mathbf{p}\|_2}{\|\mathbf{p}\|_2} < 0.25, \quad \delta_1^d: \ \max\left(\frac{\hat{d}}{d}, \frac{d}{\hat{d}}\right) < 1.25.$$

For surface normal, we evaluate on NYUv2 Nathan Silberman & Fergus (2012), iBims-1 Koch et al. (2018; 2020), Scannet Dai et al. (2017), and vkitti Cabon et al. (2020), reporting mean angular error. For boundary sharpness, we follow MoGe2 Wang et al. (2025) and evaluate on iBims-1, Sintel Butler et al. (2012), HAMMER, and Spring Mehl et al. (2023).

For all evaluation benchmarks, we resize the largest side of the input images to target image resolution and the predictions are evaluated at the original groundtruth resolution.

Table 1: **Zero-shot depth & point map accuracy.** We report the average relative error (lower is better) and δ_1 score per dataset (higher is better) and aggregate performance across datasets via the average rank (lower is better). *DepthPro is evaluated at 1536x1536 and all other models are evaluated with 2K resolution input.

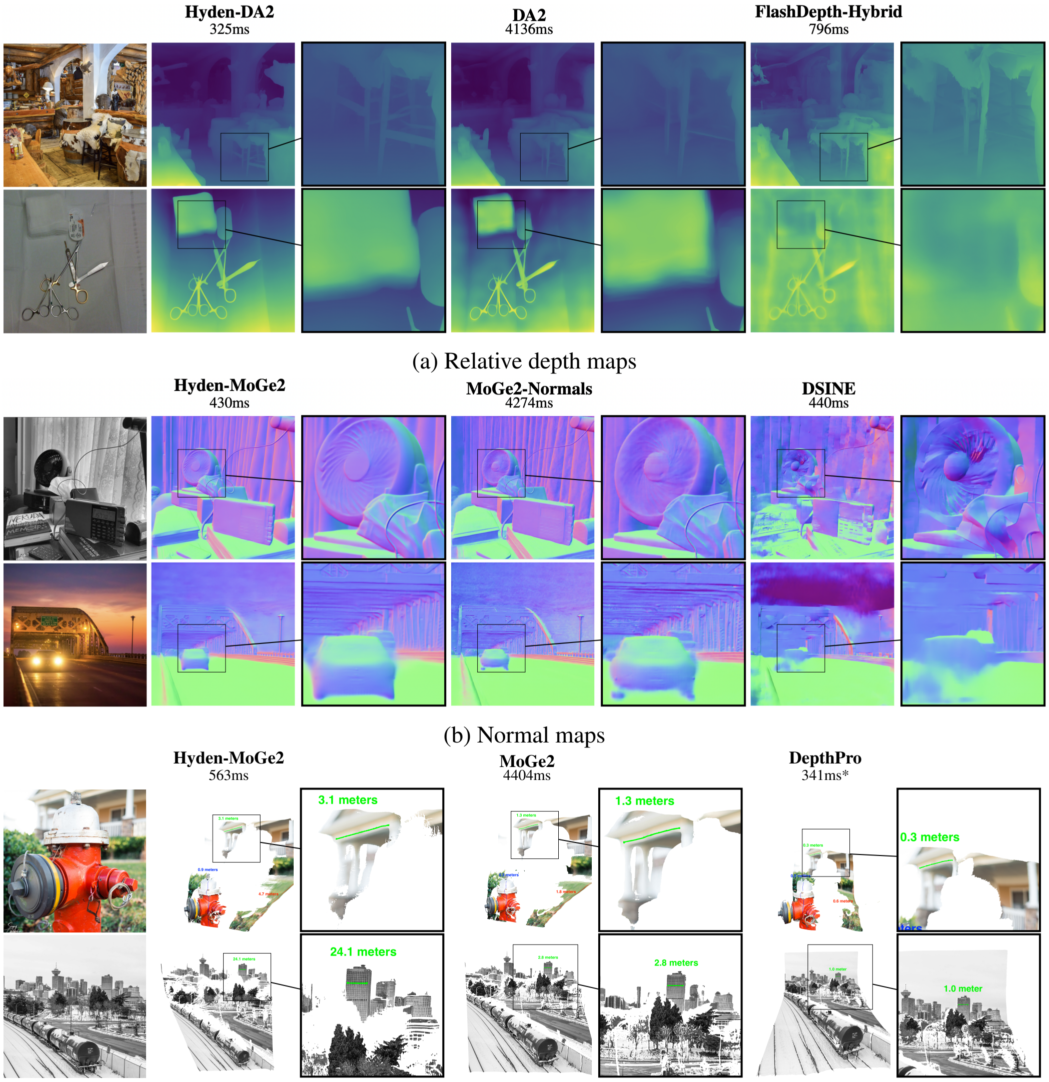
Depth	Inference	**NYUv2**		**KITTI**		**ETH3D**		**iBims-1**		**Booster**		**Middlebury**		**DDAD**		**DIODE**		**HAMMER**		**Avg.**
Model	Latency (ms)	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	**Rank↓**
Relative depth map																				
DA2 Yang et al. (2024)	408.1	5.4	92.3	8.3	92.3	5.5	94.3	4.1	95.5	3.0	98.8	7.3	87.8	15.8	82.4	5.4	95.8	6.3	96.1	4.6
DepthPro Bochkovskii et al. (2024)	341.3*	4.4	96.5	5.7	95.8	7.5	93.1	4.2	96.7	3.2	98.6	9.2	84.8	15.1	80.1	4.9	94.3	5.3	98.3	4.3
FlashDepth Chou et al. (2025)	**98.9**	8.8	90.2	12.0	91.4	8.7	91.2	8.3	87.3	5.5	95.1	11.0	78.9	19.7	75.5	7.9	90.4	7.3	87.1	6.7
Metric3Dv2 Hu et al. (2024)	606.7	5.8	92.1	5.6	95.7	5.6	94.6	5.0	93.1	3.4	98.8	12.6	75.8	**10.8**	**92.7**	**3.4**	**97.9**	3.7	98.5	4.0
MoGe2 Wang et al. (2025)	476.8	3.9	97.3	5.0	96.9	3.8	98.1	3.3	98.2	**2.1**	**99.2**	2.3	94.3	11.3	90.5	3.9	96.7	3.4	98.8	2.0
Hyden-DA2 (Ours)	**100.7**	4.6	96.5	7.6	95.3	5.1	95.8	4.1	97.8	3.0	98.7	10.4	83.2	13.4	85.7	4.7	96.2	5.4	97.3	3.9
Hyden-MoGe2 (Ours)	171.6	**3.7**	**98.5**	**4.9**	**97.8**	3.8	**98.3**	**3.2**	**98.6**	2.1	99.0	**2.0**	**95.8**	11.1	91.6	3.7	97.1	**3.2**	**99.1**	1.3
Metric depth map (w/o GT intrinsics)																				
DepthPro Bochkovskii et al. (2024)	341.3*	11.7	89.7	25.8	34.2	38.1	31.9	16.4	79.4	45.3	38.4	-	-	35.1	36.4	32.7	35.6	39.0	61.7	3.0
MoGe2 Wang et al. (2025)	476.8	9.2	92.5	**15.7**	**87.1**	19.6	82.3	14.8	88.6	24.9	35.6	-	-	25.3	55.6	18.0	69.4	**24.1**	**68.3**	1.8
Hyden-MoGe2 (Ours)	**171.6**	**7.7**	**96.1**	16.4	86.2	**18.3**	**85.7**	**12.4**	**91.2**	**20.2**	**47.3**	-	-	**24.1**	**59.2**	**17.7**	**71.4**	26.0	64.6	1.3
Metric point map																				
DepthPro Bochkovskii et al. (2024)	341.3*	12.1	88.3	26.1	67.8	40.2	61.9	18.9	75.6	72.8	32.8	-	-	36.7	34.4	32.1	34.9	40.3	58.9	3.0
MoGe2 Wang et al. (2025)	476.8	9.7	93.8	**16.8**	**85.7**	20.3	92.6	16.2	84.6	63.3	38.3	-	-	26.5	53.4	18.0	69.6	**25.2**	**69.5**	1.8
Hyden-MoGe2 (Ours)	**171.6**	**8.3**	**95.7**	17.9	83.1	**19.1**	**94.7**	**14.1**	**90.7**	**50.8**	**49.3**	-	-	**25.4**	**57.9**	**17.7**	**72.9**	28.7	66.3	1.3

4.2 PERFORMANCE COMPARISON ON IMAGES WITH DIFFERENT RESOLUTIONS

We first evaluate robustness of test-time input scaling. As illustrated in Figure 1, at low resolution (518×518), DA2 Yang et al. (2024) and MoGe2 Wang et al. (2025) surpass Hyden, but above 784×784 all baselines, including Metric3Dv2 Hu et al. (2024), degrade sharply while Hyden stays accurate. Although FlashDepth Chou et al. (2025) offers relatively low inference latency, its lightweight decoder and limited supervision result in significant performance drops. Hyden models maintain consistent depth and pointmap accuracy across resolutions, with only marginal latency increases. Trained on mixed resolutions (Section 5.2), it combines stability with efficient high-res inference. In particular, Hyden-MoGe2 delivers the best 2K+ accuracy and low latency, and at 4K (Figure 1) outperforms ViT baselines while running 10× faster.

4.3 PERFORMANCE COMPARISON ON HIGH-RESOLUTION IMAGES

Zero-shot depth & point map As shown in Figure 4, Hyden-DA2 yields sharper geometry and Hyden-MoGe2 predict better metric scale at high resolution. In Table 1, Hyden-MoGe2 attains



(a) Relative depth maps

(b) Normal maps

(c) Point maps

Figure 4: Qualitative comparison of geometry predictions on in-the-wild images at 4004 × 4004. Left: Hyden models; Middle: base models (DepthAnythingV2, MoGe2); Right: other SOTA models. Hyden models produces sharper, more accurate geometry with much lower latency. (*DepthPro is evaluated at 1536x1536 resolution due to model constraints.)

the highest average accuracy, while both Hyden variants improve accuracy and reduce latency over their baselines. At 2K, Hyden-MoGe2 achieves the best geometric accuracy and runs ∼3× faster, highlighting the effectiveness of our self-distillation for high-resolution accuracy and efficiency.

Zero-shot surface normal As shown in Table 2, Hyden-MoGe2 improves accuracy on most benchmarks while keeping low latency, with qualitative results in Figure 4. DSINE Bae & Davison (2024) runs at similar speed but trails ViT-based models. In out-of-domain tests, Hyden-MoGe2 surpasses Metric3Dv2 on three benchmarks and significantly outperforms MoGe2 at 2K resolution.

Zero-shot boundary sharpness As shown in Table 3, Hyden models achieve notably higher F1 and recall than their baselines while running faster at 2K. Hyden-MoGe2 further surpasses other SOTA methods without extra high-resolution supervision.

Table 2: **Zero-shot surface normal accuracy.** We report the mean angular errors (lower is better). *ScanNet evaluation is in-domain for Metric3Dv2, as it is included in the training set.

Surface normal Model	Inference Latency (ms)	**NYUv2** Mean↓	**iBims-1** Mean↓	**Scannet** Mean↓	**vkitti** Mean↓	**Avg.** Rank↓
DSINE Bae & Davison (2024)	149.4	17.1	18.0	16.9	30.2	4.0
Metric3Dv2 Hu et al. (2024)	606.7	15.9	15.4	**11.4***	29.6	2.3
MoGe2 Wang et al. (2025)	438.2	15.6	16.0	13.7	27.3	2.5
Hyden-MoGe2 (Ours)	**127.4**	**14.6**	**14.8**	13.0	**27.0**	1.2

Table 3: **Zero-shot boundary sharpness.** We report F1 score and recall for all datasets (higher is better).

Depth	**iBims-1**		**Sintel**		**HAMMER**		**Spring**		**Avg.**
Model	F1↑	R↑	F1↑	R↑	F1↑	R↑	F1↑	R↑	Rank↓
DA2 Yang et al. (2024)	12.7	20.0	28.7	36.4	7.7	13.4	16.3	15.3	5.0
FlashDepth Chou et al. (2025)	11.3	11.2	25.3	28.7	6.0	5.9	20.2	17.6	5.7
Metric3Dv2 Hu et al. (2024)	12.8	13.4	22.7	24.4	4.9	4.0	17.6	14.1	6.0
DepthPro Bochkovskii et al. (2024)	49.2	43.1	40.3	44.1	7.5	7.3	37.1	33.9	2.2
MoGe2 Wang et al. (2025)	49.0	45.6	38.2	41.4	7.4	7.2	34.8	32.5	3.3
Hyden-DA2 (Ours)	15.8	21.3	33.1	46.0	**10.7**	**19.3**	15.9	16.8	4.0
Hyden-MoGe2 (Ours)	**54.7**	**50.4**	**46.5**	**49.6**	7.9	7.6	**34.2**	**29.9**	1.6

Additional results and model size comparisons are provided in the supplemental material.

5 Ablation Study

5.1 Importance of Local Crop Loss

As shown in Table 4 (all models are evaluated with 2K resolution input), removing the local crop loss causes Hyden-DA2 to lose sharpness due to reliance on low-resolution global supervision. Increasing the number of crops improves sharpness, and we find that using four crops offers best trade-off between labeling cost and model performance.

Table 4: **Ablation on local crop loss.** (We report F1 score and recall for zero-shot boundary sharpness evaluation.)

Depth	**iBims-1**		**Sintel**		**HAMMER**		**Spring**	
Model	F1↑	R↑	F1↑	R↑	F1↑	R↑	F1↑	R↑
Hyden-DA2 w/o local crop loss	11.8	18.4	27.9	38.2	7.8	13.1	14.7	13.8
Hyden-DA2 w/ 2 crops	14.4	20.9	31.8	40.5	8.7	16.8	15.5	14.7
Hyden-DA2 w/ 4 crops	15.8	21.3	**33.1**	**46.0**	**10.7**	**19.3**	15.9	16.8
Hyden-DA2 w/ 8 crops	**16.1**	**22.2**	32.3	45.7	10.3	18.9	**17.1**	**18.5**

Table 5: **Ablation on mixed-resolution training & fusion network design.**

Depth Model	**NYUv2** Rel↓	**KITTI** Rel↓	**ETH3D** Rel↓	**HAMMER** Rel↓
Hyden-DA2 trained from 518-1036	5.14	8.83	5.27	7.10
Hyden-DA2 trained from 518-2072	**4.60**	**7.63**	**5.12**	**5.44**
Hyden-DA2 w/ MLP Fusion	4.72	7.92	5.31	5.93
Hyden-DA2 w/ 1-layer CNN Fusion	4.63	7.88	5.22	5.87
Hyden-DA2 w/ 2-layer CNN Fusion	**4.60**	**7.63**	**5.12**	**5.44**

5.2 Importance of Mixed-resolution Training

Table 5 compares Hyden models trained with input resolutions ranging from 518–1036 and 518–2072. Matching the training resolution to the test-time resolution improves depth accuracy, highlighting the value of mixed-resolution training. However, for ViT encoders, training at very high resolutions (e.g., over 20K tokens at 2K resolution) is computationally prohibitive. By constraining the ViT branch to low-resolution input, Hyden enables practical high-resolution training.

5.3 Feature Fusion Network Design

We evaluate several fusion designs for feature projection: a single linear layer, a single CNN layer, and two CNN layers with ReLU activation. As shown in Table 5, the two-layer CNN achieves the best performance. We adopt this configuration in our model as it offers superior accuracy without incurring excessive computational overhead.

6 Conclusion

We presented Hyden, a hybrid dual-path vision encoder that delivers high-resolution depth, point map and surface-normal estimation with low latency. By combining global and local pseudo-label self-distillation, Hyden preserves geometric accuracy while enhancing fine details, without relying on high-resolution ground truth. Integrated into leading baselines, Hyden achieves state-of-the-art accuracy across resolutions while maintaining fast inference, offering a scalable solution for dense prediction tasks.

REFERENCES

Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9545, 2024.

Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13137–13146, 2021.

Jens Behley and Cyrill Stachniss. Efficient surfel-based slam using 3d laser range data in urban environments. In *Robotics: science and systems*, volume 2018, pp. 59, 2018.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4009–4018, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

Aleksei Bochkovskii, AmaÃǤl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.) (ed.), *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, October 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020.

Weifeng Chen, Donglai Xiang, and Jia Deng. Surface normals in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1557–1566, 2017.

Gene Chou, Wenqi Xian, Guandao Yang, Mohamed Abdelfattah, Bharath Hariharan, Noah Snavely, Ning Yu, and Paul Debevec. Flashdepth: Real-time streaming video depth estimation at 2k resolution. *arXiv preprint arXiv:2504.07093*, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Yaqiao Dai, Renjiao Yi, Chenyang Zhu, Hongjun He, and Kai Xu. Multi-resolution monocular depth map fusion by self-supervised gradient-based composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 488–496, 2023.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10786–10796, 2021.

David Eigen and Rob 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*, pp. 2650–2658, 2015.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 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.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. 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.

HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, et al. On the importance of accurate geometry data for dense 3d vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 780–791, 2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 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.

Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5. California;, 2015.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. *Computer Vision and Image Understanding*, 191:102877, 2020.

Byeongjun Kwon and Munchurl Kim. One look is enough: A novel seamless patchwise refinement for zero-shot monocular depth estimation models on high-resolution images. *arXiv preprint arXiv:2503.22351*, 2025.

Jiaqi Li, Yiran Wang, Jinghong Zheng, Zihao Huang, Ke Xian, Zhiguo Cao, and Jianming Zhang. Self-distilled depth refinement with noisy poisson fusion. *Advances in Neural Information Processing Systems*, 37:69999–70025, 2024a.

Zhengqi Li and Noah Snavely. 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.

Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022.

Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10016–10025, 2024b.

Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchrefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. In *European Conference on Computer Vision*, pp. 250–267. Springer, 2024c.

Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12849–12858, 2021.

Xiaoxiao Long, Yuhang Zheng, Yupeng Zheng, Beiwen Tian, Cheng Lin, Lingjie Liu, Hao Zhao, Guyue Zhou, and Wenping Wang. Adaptive surface normal constraint for geometric estimation from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (9):6263–6279, 2024.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9685–9694, 2021.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025.

Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 283–291, 2018.

Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip HS Torr, Raquel Urtasun, and Jiaya Jia. 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.

Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21168–21178, 2022.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pp. 31–42. Springer, 2014.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas 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*, pp. 3260–3269, 2017.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8942–8952, 2021.

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

Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European conference on computer vision*, pp. 139–155. Springer, 2022.

Kaixuan Wang, Fei Gao, and Shaojie Shen. Real-time scalable dense surfel mapping. In *2019 International conference on robotics and automation (ICRA)*, pp. 6919–6925. IEEE, 2019.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

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

Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 204–213, 2021.

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022.