2K Retrofit: Entropy-Guided Efficient Sparse Refinement for High-Resolution 3D Geometry Prediction

Tianbao Zhang[1,3] Zhenyu Liang[2] Zhenbo Song[5] Nana Wang[1]
Xiaomei Zhang[5] Xudong Cai[5] Zheng Zhu[4] Kejian Wu[5]
Gang Wang[5] Zhaoxin Fan[1]✉

[1] Beihang University, Beijing, China
[2] National University of Defense Technology, Changsha, China
[3] Dim12 AI Inc, China
[4] GigaAI, China
[5] OpenHelix Robotics, China

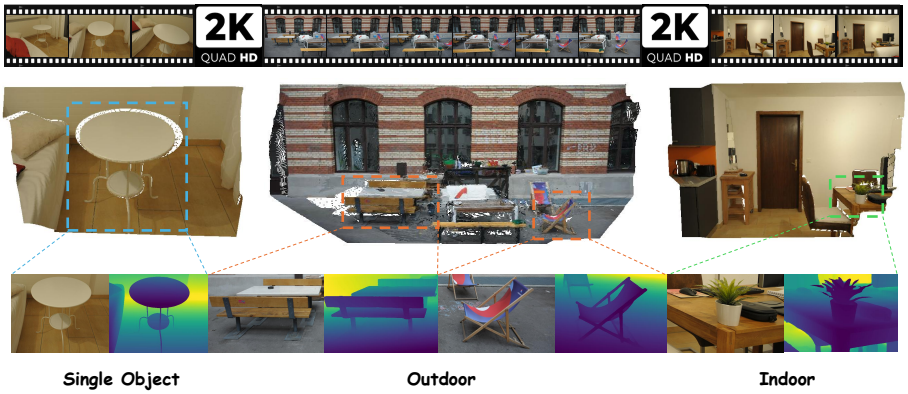


Fig. 1: 2K Retrofit efficiently enhances high-resolution geometric prediction through an entropy-guided sparse refinement strategy. Given 2K-resolution RGB inputs from diverse scenes (single-object, indoor, and outdoor), our framework first performs fast coarse geometric inference using a frozen low-resolution foundation model, then selectively refines only high-uncertainty pixels at full resolution. This design preserves global geometric consistency while recovering fine-scale 3D structures efficiently, achieving accurate and scalable 2K-level depth and pointmap estimation.

Abstract. High-resolution geometric prediction is fundamental for robust perception in autonomous driving, robotics, and augmented/mixed reality. Existing 3D foundation models demonstrate strong generalization and scalability across diverse real-world scenarios. However, we observe that their capability to handle real-world high-resolution (2K-level) scenes remains constrained by the prohibitive computational and memory requirements of dense inference, making large-scale deployment impractical. To address this limitation, we propose 2K Retrofit—to the

✉ Corresponding author.

best of our knowledge, the first framework enabling existing geometric foundation model to perform efficient 2K-resolution inference without backbone modification or retraining. Our method adopts a simple yet effective sparse refinement strategy, where a frozen foundation model provides fast coarse geometric predictions, and an entropy-guided sparse refinement selectively enhances high-uncertainty regions to restore fine-scale 3D structure. This design preserves the global consistency of the backbone while delivering precise, high-fidelity geometry at a fraction of the cost. Extensive experiments on standard benchmarks, including ARKitScenes, ScanNet++, and ETH3D, demonstrate that 2K Retrofit consistently achieves state-of-the-art accuracy and efficiency, bridging the gap between foundation model research and scalable high-resolution 3D vision applications. Code will be released upon acceptance.

Keywords: High resolution · Sparse refine · Foundation model

1 Introduction

High-resolution geometric prediction is fundamental for robust perception and interaction in modern applications such as autonomous driving [1, 11], embodied intelligence [52, 68], and augmented reality [45], since fine-grained depth maps and 3D reconstructions serve as the foundation for precise scene understanding, object localization, and manipulation. However, the spatial resolution of depth sensors in practical settings remains a key bottleneck. This inherent resolution imbalance in the captured data leads to the loss of fine-grained geometric cues that are vital for accurate 3D perception. Such resolution disparity thus may severely restrict systems' ability to perceive small-scale structures—such as railings, handles, or cables along corridors—which, from another perspective, are essential for safe navigation and precise interaction.

Addressing this bottleneck requires 3D geometric model capable of reconstructing high-fidelity 3D structure that lies beyond the native limits of current depth sensors. Although recent approaches have improved geometric prediction through advanced upsampling strategies and model optimization, their inference remains constrained to moderate resolutions (typically around 960×480), leaving true 2K-level prediction an open challenge. Meanwhile, as can be seen, though recent foundation models such as Depth Anything V1/V2 [77, 78], DUSt3R [72], VGGT [70], and others [3, 6, 27, 38, 75] trained on large-scale datasets exhibit strong generalization across diverse scenarios. They are typically limited by low-resolution training data and cannot directly deliver precise geometric predictions at high resolution. To overcome this limitation, recent works attempt to imrpove the performance through explicit high-resolution modeling. As illustrated in Fig. 2, existing approaches fall into three main categories: (1) methods that directly train native high-resolution geometric models [21, 45, 66], aiming to learn fine-grained structure through full-resolution supervision; (2) methods that perform low-resolution inference followed by upsampling or depth super-resolution [62, 67], leveraging learned priors to enhance spatial detail; and (3)

methods that apply patch-wise refinement on top of coarse predictions [34, 35], focusing computation on local regions to recover high-frequency geometry. However, as discussed, existing methods still fail to achieve efficient and accurate 2K-level 3D geometric prediction that our approach provides.

To address this challenge, we introduce 2K Retrofit—a simple yet effective paradigm that seamlessly augments existing foundation models to produce 2K-resolution geometric outputs, without any modification to the original low-resolution geometric backbone (e.g., Depth Anything, DUSt3R, or VGGT). The central insight underpinning our approach is that the discrepancies between dense low-resolution predictions and their high-resolution counterparts are concentrated in a sparse subset of pixels. By strategically directing refinement only to these high-impact regions, we can unlock precise and reliable high-resolution geometric prediction, while maintaining minimal computational and memory overhead. As shown in Fig. 1, 2K Retrofit enables efficient high-resolution geometric prediction across diverse scenarios, including single-object, indoor, and outdoor environments.

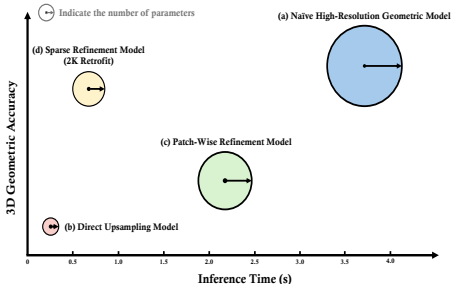


Fig. 2: Differences of 2K Retrofit with exising high-resolution geometry pipelines. (a) Native high-resolution geometric models train large networks directly on 2K data, achieving strong fidelity but with high memory and training cost. (b) Direct upsampling models are lightweight and fast but struggle to recover fine 2K structures. (c) Patch-wise refinement models require multiple low-resolution inferences and complex patch fusion, leading to slow inference and spatial inconsistencies. (d) 2K Retrofit (ours) provides efficient inference and competitive 2K geometric fidelity with far fewer parameters.

Specifically, 2K Retrofit incorporates an entropy-based pixel selection mechanism that identifies high-uncertainty regions for targeted refinement. These sparse, informative pixels are then processed by a dedicated refinement branch for feature extraction and correction, thereby circumventing the computational burden of full-resolution processing. To facilitate effective training, we further build a large-scale synthetic dataset containing 50,000 images and empirically evaluate our approach on two fundamental 3D vision tasks: depth estimation and 3D reconstruction. Extensive experiments on ScanNet++ [81], ARKitScenes [25], and ETH3D [65] demonstrate that 2K Retrofit achieves state-of-the-art accuracy

while maintaining outstanding inference efficiency, surpassing existing methods in both precision and speed. Our contributions can be summarized as follows:

- We propose 2K Retrofit, a simple yet effective method that extends 3D foundation models to 2K-resolution geometric prediction via sparse refinement, which matches or surpasses the accuracy of full 2K methods, while offering significantly faster inference.
- We introduce sparse geometry refinement, which comprises an entropy-based selector for identifying high-uncertainty pixels, a sparse feature extractor for refinement feature learning, and a gated ensembler for final fusion and refinement.
- We conduct extensive evaluations on multiple public high-resolution benchmarks demonstrating the superiority of our approach. Across diverse indoor and outdoor scenarios, our method achieves the best geometric fidelity and more reliably recovers fine-grained 3D structures compared to existing state-of-the-art methods.

2 Related Work

2.1 High-Resolution Monocular Depth Estimation

Monocular depth estimation seeks to infer scene geometry from a single RGB image, enabling a wide range of 3D vision applications. While deep learning has driven rapid progress, early methods [10, 12] were limited by poor generalization across datasets. Later approaches improve robustness via diverse training data [33,71], affine-invariant losses [59], and stronger architectures [58]. Recently, latent diffusion models [28,63] have shown impressive generalization for relative depth, yet are fundamentally limited by input resolution [2, 77], lagging behind the demands of modern high-resolution imaging. To address this gap, Guided Depth Super-Resolution (GDSR) [26, 48, 83] and tile-based methods [34, 49, 61] estimate high-resolution depth by processing image patches independently and stitching results. However, these methods often introduce boundary artifacts and redundant computation. In this paper, we propose a sparse refinement strategy that further improves both accuracy and efficiency.

2.2 Multi-View Reconstruction

Multi-View Stereo (MVS) targets dense 3D reconstruction of real-world scenes from multiple overlapping images, providing essential geometric understanding for applications in robotics, mapping, and virtual reality. Conventional MVS pipelines typically rely on known or estimated camera parameters from Structure from Motion (SfM), and can be broadly classified into handcrafted methods [15,16,64], global optimization strategies [14,51,74], and deep learning-based techniques [18, 47, 57, 79]. Recent learning-based approaches, such as MVSNet and its variants [7, 8, 27],, have significantly advanced reconstruction quality, often requiring depth super-resolution to obtain denser point clouds. Despite this

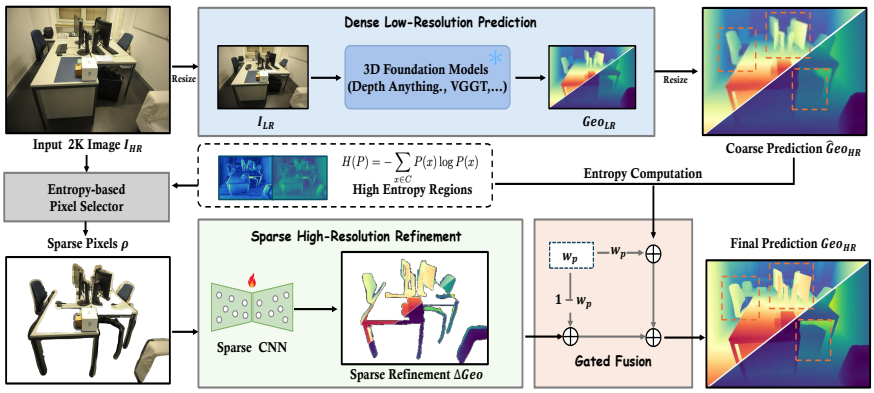


Fig. 3: Framework of 2K Retrofit. Given an HR input image I_{HR}, we first downsample it and feed the result to a frozen 3D foundation model, upsampling its output to obtain a coarse $\hat{Geo}_{\mathrm{HR}}$. An entropy-based selector then identifies a sparse set of high-uncertainty HR pixels, for which a lightweight sparse extractor predicts localize refinements ΔGeo. A gated ensembler merges the coarse map with ΔGeo, yielding high-fidelity Geo_{HR} while remaining low-latency, and memory-efficient.

progress, most methods still depend on accurate camera calibration, which can limit their flexibility in unconstrained settings. Emerging feed-forward architectures like DUSt3R [72], VGGT [70], Dense3R [13] and Pi3 [73] break this dependency, directly predicting dense, aligned point clouds from just a pair of images without explicit camera parameters, thus broadening the applicability of MVS. In contrast to existing high-resolution pipelines, we introduce a model-agnostic sparse refinement framework that seamlessly integrates with modern 3D foundation models to enable efficient and accurate 2K-resolution reconstruction.

2.3 Sparse Inference in Computer Vision

Sparse inference—where computation is selectively applied to only a subset of critical activations or parameters—has emerged as an effective strategy for reducing computational cost in computer vision. This selective computation arises naturally in various visual domains, such as videos [53, 54], point clouds [20, 43], and masked images for self-supervised pre-training [17, 22, 23]. It can also be induced through techniques like activation pruning [30, 37, 55, 60], token merging [4, 5], or clustering [46]. However, most activation-sparsity approaches are tailored for classification or detection tasks, where maintaining complete spatial structure is less important than capturing global semantics. Beyond feature sparsity, another complementary line of work [29, 82] focuses on spatially selective processing, often referred to as selective refinement, which progressively enhances uncertain or structurally critical regions. In segmentation, for instance,

methods such as PointRend [29] and RefineMask [82] identify ambiguous pixels at coarse resolutions and selectively refine them using higher-resolution features, significantly improving mask quality. SparseRefine [44] is the first work to explicitly introduce the concept of sparse refinement, demonstrating that selectively refining ambiguous regions can significantly improve prediction quality. These approaches have shown strong results in segmentation and other 2D prediction tasks, but they are rarely applied to geometric reasoning. Extending selective refinement to 3D geometry remains underexplored. Inspired by SparseRefine [44], we take a new direction by extending the concept of sparse inference beyond recognition tasks and integrating selective refinement into high-resolution 3D geometry prediction.

3 Methodology: 2K Retrofit

3.1 Overview

We target high-quality geometric prediction—such as depth estimation and 3D reconstruction—from high-resolution (2K) images $\mathbf{I}_{\text{HR}} \in \mathbb{R}^{H \times W \times 3}$. Direct dense inference at this resolution is computationally expensive for existing geometric foundation models. As shown in Fig. 4, prediction errors at high resolution are typically concentrated on sparse, semantically critical regions (e.g., object boundaries and small structures). This observation motivates a sparse refinement strategy that focuses computation on these regions.

To this end, we propose 2K Retrofit, a framework applicable to both monocular and multi-view settings and compatible with diverse geometric backbones for depth and point map prediction. As illustrated in Fig. 3, the pipeline consists of two stages: (i) coarse geometric prediction and (ii) sparse high-resolution refinement.

Specifically, the high-resolution image $\mathbf{I}_{\text{HR}}$ is first downsampled to a tractable resolution (e.g., 256 pixels on the longer side) and processed by a frozen backbone $\mathcal{F}$ [72, 78]. The prediction is then upsampled to 2K resolution via nearest-neighbor interpolation, producing a dense coarse estimate $\hat{\mathbf{Y}}_{\text{HR}}$.

An entropy-based selector $\mathcal{S}$ computes an uncertainty map and selects high-uncertainty pixels $\mathcal{P}$ (threshold α). For each $p \in \mathcal{P}$, a lightweight refinement module $\mathcal{R}$ predicts a local correction from the corresponding high-resolution patch. The correction is fused with the coarse estimate through a gated module $\mathcal{G}$ that adapts weights according to prediction confidence.

The full pipeline—$\mathcal{F}$, $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{G}$—is used in both training and inference. With $\mathcal{F}$ frozen, 2K Retrofit enables 2K inference without retraining and generalizes to monocular or multi-view inputs with different backbones (e.g., Depth Anything [78], VGGT [70]).

3.2 Dense Low-Resolution Geometric Prediction

A naïve approach to high-resolution (2K resolution in this paper) geometric prediction would be to directly adapt existing foundation models [70, 78] to

process 2K-scale inputs. However, such a strategy is fundamentally constrained by both the immense computational overhead and the limited generalization ability of current models, which are rarely trained or optimized for ultra-high-resolution data.

To tackle the issue, we adopt a principled two-stage paradigm. Rather than relying on resource-intensive dense prediction at full resolution, we first perform geometric inference on a substantially downsampled version of the input. This low-resolution prediction stage offers high computational efficiency. It also can well leverage the strong global priors learned by foundation models. Hence, the resulting coarse geometric map effectively captures the global structure and semantics of the scene, serving as a robust initialization for subsequent processing.

Specifically, let the high-resolution input image be denoted as $\mathbf{I}_{\text{HR}} \in \mathbb{R}^{H \times W \times 3}$, where H, W indicate the spatial resolution. We first downsample $\mathbf{I}_{\text{HR}}$ to obtain a low-resolution image $\mathbf{I}_{\text{LR}} \in \mathbb{R}^{h \times w \times 3}$, where $h \ll H$ and $w \ll W$. A frozen geometric model $\mathcal{F}$ is then applied to produce a dense low-resolution prediction:

$$\mathbf{Y}_{\text{LR}} = \mathcal{F}(\mathbf{I}_{\text{LR}}). \tag{1}$$

This prediction is then upsampled to the original resolution via nearest-neighbor interpolation, yielding a coarse high-resolution estimate $\hat{\mathbf{Y}}_{\text{HR}} \in \mathbb{R}^{H \times W \times C}$, where C denotes the output dimension (e.g., $C = 1$ for monocular depth, $C = 3$ for 3D point maps).

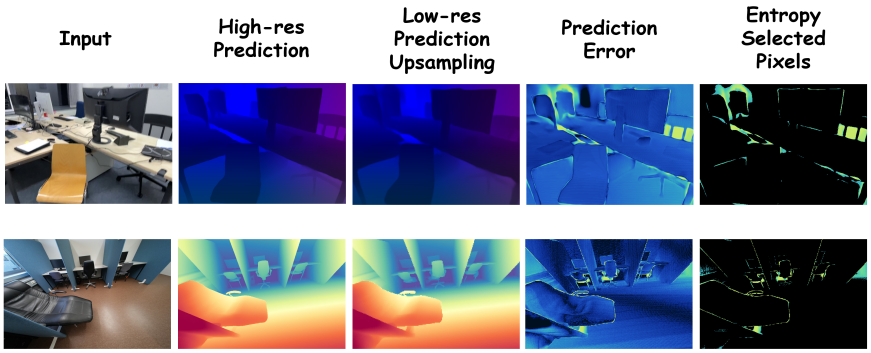


Fig. 4: Analysis of the correlation between prediction error and entropy-based uncertainty, which confirms a strong correspondence between the sparse pixels that we selected and the error map.

3.3 Sparse High-Resolution Geometric Refinement

While low-resolution predictions deliver substantial gains in computational efficiency, they inevitably struggle to capture fine-grained geometric details, particularly around distant objects, intricate structures, and semantic boundaries.

Our systematic error analysis reveals a clear pattern: the vast majority of errors in coarse predictions are not randomly distributed, but instead cluster within sparse, semantically critical regions—such as object contours, thin structures, and areas of category transition (see Fig. 4). These localized regions play an outsized role in scene understanding; thus, selectively improving predictions in these areas can yield significant quality gains with only marginal additional computational cost.

Building on this observation, we posit that high-fidelity 2K geometric prediction can be achieved by selectively refining only those regions with high error. To this end, we design a targeted refinement approach comprising three core components: an entropy-based selector, a sparse feature extractor, and a gated ensembler. We elaborate on each module in the following sections.

(i) Entropy Selector. The first step of our refinement pipeline is to identify a sparse subset of pixels that are most in need of correction. Locating these high-error pixels is non-trivial. Fortunately, as shown in Fig. 4, the regions with the largest prediction errors align closely with high-entropy responses. This strong correlation allows us to use entropy as a principled signal for selecting pixels that truly require refinement. We leverage prediction entropy as a principled measure of model uncertainty, following the intuition that regions with high entropy often correspond to visually or geometrically ambiguous areas such as object boundaries, thin structures, or depth discontinuities. Importantly, since the backbone models are regression models, entropy is *not* computed from the final regression outputs. Instead, we compute entropy from the backbone head features before the final regression projection.

Specifically, based on the coarse prediction, we derive a per-pixel uncertainty map from the backbone logits and apply an entropy-based selector $\mathcal{S}$ to extract a sparse set of high-uncertainty pixels $\mathcal{P} \subseteq \{1, \ldots, H\} \times \{1, \ldots, W\}$. For each pixel $p \in \mathcal{P}$, the entropy is computed from the softmax-normalized logits $\mathbf{q}_p \in \mathbb{R}^C$ as:

$$\mathcal{H}(p) = -\sum_{c=1}^{C} \mathbf{q}_p^{(c)} \log \mathbf{q}_p^{(c)}. \tag{2}$$

Pixels with entropy above a predefined threshold α, i.e., satisfying $\mathcal{H}(p) > \alpha$, are then selected for refinement. This design allows our framework to focus computation on semantically critical yet spatially sparse regions, rather than uniformly refining all pixels. Empirically, our entropy-based selector recovers around 80% of the high-error pixels while refining only the top 10% most uncertain ones (Fig. 4), substantially outperforming magnitude-based selection and approaching the performance of a learnable selector at a fraction of the cost (2.3 ms vs. 4.0 ms on RTX 4090).

(ii) Sparse Refinement Module. After identifying the high-uncertainty pixels $\mathcal{P}$, we apply a lightweight sparse refinement module $\mathcal{R}$ to improve local predictions. Since the selected pixels are distributed irregularly across the image—

similar to sparse 3D point sets with uneven density, occlusion, and complex structure—a dense CNN becomes inefficient and unnecessary. Motivated by this analogy, we employ a modified MinkowskiUNet [9] as our sparse feature extractor, which uses sparse convolutions to operate only on active locations, maintaining sparsity throughout the network and enabling efficient high-resolution processing. The refined residual correction for each selected pixel is computed as:

$$\Delta\mathbf{Y}_p = \mathcal{R}(\mathbf{I}_{\mathrm{HR}}|_p), \quad \forall p \in \mathcal{P}. \tag{3}$$

(iii) Gated Fusion. After obtaining the sparse refinements, the most straightforward approach is to overwrite the coarse prediction at the refined pixel locations. However, sparse high-resolution predictions lack the global context captured by the dense low-resolution estimates, this naïve substitution often leads to inconsistent or noisy results. To effectively combine global consistency with local precision, we introduce a gated fusion mechanism that adaptively balances the two sources of information.

To combine the sparse refinements $\Delta\mathbf{Y}$ with the initial dense estimate $\hat{\mathbf{Y}}_{\mathrm{HR}}$, we employ a pixel-wise gating strategy. For each selected pixel $p \in \mathcal{P}$, the final prediction is computed as:

$$\mathbf{Y}_p = f\left(w_p \cdot \hat{\mathbf{Y}}_p + (1 - w_p) \cdot \Delta\mathbf{Y}_p\right), \tag{4}$$

where the fusion weight w_p is defined as:

$$w_p = \sigma\left(\mathrm{MLP}\left([\hat{\mathbf{Y}}_p; \Delta\mathbf{Y}_p; \mathcal{H}(\hat{\mathbf{Y}}_p); \mathcal{H}(\Delta\mathbf{Y}_p)]\right)\right), \tag{5}$$

with $\mathrm{MLP}(\cdot)$ a two-layer perceptron and $\sigma(\cdot)$ the sigmoid activation.

Intuitively, this design allows the network to rely more on the coarse estimate when uncertainty is low, while assigning greater weight to the refined prediction in ambiguous or high-entropy regions. By conditioning on both the predictions and their uncertainty, the gated fusion achieves a smooth balance between global structure and local detail, producing coherent and high-fidelity geometric outputs.

4 Experiments

4.1 Implementation Details

Synthetic Training Dataset. As the first to address 2K-level geometric prediction, we construct a dedicated synthetic dataset to support our experiments. Inspired by FoundationStereo [75], we use NVIDIA Omniverse to generate 50K high-quality rendered frames for training both depth estimation and 3D reconstruction models. The dataset covers structured scenes with diverse geometries and textures under complex, realistic lighting. Further dataset details and example scenes are provided in the appendix.

Zero Shot	Method	ARKitScenes			Scannet++		
		AbsRel ↓	RMSE ↓	$\delta_{0.5}$ ↑	AbsRel ↓	RMSE ↓	$\delta_{0.5}$ ↑
No	MSPF [76]	0.0149	0.0363	0.9721	0.0226	0.0975	0.9674
	DepthAnything v2* [78]	0.0326	0.0764	0.9615	0.0371	0.1010	0.9437
	PromptDA [40]	0.0131	0.0316	0.9825	0.0175	0.0829	0.9781
	2K Retrofit (Ours)	**0.0118**	**0.0275**	**0.9866**	**0.0146**	**0.0774**	**0.9825**
Yes	Depth Prompting [56]	0.0212	0.0422	0.9710	0.0242	0.0983	0.9657
	BPNet [67]	0.1261	0.2100	0.9513	0.1419	0.1963	0.9357
	DepthAnything v2 [78]	0.0411	0.0792	0.9529	0.0402	0.1145	0.9404
	Marigold [28]	0.0609	0.1065	0.9087	0.0603	0.1412	0.8718
	PatchRefiner [35]	0.0206	0.0418	0.9722	0.0237	0.0976	0.9683
	PromptDA [40]	0.0142	0.0376	0.9813	0.0224	0.0966	0.9700
	2K Retrofit (Ours)	**0.0129**	**0.0324**	**0.9831**	**0.0197**	**0.0853**	**0.9801**

Table 1: Quantitative comparisons on ARKitScenes, Scannet++ dataset. Methods marked with * are finetuned with their released models and code on ARKitScenes [25] and ScanNet++ [81] datasets.

Training and Latency Details. 2K Retrofit is trained independently from the baseline models, while the loss function is inherited from the corresponding base model. Training proceeds for 13 epochs with a batch size of 8, requiring approximately a day on two NVIDIA RTX A6000 GPUs. Inference latency is evaluated on a single RTX 4090 using FP16 precision.

4.2 Monocular 2K Depth Estimation

We first evaluate 2K Retrofit on monocular 2K depth estimation to verify its ability of enhance high-resolution geometric prediction without retraining the base model. Monocular depth estimation serves as a fundamental benchmark for geometric perception, providing a direct and quantitative way to assess a model's ability to recover fine-grained 3D details from high-resolution inputs. Depth Anything v2 [78] is used as the frozen backbone. Given a single high-resolution RGB image, our method produces a dense 2K depth map by combining a coarse low-resolution prediction with entropy-guided sparse refinement.

Experiments are conducted on the 2K Retrofit synthetic dataset and two real-world benchmarks—ScanNet++ [81] and ARKitScenes [25]—both providing 1440×1920 depth annotations. Following standard protocols, we use their official training/validation splits, and additionally test on ETH3D [65] in a zero-shot setting to assess generalization. Performance is measured using standard monocular depth metrics: AbsRel, $\delta_{0.5}$, and RMSE.

Results. We compare 2K Retrofit with state-of-the-art (SOTA) depth estimation methods from two categories: monocular depth estimation (MDE) and depth completion/upsampling. For MDE, we include Depth Anything v2 [78] and Marigold [28]; for completion and upsampling, we evaluate BPNet [67], Depth Prompting [56], MSPF [76], BoostingDepth [50], PatchFusion [34], PatchRe-

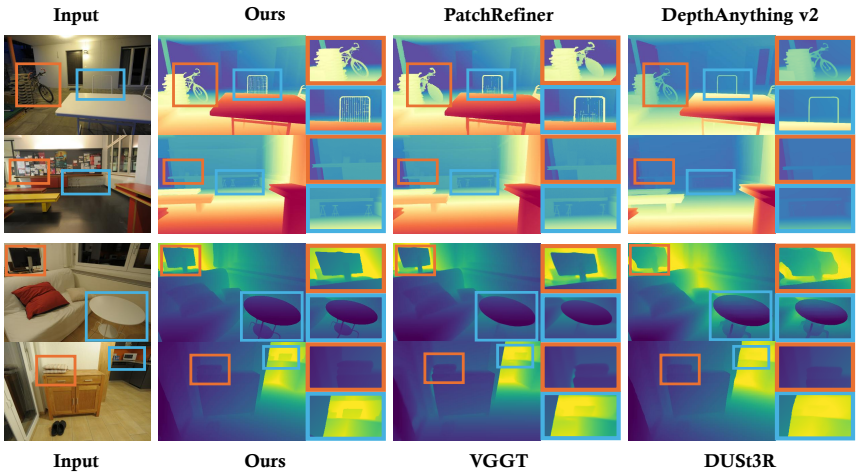


Fig. 5: Qualitative comparisons of 2K Retrofit with state-of-the-art methods on the ETH3D dataset. The top two rows show results on the **depth estimation** task, while the bottom two rows present results on the **point map estimation** task. 2K Retrofit consistently produces sharper high-resolution geometry and recovers fine structures more accurately than competing methods.

finer [35], PRO [31], and PromptDA [40]. All methods are evaluated under two settings: *zero-shot* (trained only on our synthetic 2K Retrofit dataset) and *non zero-shot*.

As shown in Table 1, Table 2, and Fig. 5, 2K Retrofit achieves the best overall performance on ARKitScenes, ScanNet++, and ETH3D. Compared with the strongest baseline PRO [31], our method reduces AbsRel by over 30% while running more than three times faster (8.1 vs. 6.2 FPS), demonstrating advantages in both accuracy and efficiency.

Qualitative comparisons in Fig. 5 show that 2K Retrofit reconstructs sharper object boundaries and cleaner thin structures, especially near occlusions and high-frequency regions. These improvements stem from entropy-guided sparse refinement, which focuses computation on uncertain regions while avoiding redundant processing.

4.3 Multi-View 2K Pointmap Estimation

We further evaluate 2K Retrofit on high-resolution multi-view 3D pointmap estimation to examine its generalization beyond monocular depth prediction. Multi-view pointmap estimation serves as a more comprehensive benchmark for geometric reasoning, as it requires consistent 3D reconstruction across multiple views rather than single frame depth prediction. VGGT [70] is adopted as the frozen backbone. Given a set of multi-view RGB images, our framework refines

Method	Monocular Depth			
	AbsRel ↓	RMSE ↓	$\delta_{0.5}$ ↑	FPS
BoostingDepth [50] (CVPR 2021)	0.0375	0.1201	0.9503	5.0
DepthAnything v2* [78] (NeurIPS 2024)	0.0342	0.1129	0.9550	4.7
Marigold [28] (CVPR 2024)	0.0711	0.1305	0.8849	1.3
BPNet [67] (CVPR 2024)	0.1494	0.2561	0.9137	2.6
PatchFusion$_{P=177}$ [34] (CVPR 2024)	0.0255	0.1104	0.9618	3.0
PatchRefiner [35] (ECCV 2024)	0.0240	0.0987	0.9659	3.1
PRO [31] (ICCV 2025)	0.0217	0.0902	0.9688	6.2
2K Retrofit (Ours)	**0.0192**	**0.0877**	**0.9700**	**8.1**

(a) Monocular depth estimation.

Method	Point Map Estimation			
	Acc. ↓	Comp. ↓	Overall ↓	FPS
DUSt3R [72] (CVPR 2024)	1.462	0.927	1.195	1.1
MASt3R [32] (ECCV 2024)	1.267	0.784	1.026	1.8
VGGT [70] (CVPR 2025)	1.167	0.639	0.903	2.6
WinT3R [36] (ArXiv 2025)	1.131	0.617	0.891	3.2
StreamVGGT [84] (ICLR 2026)	1.140	0.613	0.870	3.5
2K Retrofit (Ours)	**0.935**	**0.602**	**0.839**	**5.5**

(b) Point map estimation.

Table 2: Quantitative comparisons on ETH3D [65] dataset.

the coarse pointmaps produced by the backbone through entropy-guided sparse correction, enabling accurate and consistent 2K-level 3D reconstruction.

Experiments are conducted on both synthetic (MVS-Synth [24] and our 2K Retrofit synthetic dataset) and real-world datasets (DL3DV [41] and BlendMVS [80]), with an additional zero-shot evaluation on ETH3D [65]. We follow standard pointmap metrics [69]: average accuracy, average completeness, and the overall average error.

Results. We benchmark 2K Retrofit against state-of-the-art feedforward 3D reconstruction models, including DUSt3R [72], MASt3R [32], WinT3R [36], StreamVGGT [84], and VGGT [70]. As shown in Table 2, our method consistently improves pointmap estimation on ETH3D. Compared with the strongest baseline VGGT [70], 2K Retrofit reduces average accuracy error and completeness by roughly 20% and 6%, respectively, yielding an overall error reduction of about 15%. These results demonstrate that selective refinement significantly improves geometric precision without retraining the backbone.

Qualitative comparisons in Fig. 5 further support these findings. 2K Retrofit reconstructs finer structures and cleaner surfaces, particularly around thin objects and occlusion boundaries where foundation models typically degrade at 2K resolution.

In addition to accuracy gains, 2K Retrofit runs at 5.5 FPS—over 3× faster than VGGT (2.6 FPS) and nearly 8× faster than DUSt3R (1.1 FPS)—showing a favorable balance between accuracy and computational efficiency for high-resolution 3D reconstruction.

We further compare our method with multi-view stereo (MVS) approaches. We note that many MVS methods, such as RAFT-Stereo [42] and MVSFormer++ [8], also follow a coarse-to-fine paradigm. Therefore, we conduct additional evaluations on the ETH3D benchmark. For fairness, MVSFormer++ [8] is evaluated with both its original weights and weights fine-tuned on our synthetic dataset, denoted as MVSFormer++†. Results in Table 3 show that our method outperforms these approaches and remains competitive with MVSFormer++†, while being substantially faster.

Methods	Precision↑	F1-Score↑	Runtime(s)↓
CascadeMVS [19] (CVPR2020)	79.17	80.84	0.4
RAFT-Stereo [42] (3DV 2021)	80.36	81.42	0.8
MVSFormer++ [8] (ICLR2024)	82.23	83.75	1.5
MVSFormer++† [8] (ICLR2024)	**84.49**	**85.13**	1.5
2K Retrofit (Ours)	84.01	84.88	**0.3**

Table 3: Dense MVS Estimation on the ETH [65] Dataset.

4.4 Efficiency Analysis on ETH3D.

We compare the efficiency of 2K Retrofit with retraining VGGT at 2K resolution (Table 4). Full retraining requires 76.5 GB memory and 495 GFLOPs per forward pass, whereas our method reduces these costs by over 50% (32.8 GB, 172 GFLOPs). While retrained VGGT achieves slightly higher accuracy, 2K Retrofit runs at 8.5 FPS versus 0.5 FPS, providing a 17× speedup with competitive performance, demonstrating a practical balance between accuracy and efficiency for high-resolution 3D geometry prediction.

Method	**VRAM (GB) ↓**	**GFLOPs ↓**	**Accuracy ↓**	**Completeness ↓**	**FPS ↑**
VGGT at 518p, upsampled to 2K	30.2	133	1.267	0.730	**6.1**
VGGT retrained at 2K	76.5	495	**0.911**	**0.593**	0.8
2K Retrofit (Ours)	**32.8**	**172**	0.935	0.602	5.5

Table 4: Efficiency comparison on ETH3D point map estimation task.

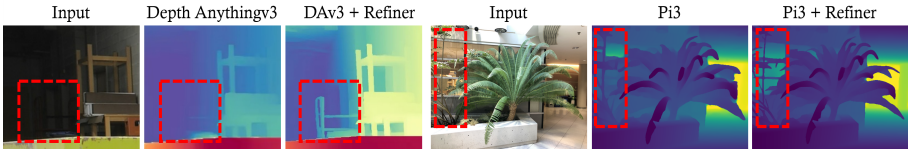
4.5 Generalization and Robustness Analysis

To further evaluate the generalization and robustness of our approach, we extend our experiments to additional monocular depth models, including Pi3 [73] and Depth Anything v3 [39](Fig. 6a). These models are tested without retraining, demonstrating the plug-and-play compatibility of our framework across different architectures and consistently improving reconstruction quality.

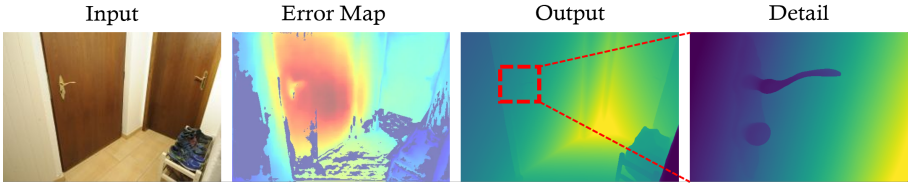
We also present representative failure cases in Fig. 6b. Our method may still struggle in extremely textureless or highly reflective regions—common challenges in depth estimation and 3D reconstruction—where errors can become dense. Nevertheless, our approach remains capable of refining high-frequency structural details in most regions, resulting in improved geometric fidelity.

4.6 Ablation Study

We analyze alternative designs for key components of our method, focusing on point map estimation. Detailed breakdowns of accuracy and efficiency improvements are provided below.



(a) Generality across different backbones.



(b) Robustness visualization under challenging condition.

Fig. 6: Generalization and Robustness Analysis.

Impact of Entropy Selectors. We compare our entropy-based pixel selector with several alternatives, including a random selector, a learnable selector, edge-based heuristics, and bilateral upsampling (Table 5a). The random selector fails to surpass the low-resolution baseline, showing that unguided refinement is ineffective at 2K resolution. The learnable selector slightly improves recall but nearly doubles latency and memory. Edge-based heuristics and bilateral upsampling perform better but still lag behind entropy-guided selection. Overall, entropy-based selection achieves the best accuracy–efficiency trade-off.

Pixel Selector	Acc. ↓	Comp. ↓	FPS
Random	1.126	0.835	4.1
Edge-based heuristics	0.942	0.625	4.1
Bilateral upsampling	0.940	0.617	4.0
Learnable	**0.921**	0.635	3.8
Entropy	0.935	**0.602**	**5.5**

(a) The pixel selector.

Fusion strategy	Acc. ↓	Comp. ↓	FPS
Direct	1.113	0.735	**5.9**
Entropy	0.940	0.607	5.0
Gated	**0.935**	**0.602**	5.5

(b) The Fusion strategy.

α	Acc. ↓	Comp. ↓	FPS
0.8	1.135	0.633	8.0
0.6	0.962	0.621	7.3
0.1	**0.917**	**0.595**	3.4
0.3	0.935	0.602	**5.5**

(c) The entropy threshold.

Sparse Feature Extractor	Acc. ↓	Comp. ↓	FPS
Point Transformer	**0.919**	**0.587**	1.7
MinkowskiUNet	0.935	0.602	**5.5**

(d) The sparse feature extractor.

Table 5: Ablation experiments to validate our design choices.

Impact of Fusion Strategies. As shown in Table 5b, direct replacement performs poorly due to the loss of global context, while entropy-based fusion im-

proves stability but still lacks feature-level adaptivity. Our gated fusion achieves the lowest errors (Acc. 0.935, Comp. 0.602) with comparable FPS, confirming its effectiveness in selectively integrating coarse and refined estimates.

Impact of Entropy Thresholds. We further study the influence of the entropy threshold α on accuracy and latency (see Table 5c). Lowering the threshold increases the number of selected pixels, leading to better reconstruction accuracy but also higher latency. In contrast, higher thresholds reduce computational load at the expense of finer details. We choose $\alpha = 0.3$ as a moderate setting that achieves strong accuracy while maintaining the largest speedup.

Impact of Sparse Extractor Capacity. We also evaluate different backbone choices for sparse feature extraction (Table 5d). Using a point transformer as the sparse feature extractor yields marginally higher accuracy but considerably reduces inference speed. In contrast, MinkowskiUNet offers an optimal balance of efficiency and accuracy, and is therefore adopted as our default extractor.

5 Conclusion

This paper proposes 2K Retrofit, a general framework designed to overcome the scalability limitations of current geometric foundation models for high-resolution prediction. By combining fast coarse inference with entropy-based sparse refinement, our method efficiently bridges the gap between low-resolution backbones and 2K-level outputs. Unlike previous approaches that require network modification, patch-wise processing, or retraining, 2K Retrofit directly adapts existing foundation models to high-resolution scenarios without altering their original architectures. This modular and easily integrable design makes the method broadly applicable to a range of geometric prediction tasks, including depth estimation and 3D reconstruction. Extensive experiments on both 2K depth estimation and 2K pointmap reconstruction demonstrate that 2K Retrofit achieves superior accuracy and significantly higher efficiency compared with existing methods.

Limitations. 2K Retrofit may still struggle in challenging scenarios such as reflective surfaces, transparent objects, and low-texture regions, where geometric cues are inherently ambiguous, leaving a future research direction of incorporating stronger geometric priors into our model.

Acknowledgements

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0122603). It was also supported by the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant No. 2024M764093 and Grant No. BX20250485, the Beijing Natural Science Foundation under Grant No. 4254100, and by Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

References

1. 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)
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. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073 (2024)
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
5. Bolya, D., Hoffman, J.: Token merging for fast stable diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4599–4603 (2023)
6. Cai, X., Wang, S., Wang, P., Wang, Y., Fan, Z., Li, W., Zhang, T., Tao, J., Jin, Y., Li, D.: Mem4d: Decoupling static and dynamic memory for dynamic scene reconstruction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 2552–2560 (2026)
7. Cao, C., Ren, X., Fu, Y.: Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth. arXiv preprint arXiv:2208.02541 (2022)
8. Cao, C., Ren, X., Fu, Y.: Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. arXiv preprint arXiv:2401.11673 (2024)
9. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
10. Eigen, D., Fergus, R.: 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)
11. 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)
12. 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)
13. Fang, X., Gao, J., Wang, Z., Chen, Z., Ren, X., Lyu, J., Ren, Q., Yang, Z., Yang, X., Yan, Y., et al.: Dens3r: A foundation model for 3d geometry prediction. arXiv preprint arXiv:2507.16290 (2025)
14. Fu, Q., Xu, Q., Ong, Y.S., Tao, W.: Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. Advances in Neural Information Processing Systems **35**, 3403–3416 (2022)
15. Furukawa, Y., Hernández, C., et al.: Multi-view stereo: A tutorial. Foundations and trends® in Computer Graphics and Vision **9**(1-2), 1–148 (2015)
16. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the IEEE international conference on computer vision. pp. 873–881 (2015)
17. Gao, P., Ma, T., Li, H., Lin, Z., Dai, J., Qiao, Y.: Convmae: Masked convolution meets masked autoencoders. arXiv preprint arXiv:2205.03892 (2022)
18. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2020)

19. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2020)
20. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. arxiv. 2022. arXiv preprint arXiv:2209.08575 (2022)
21. Han, S.H., Park, M.G., Yoon, J.H., Kang, J.M., Park, Y.J., Jeon, H.G.: High-fidelity 3d human digitization from single 2k resolution images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12869–12879 (2023)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
23. Huang, L., You, S., Zheng, M., Wang, F., Qian, C., Yamasaki, T.: Green hierarchical vision transformer for masked image modeling. Advances in Neural Information Processing Systems **35**, 19997–20010 (2022)
24. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2821–2830 (2018)
25. Huang, Y.H., Proesmans, M., Georgoulis, S., Van Gool, L.: Uncertainty based model selection for fast semantic segmentation. In: 2019 16th International Conference on Machine Vision Applications (MVA). pp. 1–6. IEEE (2019)
26. Hui, T.W., Loy, C.C., Tang, X.: Depth map super-resolution by deep multi-scale guidance. In: European conference on computer vision. pp. 353–369. Springer (2016)
27. Izquierdo, S., Sayed, M., Firman, M., Garcia-Hernando, G., Turmukhambetov, D., Civera, J., Mac Aodha, O., Brostow, G., Watson, J.: Mvsanywhere: Zero-shot multi-view stereo. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11493–11504 (2025)
28. 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)
29. Kirillov, A., Wu, Y., He, K., Girshick, R.: Pointrend: Image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020)
30. Kong, Z., Dong, P., Ma, X., Meng, X., Sun, M., Niu, W., Shen, X., Yuan, G., Ren, B., Qin, M., et al.: Spvit: enabling faster vision transformers via soft token pruning (2022). URL https://arxiv. org/abs/2112.13890 (2021)
31. Kwon, B., Kim, M.: One look is enough: Seamless patchwise refinement for zero-shot monocular depth estimation on high-resolution images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8077–8087 (2025)
32. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision. pp. 71–91. Springer (2024)
33. 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)
34. Li, Z., Bhat, S.F., Wonka, P.: 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 (2024)
35. Li, Z., Bhat, S.F., Wonka, P.: Patchrefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. In: European Conference on Computer Vision. pp. 250–267. Springer (2024)
36. Li, Z., Zhou, J., Wang, Y., Guo, H., Chang, W., Zhou, Y., Zhu, H., Chen, J., Shen, C., He, T.: Wint3r: Window-based streaming reconstruction with camera token pool. arXiv preprint arXiv:2509.05296 (2025)
37. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. arXiv preprint arXiv:2202.07800 (2022)
38. Liang, Z., Wang, L., Diao, Y., Wang, Y., Mu, H., Zuo, L., Gao, H., Fan, Z., Yang, X.: Understanding the adversarial robustness of deep learning-based single-pixel imaging. Pattern Recognition p. 112555 (2025)
39. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
40. Lin, H., Peng, S., Chen, J., Peng, S., Sun, J., Liu, M., Bao, H., Feng, J., Zhou, X., Kang, B.: Prompting depth anything for 4k resolution accurate metric depth estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17070–17080 (2025)
41. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
42. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International conference on 3D vision (3DV). pp. 218–227. IEEE (2021)
43. Liu, J., Chen, Y., Ye, X., Tian, Z., Tan, X., Qi, X.: Spatial pruned sparse convolution for efficient 3d object detection. Advances in neural information processing systems **35**, 6735–6748 (2022)
44. Liu, Z., Zhang, Z., Khaki, S., Yang, S., Tang, H., Xu, C., Keutzer, K., Han, S.: Sparse refinement for efficient high-resolution semantic segmentation. In: European Conference on Computer Vision. pp. 108–127. Springer (2024)
45. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp. 347–353 (1998)
46. Ma, X., Zhou, Y., Wang, H., Qin, C., Sun, B., Liu, C., Fu, Y.: Image as set of points. arXiv preprint arXiv:2303.01494 (2023)
47. Ma, Z., Teed, Z., Deng, J.: Multiview stereo with cascaded epipolar raft. In: European Conference on Computer Vision. pp. 734–750. Springer (2022)
48. Metzger, N., Daudt, R.C., Schindler, K.: Guided depth super-resolution by deep anisotropic diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18237–18246 (2023)
49. Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S., Aksoy, Y.: 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)
50. Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S., Aksoy, Y.: 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)
51. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3504–3515 (2020)
52. Ouyang, W., Song, X., Feng, B., Xu, Z.: Octocc: High-resolution 3d occupancy prediction with octree. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4369–4377 (2024)
53. Pan, B., Lin, W., Fang, X., Huang, C., Zhou, B., Lu, C.: Recurrent residual module for fast inference in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1536–1545 (2018)
54. Pan, B., Panda, R., Fosco, C., Lin, C.C., Andonian, A., Meng, Y., Saenko, K., Oliva, A., Feris, R.: Va-red^2: Video adaptive redundancy reduction. arXiv preprint arXiv:2102.07887 (2021)
55. Pan, B., Panda, R., Jiang, Y., Wang, Z., Feris, R., Oliva, A.: Ia-red^2: Interpretability-aware redundancy reduction for vision transformers. Advances in neural information processing systems **34**, 24898–24911 (2021)
56. Park, J.H., Jeong, C., Lee, J., Jeon, H.G.: Depth prompting for sensor-agnostic depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9859–9869 (2024)
57. Peng, R., Wang, R., Wang, Z., Lai, Y., Wang, R.: Rethinking depth estimation for multi-view stereo: A unified representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8645–8654 (2022)
58. 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)
59. 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)
60. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems **34**, 13937–13949 (2021)
61. Rey-Area, M., Yuan, M., Richardt, C.: 360monodepth: High-resolution 360deg monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3762–3772 (2022)
62. Riegler, G., Rüther, M., Bischof, H.: Atgv-net: Accurate depth super-resolution. In: European conference on computer vision. pp. 268–284. Springer (2016)
63. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
64. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European conference on computer vision. pp. 501–518. Springer (2016)
65. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: 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)

66. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)
67. Tang, J., Tian, F.P., An, B., Li, J., Tan, P.: Bilateral propagation network for depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9763–9772 (2024)
68. Tang, S., Chen, J., Wang, D., Tang, C., Zhang, F., Fan, Y., Chandra, V., Furukawa, Y., Ranjan, R.: Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In: European Conference on Computer Vision. pp. 175–191. Springer (2024)
69. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on pattern analysis and machine intelligence **13**(4), 376–380 (2002)
70. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
71. Wang, Q., Zheng, S., Yan, Q., Deng, F., Zhao, K., Chu, X.: Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. arXiv preprint arXiv:1912.09678 (2019)
72. 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)
73. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
74. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5610–5619 (2021)
75. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5249–5260 (2025)
76. Xian, C., Qian, K., Zhang, Z., Wang, C.C.: Multi-scale progressive fusion learning for depth map super-resolution. arXiv preprint arXiv:2011.11865 (2020)
77. 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)
78. 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)
79. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018)
80. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1790–1799 (2020)
81. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
82. Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., Hu, X.: Refinemask: Towards high-quality instance segmentation with fine-grained features. In: Proceedings of

the IEEE/CVF conference on computer vision and pattern recognition. pp. 6861–6869 (2021)
83. Zhong, Z., Liu, X., Jiang, J., Zhao, D., Ji, X.: Guided depth map super-resolution: A survey. ACM Computing Surveys **55**(14s), 1–36 (2023)
84. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025)