CEIG - Spanish Computer Graphics Conference (2026)
E. Garces and I. Muñoz-Pandiella (Editors)

FRED: Full-Resolution Equirectangular Depth Estimation

Uzair Shah[1], Giovanni Pintore[2], Muhammad Tukur[1], Zahoor Ahmad[1], Jens Schneider[1]
Matteo Sgrenzaroli[3], Giorgio Vassena[4], Pere Pau Vázquez[5], Enrico Gobbetti[2], Marco Agus[1]

[1]HBKU, Doha, Qatar [2]CRS4, Cagliari, Italy [3]GEXCEL, Brescia, Italy [4]U. Brescia, Italy [5]UPC, Barcelona, Spain

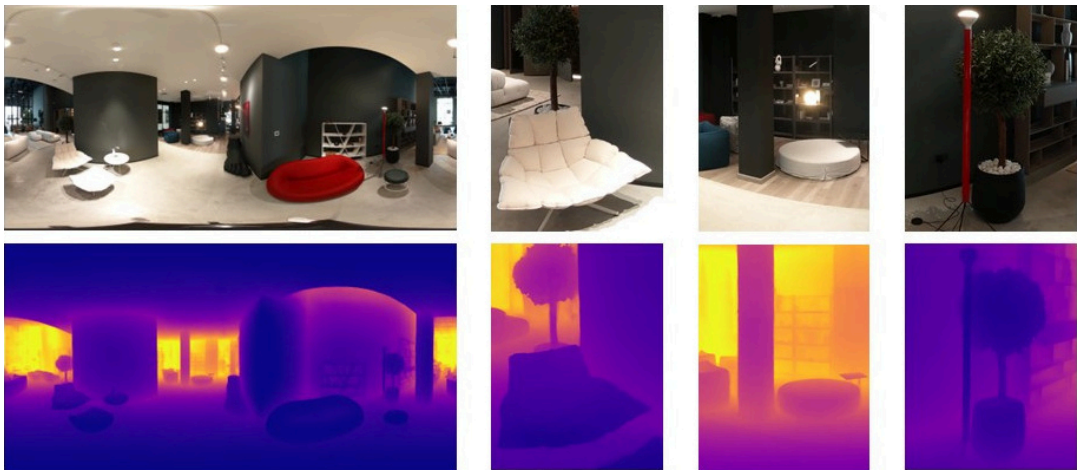


Figure 1: ***FRED inference example:*** *Our training-free AI-based processing pipeline enables zero-shot native resolution equirectangular depth estimation, illustrated here on a casually captured 11008x5504 indoor scene. Image acquired at BLAUUCASA showroom in Msheireb Downtown Doha with Ricoh Theta X camera. Courtesy of Doha Design District.*

Abstract

Monocular 360° *depth estimation is a key component of holistic 3D scene understanding. While deep learning and holistic panoramic methods leverage global context to produce structurally consistent predictions, their high computational cost and reliance on hard-to-obtain high-resolution training data of current approaches restrict them to resolutions far below those of modern omnidirectional cameras and head-mounted displays. Post-hoc upsampling and multi-view stitching only partially address these limitations, often trading local detail for global consistency. To overcome these limitations, we introduce* FRED *(Full-Resolution Equirectangular Depth), a training-free framework that combines coarse holistic* 360° *depth estimation with high-resolution perspective refinement.* FRED *first extracts a low-resolution panoramic prior from a pretrained model, then performs prompt-guided local depth inference on overlapping gnomonic projections distributed on a midpoint-icosahedral sphere. Conditioning local predictions on a global panoramic prior enables their effective fusion through a low-complexity, seam-robust spherical blending strategy. The approach scales linearly with resolution, enabling accurate depth estimation at 8K–60 MP with modest computational requirements. We also present the first benchmark of indoor panoramas with metric 8K depth annotations. Experiments demonstrate that* FRED *outperforms prior methods in both accuracy and perceptual quality on high-resolution panoramic imagery.*

1. Introduction

Automatic 3D modeling of indoor environments from captured data has become a central topic in computer vision, driven by applications in immersive visualization, mapping, and robotics [PMG*20]. While early methods exploited strong geometric regularities in man-made spaces, recent approaches increasingly focus on data-driven inference to handle cluttered and diverse interiors under challenging illumination [GPA24]. Accurate depth estimation remains fundamental, as it enables the recovery of metric scene structure and supports tasks such as AR/VR rendering [SJT*25], navigation [PJAG25], and digital twin creation [TUP*23, ACW25, PJVH*24, PASG26].

Spherical imaging (also known as omnidirectional, panoramic, or 360° imagery) has become a de facto standard in industrial settings and a central research topic. Commodity panoramic cameras now provide cost-effective single-shot acquisition of full-scene context [JOV19], offering rich cues for monocular depth inference and scene understanding [YJL*18]. By effectively combining global context with local cues, panoramic methods infer structurally consistent depth even in environments dominated by textureless regions, occlusions, and clutter [ACW25]. However, high memory and computational costs, together with the scarcity of large-scale high-resolution training data, limit existing holistic 360° approaches to low input resolutions. Recent monocular panoramic depth models—from distortion-aware CNNs to spherical transformers—still operate at around 1024×512 pixels (0.5MP) [ACW25, PASG26], far below both modern VR head-mounted displays (e.g., HTC Vive at 2448 × 2448 per eye) and current omnidirectional sensors, which reach 8K (8192×4096) or even 60MP (e.g., Ricoh Theta X at 11008×5504). This represents a gap of roughly 60× to 120× in pixel count compared to state-of-the-art methods [TBJ*24]. Post-hoc upsampling only partially mitigates this gap. In particular, thin structures and small or distant objects are easily lost when depth is predicted at low resolution and only later upsampled, limiting downstream tasks such as SLAM, meshing, or neural rendering. For this reason, existing high-resolution methods compose inferred depths from multiple limited-FoV perspective views or spherical tiles (Sec. 2), improving memory efficiency and local accuracy but often compromising robustness and global consistency or requiring heavy post-processing.

This work introduces FRED (Full-Resolution Equirectangular Depth), a unified framework that combines holistic 360° inference with high-fidelity local refinement for full-resolution panoramic depth estimation (Sec. 3). Compared to previous works, this unified formulation explicitly connects coarse panoramic reasoning with fine-grained local refinement, enabling a geometry-consistent full-resolution pipeline. Rather than training a new high-resolution model, FRED integrates pretrained components at relevant scales within a geometry-aware tiling, prompting, and fusion scheme that scales to tens of megapixels and is explicitly validated at native sensor resolution. A coarse global depth prior is first obtained from a low-resolution panoramic estimator, providing a structurally consistent foundation that stabilizes local predictions and mitigates scale drift, while enabling direct reuse of existing models trained on large panoramic datasets [STA*24] (Sec. 3.1). The high-resolution panorama is then decomposed into overlapping perspective tiles via gnomonic projections on a midpoint-icosahedral sphere, ensuring near-uniform angular coverage. Gnomonic projection is essential as it transforms spherical patches into rectilinear perspective images—the native input domain of foundation models like PromptDA [LPC*25], trained exclusively on perspective imagery. Each tile, paired with its corresponding coarse depth region, is refined by this pretrained prompt-guided model, where *prompting* means conditioning depth estimation with sparse depth values at specific pixels—analogous to seed points in interactive segmentation—anchoring predictions to the global structure. PromptDA [LPC*25] leverages these sparse depth "prompts" to sharpen local geometry while preserving global coherence, without retraining on panoramic data (Sec. 3.2). Refined tiles are reprojected and fused into a full-resolution equirectangular depth map. Because each patch is conditioned on the global structure, predictions are already near-final, enabling seamless fusion without heavy optimization. A lightweight spherical blending strategy, employing prompt-consistency weighting, raised-cosine windows, and explicit $\pm\pi$ meridian wrap-around handling, manages alignment, confidence, and overlaps while preserving sharp discontinuities (Sec. 3.3). The resulting depth maps are globally consistent, seamless, and at native sensor resolution (Fig. 1).

The main contributions of this work are twofold:

- **Native-resolution unified training-free panoramic pipeline.** FRED presents the first unified pipeline for generating native-resolution (8K–60MP) equirectangular depth maps without additional supervised training or fine-tuning. While current transformer-based panoramic estimators [SZL*23, YWZ*25] are limited to 1K resolution, FRED leverages pretrained models and geometry-aware fusion, beyond prompt injection, to scale linearly with resolution. Unlike projection-and-fuse methods such as *360MonoDepth* [RAYR22] and *PanDA* [CZZ*25], it generalizes across domains and cameras without dataset-specific tuning.
- **Low-complexity geometry-consistent global–local depth fusion.** FRED combines global structural reasoning with high-detail refinement via a coarse panoramic prior capturing large-scale structure and long-range context from ERP models. Injected during per-tile refinement, it acts as a geometric consistency field aligning all high-resolution patches. This enables lightweight spherical fusion with confidence- and structure-aware blending, avoiding heavy optimization used in *360MonoDepth* [RAYR22]. Unlike simple stitching approaches (e.g., *PanDA* [CZZ*25]), it preserves depth discontinuities, handles wrap-around, and produces seamless reconstructions with stable scaling across overlaps (Fig. 4).

Together, these components enable FRED to surpass prior panoramic and projection-based methods in accuracy, perceptual quality, and scalability, combining global awareness with perspective-based geometric precision. The quality and performance of our approach are demonstrated both on commonly available datasets and on a metrically calibrated real-world dataset of RGB-D indoor panoramic scenes at 8192×4096, captured with professional LiDAR and a calibrated 360° camera.

2. Related Work

Depth estimation from monocular 360° imagery and 3D reconstruction of indoor scenes are long-standing problems in computer vision that have gained renewed momentum with the advent of deep learning. A comprehensive review is beyond the scope of this paper. In the following, we discuss the approaches most closely related to our work, referring readers to existing surveys for the broader context [PMG*20, dSPMLJ22, GPA24, ACW25, PASG26].

Panoramic depth estimation. Spherical cameras have become standard for panoramic capture due to their simplicity and low cost [JOV19]. They produce seamless full-view images with an approximate single projection center [IHR*16, HCCJ17], modeled as a unit sphere defined by extrinsic parameters [dSPMLJ22]. Most

devices output a $360^\circ \times 180^\circ$ *equirectangular projection (ERP)* mapping longitude and latitude to image axes [ESLF20], and depth estimation must assign a value to every pixel in the ERP. Early deep approaches adapted perspective CNNs to the equirectangular domain either by *distortion-aware* operators in ERP [SG17, TNT18, ZKZD18, SG19] or by projection to cubemaps [CCD*18], sometimes combining both branches [WYS*20, JSZ*21, HSC*25]. A complementary line introduced *structure-aware* designs that encode indoor priors (gravity, Manhattan cues, room layout) directly in the network, improving efficiency and boundary fidelity [SSC21, PAA*21, PJVH*24, PASG25]. With transformers, methods increasingly leverage long-range context and patch/token workflows tailored to panoramas. PanoFormer [SLL*22] divides patches on the spherical tangent domain and reshapes the self-attention module with a learnable token flow. MultiPanoWise [STA*24] expands the architecture to simultaneously predict depth, normals, and other non-geometric information. Other perspective-patch pipelines process ERP via local views and fuse predictions globally [LGY*22, RAYR22, SZL*23, AW24]. Hybrid representation fusion further balances ERP faithfulness and local detail [YWZ*25]. Despite progress, most methods remain constrained to low resolutions (typically 1024×512) due to the high computational cost of panoramic networks and the lack of high-resolution annotated data, compromising scalability to native 8K–60 MP panoramas and the recovery of fine detail [ACW25, PASG26]. *Self- and weakly supervised* strategies have been explored to mitigate reliance on dense ground truth, including photometric or view-synthesis losses [WHC*18, ZKZ*19, WYT*22], cross-domain distillation [WL24], and structural or dual-domain consistency priors [CAVW24, WHZ*25, CZZ*25]. Cao et al. [CZZ*25], in particular, introduced a Scene Structural Knowledge Transfer module to enhance structural cues under weak supervision. To our knowledge, we are the first to achieve native-resolution ERP depth using a modular framework unifying a holistic prior with local refinement.

Exploiting foundation models for depth estimation. Pretrained foundation depth models have advanced zero-shot monocular depth estimation, improving cross-domain robustness and metric consistency. For panoramic imagery, two main strategies have emerged. *Distillation/transfer* methods use perspective models to supervise equirectangular learners [WL24], addressing the scarcity of annotated panoramic data, but resulting networks struggle to scale to high resolutions on standard hardware. *Projection-and-fuse* approaches apply pretrained depth models on overlapping perspective patches and reproject via seam-aware fusion [LGY*22, RAYR22, SZL*23, ACC*23]. Yet, independent predictions can be inconsistent and difficult to merge, and often require heavy postprocessing [RAYR22, SZL*23]. ST^{2}360D [CZA*25] transforms panoramas into perspective video frames and predicts depth with foundational video models [LHS*20, CGZ*25], using temporal continuity to improve consistency among perspective patches. However, frame prediction lacks holistic context and may not produce a seamless closed loop. Unlike ST2360D and SphereFusion, our pipeline enforces explicit spherical geometry-consistency during refinement and fusion. Peng and Zhang [PZ23] improve registration by aligning perspective predictions to a panoramic target created by a panoramic depth prediction method at a low resolution. This helps produce a more consistent global result, whose quality still depends on the quality of the independent perspective predictions. Moreover, the employed composition methods may lead to discontinuities and blurring around patch boundaries [CZZ*25]. We address these limitations with a geometry-consistent framework: a coarse panoramic prior enforces global metric alignment, while guiding high-resolution patch refinement by prompting a foundational depth estimation model (e.g., PromptDA [LPC*25]), and a low-complexity seam-robust spherical weighted-blending fusion handles depth combination without requiring heavy post-processing. This modular, training-free approach produces globally coherent, native-resolution depth maps. We also contribute a dataset of indoor panoramas with metric 8K depth annotations.

3. Method

Our method estimates metrically consistent depth maps directly at the native resolution of modern 360° imagery through a geometry-driven pipeline integrating *global context*, *local refinement*, and *spherical consistency* (Fig. 2). The framework operates in three stages: a *coarse panoramic prior*, a *geometry-consistent refinement*, and a *spherical fusion module*.

In the first stage (Sec. 3.1), a pretrained low-resolution ERP depth estimator leverages long-range contextual information from the holistic view and semantic–geometric relationships learned from large panoramic datasets to infer a structurally consistent global depth map that captures layout, scale, and depth ordering. Despite its low resolution, this coarse panoramic prior provides a stable geometric scaffold that regularizes subsequent refinement and enforces global metric alignment.

In the second stage (Sec. 3.2), the full-resolution panorama is decomposed into overlapping perspective patches placed on a midpoint-icosahedral tessellation for uniform angular coverage. Each patch combines high-resolution RGB data with its corresponding segment of the global prior, enabling a pretrained prompt-guided foundation model (e.g., PrompDA [LPC*25]) to refine local details while remaining globally coherent. This effectively fuses holistic spherical reasoning with fine-grained local inference under a unified geometric formulation.

Finally (Sec. 3.3), the refined predictions are reprojected and merged into a seamless full-resolution ERP map through an efficient spherical fusion scheme. Because each patch is already guided by the global prior, simple confidence-aware terms and raised-cosine windows suffice, avoiding costly optimization. The strategy maintains metric consistency, handles wrap-around at the $\pm\pi$ meridian, and prevents boundary artifacts. The result is a globally coherent, artifact-free panoramic depth map scalable up to 8K–60MP—without retraining.

3.1. Coarse panoramic prior

FRED starts its processing by estimating *wide-context coarse depth prior* that provides a globally consistent but low-detail representation of the scene geometry seen in the input high-resolution equirectangular panorama $\mathbf{I} \in [0,1]^{H\times W\times 3}$. To this end, we apply a panoramic depth estimator f_ℓ to a downsampled version of $\mathbf{I}$:

$$\mathbf{D}^\ell = f_\ell\big(\mathrm{down}_s(\mathbf{I})\big), \qquad s = \max\big(\tfrac{H}{h}, \tfrac{W}{w}\big), \tag{1}$$

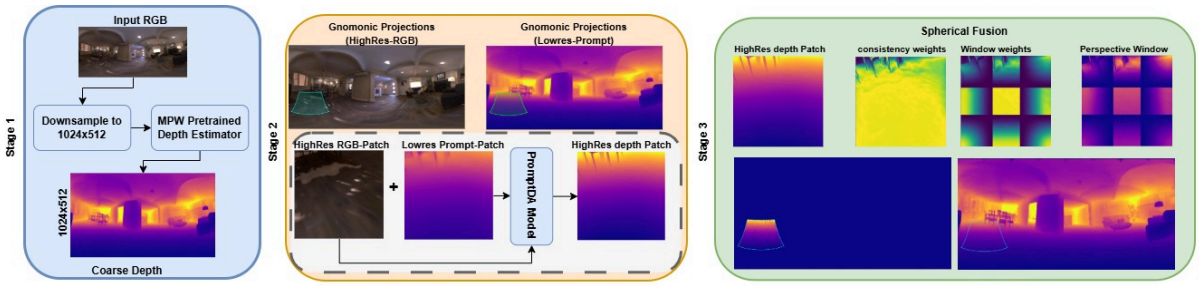


Figure 2: ***Overview of the* FRED *pipeline.*** *Starting from a high-resolution 360° RGB image,* FRED *predicts dense depth at native equirectangular resolution by combining a wide-context global prior, a geometry-consistent local refinement, and a seam-robust spherical fusion.*

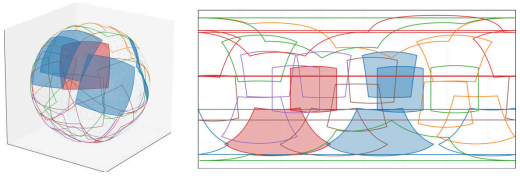


Figure 3: ***Gnomonic sampling.*** *A spherical tessellation provides K evenly distributed viewing directions.*

where (h,w) denotes the standard inference resolution (typically 1024×512) commonly used in annotated 360° datasets, and down_s denotes bicubic downsampling by factor s. We chose bicubic interpolation over alternatives such as nearest-neighbor or area averaging because it provides smooth anti-aliased results that preserve image gradients and color fidelity, which are important for the subsequent depth estimation network. In all experiments, we employ *MultiPanoWise* [STA*24] as the global coarse prior. This supervised panoramic model provides stable, large-scale geometry and coherent depth ordering, which are sufficient to anchor the subsequent high-resolution refinement stage. Importantly, the choice of the coarse estimator has reduced impact on overall performance: its role is primarily to supply long-range structural cues that ensure global consistency across perspective patches, rather than to contribute high-frequency detail. In fact, FRED can operate with any panoramic model to produce a coherent low-resolution depth map. Although the coarse depth $\mathbf{D}^\ell$ lacks high-frequency information, it preserves the dominant layout and scene gradients, offering a geometric scaffold for refinement.

3.2. Localized high-resolution refinement

To enable localized processing, the panorama is decomposed into K overlapping perspective *patches*. We uniformly sample $K\approx20\cdot4^L$ virtual camera orientations on the unit sphere via midpoint icosahedral tessellation (Fig. 3). Each camera k is characterized by a rotation matrix $\mathbf{R}_k\in SO(3)$, an image size $P\times P$, and a field-of-view:

$$\text{fov}_k = 2\,\theta_k + \alpha\cdot\frac{\pi}{180}, \tag{2}$$

where θ_k is the half-diagonal angular extent of the tessellated face and α is an overlap margin ensuring smooth inter-patch transitions. Each virtual camera extracts a perspective RGB patch $\mathbf{I}_k$ by gnomonic projection of the panorama $\mathbf{I}$. For each pixel (x,y) in the patch plane, the corresponding unit ray in camera coordinates is:

$$f_k = \frac{P-1}{2\tan(\text{fov}_k/2)},$$
$$\mathbf{d}^{\text{cam}} = \frac{1}{\sqrt{(x/f_k)^2+(y/f_k)^2+1}}\begin{bmatrix} x/f_k \\ y/f_k \\ 1\end{bmatrix}. \tag{3}$$

The ray is then rotated into the world frame:

$$\mathbf{d}^{\text{world}} = \mathbf{R}_k\,\mathbf{d}^{\text{cam}}. \tag{4}$$

By spherical-to-ERP reprojection, the corresponding pixel coordinates (u,v) in the equirectangular domain are:

$$\lambda = \text{atan2}(d_z,d_x), \qquad \varphi = \arcsin(d_y),$$
$$u = \tfrac{\lambda+\pi}{2\pi}(W-1), \qquad v = \tfrac{\varphi+\frac{\pi}{2}}{\pi}(H-1). \tag{5}$$

To associate the coarse prior $\mathbf{D}^\ell$ with each patch, we sample its values at the corresponding downsampled coordinates $(u^\ell,v^\ell) = \left(u\frac{w}{W}, v\frac{h}{H}\right)$, obtaining a *prompt depth map*:

$$\mathbf{P}_k = \text{sample}\left(\mathbf{D}^\ell, u^\ell, v^\ell\right). \tag{6}$$

Thus, each patch is represented by the pair $(\mathbf{I}_k,\mathbf{P}_k)$, where $\mathbf{P}_k$ encodes coarse geometric hints aligned with the global structure but expressed in the local perspective domain. These pairs are processed by a prompt-guided refinement network g_ϕ [LPC*25]:

$$\widehat{\mathbf{D}}_k = g_\phi(\mathbf{I}_k,\mathbf{P}_k), \tag{7}$$

which outputs a high-resolution local depth map $\widehat{\mathbf{D}}_k$.In this context, *prompting* refers to conditioning the depth estimation by providing sparse depth values at specific pixels—analogous to seed points in interactive estimation—which anchor the predicted depth map to the global structure.

3.3. Spherical fusion

All refined depth patches $\widehat{\mathbf{D}}_k$ are reprojected from their local perspective domains back onto the equirectangular canvas, restoring the original resolution $(H\times W)$, and are blended to compute the final high-resolution depth map. Since they are conditioned to be already consistent with the overall structure of the environment, we

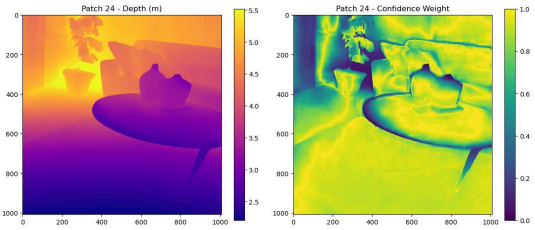


Figure 4: ***Patch reprojection and prompt-consistency–weighted fusion.*** *Each refined patch is reprojected and blended into the ERP using Tukey windows and prompt-consistency terms for seamless, coherent fusion. Left: patch depth. Right: per-pixel consistency weight.*

do not need complex and heavy post-hoc optimization to achieve a seamless high-resolution result. We therefore propose a very efficient low-cost solution based on geometry-aware weighting.

Each patch contributes a weighted vote to every ERP pixel uv according to:

$$\mathbf{W}_k[uv] = \mathbf{W}_k^{\mathrm{tk}}[uv] \cdot \mathbf{w}_k^{\mathrm{pc}}[uv], \tag{8}$$

$\mathbf{W}_k^{\mathrm{tk}}$ is a Tukey window, defined as a tensor product of tapered cosine functions along each patch axis, rising from zero at the patch boundaries to one at a distance of $r/2$ from the edge, with central pixels receiving full weight. In essence, this creates a smooth falloff that assigns maximum confidence to the patch center and gradually reduces weight toward the edges, ensuring seamless blending in overlap regions without hard transitions. In all our experiments, the parameter r is set to 0.35. This weighting not only smooths the combination of overlapping patches but also encodes confidence in the per-patch depth estimates, which is highest at the patch centers and decreases toward the boundaries due to reduced contextual information.

$\mathbf{w}_k^{\mathrm{pc}}$ is the prompt-consistency weight derived from Sec. 3.2, encouraging alignment between the refined prediction and the global coarse prior:

$$\mathbf{w}_k^{\mathrm{pc}}[uv] = \exp\left(-|\widehat{\mathbf{D}}_k[uv] - \mathbf{P}_k[uv]|\right), \tag{9}$$

This weighting encourages geometrically consistent refinement, preserving fine RGB-induced details while penalizing deviations from the global depth prior (Fig. 4).

Accumulating patch contributions yields for each pixel uv:

$$\mathbf{S}[uv] \mathrel{+}= \mathbf{W}_k \widehat{\mathbf{D}}_k, \qquad \mathbf{A}[uv] \mathrel{+}= \mathbf{W}_k, \tag{10}$$

and the predicted panoramic depth is obtained by normalization:

$$\widehat{\mathbf{D}}[uv] = \frac{\mathbf{S}[uv]}{\mathbf{A}[uv] + \varepsilon}. \tag{11}$$

The resulting $\widehat{\mathbf{D}}$ combines local high-frequency cues with global geometric consistency, producing a depth map faithful to the native image resolution and structurally coherent across the entire panorama.

4. Results

We evaluate FRED on a diverse set of *high-resolution* panoramic scenes (4K and beyond). Quantitative results are reported on *Replica360* [RAYR22], containing 13 indoor scenes and 130 RGB-D panoramas with accurate ground truth at 4K resolution. We also include results for a real-world benchmark of 52 panoramas at 8192×4096, captured with a professional LiDAR and a calibrated 360° RGB camera. Qualitative evaluation further includes zero-shot examples from public datasets [CZZ*25] and casual user captures (see Fig. 1 and supplementary material). For quantitative comparison, we selected only baselines for which *complete and reproducible code* is publicly available, ensuring faithful inference and, when required, retraining under identical input conditions. Several additional related methods [CZA*25, PZ23, WL24], discussed in Sec. 2, could not be included quantitatively due to missing pretrained models or incomplete code releases, which prevented a fair and controlled evaluation. Accordingly, we include *360MonoDepth* [RAYR22], a representative projection–fusion pipeline that combines multi-view predictions via cross-projection blending, and *PanDA* [CZZ*25], a transformer-based panoramic model adapted from foundation-depth priors. These methods cover the two dominant paradigms in high-resolution 360° depth estimation—projection-based and transformer-driven pipelines—and provide strong supervised and zero-shot references. Our framework employs *MultiPanoWise* [STA*24] as the coarse prior. The choice of the coarse model, however, has a very limited influence on the final performance, as it serves only to provide a globally consistent depth scaffold that anchors perspective refinements within a coherent spherical context. All experiments are run in PyTorch on an NVIDIA RTX 4090 (24 GB), using 1024×1024 perspective patches with a 100° field of view, 6° overlap, and Tukey-window inverse-depth spherical fusion. For hardware parity, *PanDA* was tested on an NVIDIA H200 (140 GB), where the largest feasible patch size under identical settings was 2016×2016.

4.1. Computational Assessment

To analyze the computational footprint of FRED, we report the parameters and FLOPs of its two components: the coarse 360° prior (MultiPanoWise [STA*24]) and the high-resolution patch estimator (PromptDA [LPC*25]). MultiPanoWise, derived from the PanoFormer [SLL*22] architecture, is evaluated once at 1024×512 and contributes a fixed cost of ∼25–30M parameters and ∼80–95 GFLOPs. The high-resolution stage dominates the overall complexity. An input panorama is decomposed into N_p overlapping gnomonic crops, each independently processed by PromptDA. For PromptDA-S, each patch requires ∼50–55M parameters and ∼15–20 GFLOPs, and the total inference cost scales linearly as $N_p\, G_{\mathrm{patch}}$. No additional training overhead is introduced, and composition is obtained through simple blending rather than iterative optimization. Despite operating at native 8K–60MP resolution, FRED remains tractable thanks to patch-level parallelism and the lightweight nature of its components.

4.2. Quantitative Assessment

We evaluate performance using standard depth metrics—Absolute Relative Error (AbsRel), Root Mean Squared Error (RMSE), and

Table 1: ***Quantitative results on*** **Replica360-4K** ***[SWM*19].*** *Comparison with native high-resolution panoramic methods (360MonoDepth, PanDA). FRED achieves best performance across all metrics—global accuracy (AbsRel, RMSE) and local consistency (σ thresholds)—while remaining efficient and training-free.* ^{504}, ^{1008}*: PanDA patch sizes;* ^{M}/$_{S}$*: trained on Matterport/Stanford;* ^{M2}/$_{M3}$*: MiDaS v2/v3 backbone.*

Method	AbsRel↓	MAE↓	RMSE↓	RMSELog↓	$\sigma < 1.25$↑	$\sigma < 1.25^2$↑	$\sigma < 1.25^3$↑
360MonoDepth^{M2}	0.150	0.335	0.558	0.081	0.813	0.953	0.983
360MonoDepth^{M3}	0.161	0.363	0.607	0.085	0.781	0.951	0.984
PanDA^{504}	0.182	0.348	0.587	0.107	0.757	0.861	0.917
PanDA^{1008}	0.172	0.342	0.567	0.101	0.787	0.905	0.935
FRED(ours)	**0.135**	**0.263**	**0.404**	**0.081**	**0.844**	**0.959**	**0.984**

Table 2: ***Quantitative results on a 8K dataset.*** *FRED outperforms all competing pipelines at native 8K resolution (*8192×4096*), demonstrating high geometric fidelity and robust accuracy on challenging real indoor scenes. Panda*^{504}*, Panda*^{1008}*: PanDA input patch sizes before upscaling.*

Method	AbsRel↓	MAE↓	RMSE↓	RMSELog↓	$\sigma < 1.25$↑	$\sigma < 1.25^2$↑	$\sigma < 1.25^3$↑
360MonoDepth	0.183	1.260	1.462	0.095	0.667	0.967	0.992
PanDA^{504}	0.271	1.836	2.251	0.265	0.565	0.766	0.821
PanDA^{1008}	0.244	1.576	1.978	0.260	0.643	0.786	0.834
FRED (ours)	**0.181**	**0.666**	**0.992**	**0.094**	**0.754**	**0.968**	**0.993**

accuracy thresholds δ_1, δ_2, δ_3—computed only on valid ground-truth pixels for fair comparison. Results in Table 1 and Table 2 report accuracy on the synthetic *Replica360-4K* [RAYR22] and a real-world 8K benchmark, covering both medium- and ultra-high-resolution scenarios. On *Replica360-4K*, FRED surpasses competing high-resolution baselines such as *360MonoDepth* [RAYR22] and *PanDA* [CZZ*25], achieving the best results across all metrics. On the real-world 8K dataset (Table 2), FRED further widens the margin, outperforming both conventional and foundation-model approaches. It maintains strong geometric fidelity at 8K resolution, yielding lower AbsRel and RMSE and demonstrating excellent scalability without retraining, dataset-specific tuning, or heavy optimization-based post-processing. Overall, these results confirm that FRED combines superior accuracy, efficiency, and scalability for high-resolution 360° depth estimation.

4.3. Qualitative Assessment

Fig. 5 and Fig. 6 present representative qualitative comparisons on *Replica360-4K* and a real-world 8K dataset. On *Replica360*, FRED produces visibly sharper and more coherent depth maps than *360MonoDepth* [RAYR22], particularly along object boundaries, thin structures, and large planar regions. While the competing method often exhibits cross-projection artifacts and local discontinuities between adjacent patches, FRED achieves smooth yet accurate transitions across the full 360° panorama, successfully preserving both the global scene layout and fine geometric details.

The improvements are even more evident on the real-world 8K dataset, which contains 58 real panoramic captures at ultra-high resolution, with ground-truth provided by a Lidar acquistion. Here, FRED demonstrates strong robustness to visual complexity, specular surfaces, and large depth ranges, maintaining high-frequency detail and correct spatial ordering across distant regions. This behavior confirms the effectiveness of our geometry-consistent fusion prior, which enforces local–global coherence even under severe equirectangular distortions. Moreover, our predictions remain stable across varying lighting conditions and scene textures, avoiding the smoothing and bleeding artifacts observed in projection-based baselines.

Table 3: ***Ablation on*** **Replica360-4K.** *Different weighting solutions for fusion are compared.*

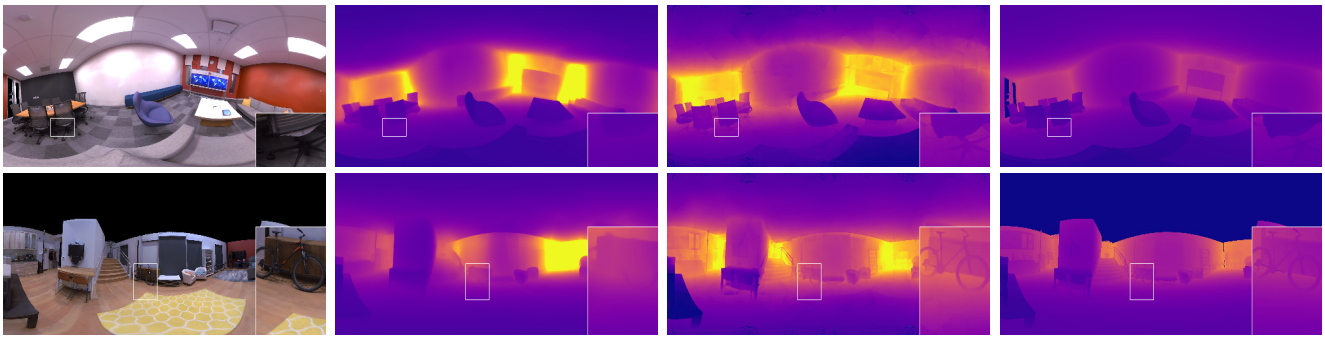
Configuration	MAE↓	RMSE↓	δ_1↑
Uniform weights	0.305	0.468	0.806
Tukey window only	0.284	0.431	0.832
Prompt+Tukey (ours)	**0.269**	**0.412**	**0.841**

Overall, qualitative results reinforce the quantitative findings reported in Sec. 4.2, showing that FRED consistently delivers sharper, geometrically accurate, and perceptually faithful depth reconstructions across both synthetic and real-world high-resolution panoramas.

4.4. Ablation Experiments

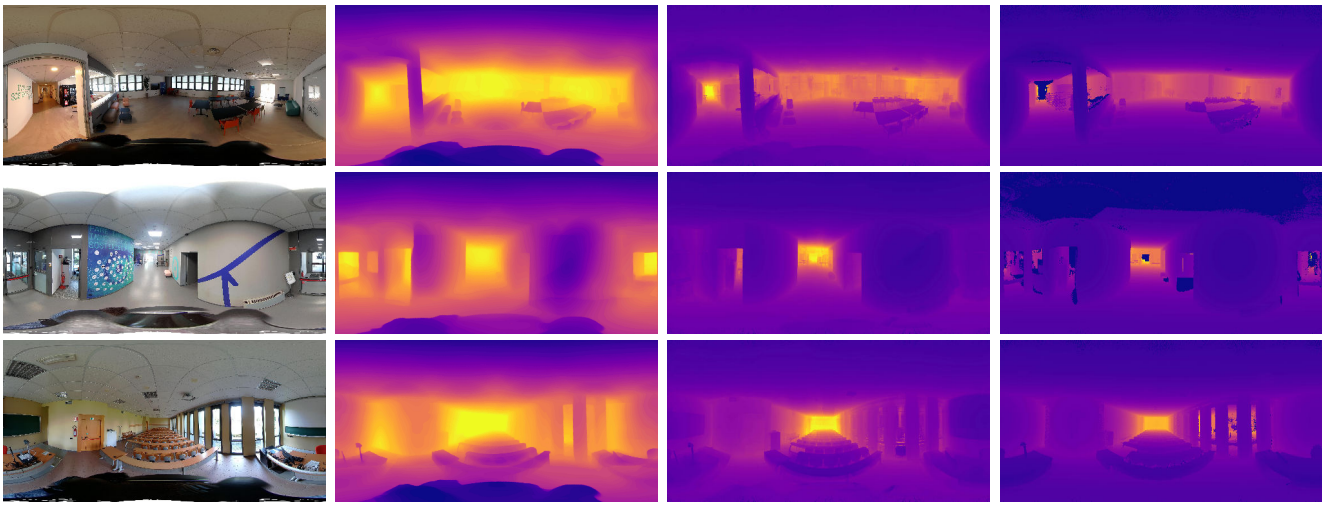
Replica360-4K. We assess the contribution of each component of FRED on the *Replica360-4K* dataset under identical, training-free conditions. As reported in Table 3, removing either the prompt-guided refinement or the Tukey window disrupts cross-patch coherence: uniform weights create visible seams, whereas Tukey-only blending reduces noise but still suffers from local inconsistencies. Their combination provides the smoothest transitions and the largest RMSE drop. All the examples use the default setup with 1024×1024 gnomonic patches with a 100° field of view and 6° overlap.

Additional ablations. We added additional ablation exper-



RGB Input | 360MonoDepth [RAYR22] | **FRED (ours)** | Ground Truth

Figure 5: ***Qualitative comparison on the* Replica360-4K *dataset.*** *Each row shows a representative panoramic scene with its RGB input, the prediction from* 360MonoDepth *[RAYR22], our prediction, and the corresponding ground-truth depth.* FRED *preserves both global geometry and fine spatial detail, producing sharper, more consistent, and artifact-free high-resolution depth maps.*



RGB Input | 360MonoDepth [RAYR22] | **FRED (ours)** | Ground Truth

Figure 6: ***Qualitative comparison on a real-world 8K dataset and on zero-shot scenes.*** FRED *shows superior reconstruction fidelity at ultra-high resolutions (*8192×4096*), maintaining structural consistency and fine-grained geometry even in challenging lighting and reflective conditions.*

Table 4: ***Ablation results on the real-world 8K dataset.***

Method	**AbsRel**↓	**MAE**↓	**RMSE**↓	**RMSELog**↓	$\sigma < 1.25$↑	$\sigma < 1.25^2$↑	$\sigma < 1.25^3$↑
SliceNet [PAA*21]	0.183	0.75	1.196	0.125	0.732	0.901	0.954
Unifuse [JSZ*21]	0.159	0.621	0.951	0.094	0.791	0.945	0.984
GT as lowres	**0.021**	**0.072**	**0.251**	**0.061**	**0.997**	**0.999**	**0.999**

iments comparing different low-resolution estimators, like SliceNet [PAA*21] and UniFuse [JSZ*21], and two additional experiments: FRED without low-resolution prompt guidance, and FRED with ground truth low-resolution prompts.

Prompting vs. no prompting (DINOv2). In the *no prompting* variant, we employ DINOv2 [ODM*24] features for per-patch estimation but intentionally do *not* inject any external LR depth prompt. As shown in Fig. 7 (left), this yields locally plausible detail yet suffers from cross-patch scale drift and imperfect blending. When a low-resolution metric prompt is provided, the global scale is stabilized and seam artifacts are strongly reduced (Fig. 7, right).

Upper bound with GT-as-prompt. Replacing the LR prompt with downsampled ground-truth (used only as a prior) yields the near-perfect scores in Table 4, indicating that our tiling+fusion faithfully transfers LR metric cues to native resolution without sacrificing high-frequency detail.

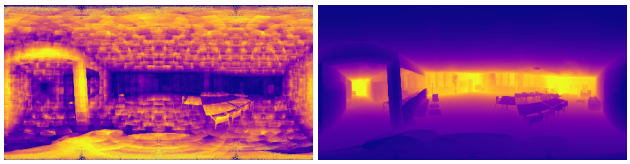


Figure 7: ***Effect of prompting.*** *Left:* No prompting (DINOv2) *— we run the per-patch estimator without injecting any external LR depth. Predictions are locally detailed but exhibit cross-patch scale drift and imperfect blending. Right:* With prompting *— a low-resolution metric prompt provides a global scaffold that stabilizes absolute scale and reduces quilting artifacts, yielding sharper, more consistent depth.*

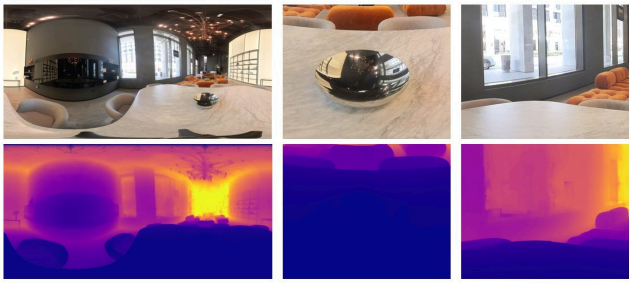


Figure 8: ***Failure case.*** *Highly reflective and transparent surfaces remain difficult. In this example, a metallic bowl is largely missed and large windows are partially misestimated, yielding local depth flattening and halos. Image acquired at BLAUUCASA showroom in Msheireb Downtown Doha with Ricoh Theta X camera. Courtesy of Doha Design District.*

Takeaways. (1) Prompt guidance is crucial for cross-patch scale consistency, even when strong features (DINOv2) are available. (2) The Tukey-weighted fusion effectively suppresses seams in the presence of a metric prior. (3) With a strong LR prior (GT-as-prompt), the pipeline scales to full resolution with minimal loss.

5. Limitations

Despite strong results on in-the-wild images, FRED inherits several limitations:

- **Specular/transparent materials.** Depth priors and foundation estimators struggle with mirrors, glass, glossy metals, and refractive media; our fusion cannot fully compensate when all patches are unreliable (Fig. 8). Note, however, that humans also struggle in these environments; most of us have personally seen persons walk into glass windows and closed doors occasionally.
- **Textureless regions.** Sky, uniform walls, and glossy floors may exhibit residual bias or over-smoothing. While Tukey/raised-cosine tapering reduces seams, absolute depth can drift without reliable cues.
- **Dynamic content and rolling shutter.** Moving people/objects and exposure changes across captures can introduce local inconsistencies across patches and around poles.
- **Prompt mismatch.** When a low-resolution prompt is highly inaccurate (due to domain shift or extreme lighting), consistency weighting may overtrust it. We mitigate this with robust normalization and optional consistency disabling, but large prompt errors can still bias local predictions.
- **Poles and extreme FoV.** Gnomonic sampling near poles or with very wide patch FoVs amplifies reprojection sensitivity; slight pose/grid errors can create faint quilting if overlap is too small.
- **Scale assumptions.** FRED produces metric depth when the prompt (or pseudo-prompt) is metric. If the prompt is non-metric, global scale may drift; a simple one-parameter scale fit to sparse GT or LiDAR can correct this.
- **Compute/memory.** Although embarrassingly parallel across patches, very high-resolution ERPs still require non-trivial GPU memory and I/O bandwidth. Batch size and patch size must respect ViT constraints (multiples of 14) and device limits.

Failure patterns and mitigations. For specular/transparent regions, increasing overlap and using a more conservative Tukey radius can reduce seam artifacts. If a prompt is available, per-patch confidence gating helps down-weight unreliable priors; otherwise, a pseudo-prompt can stabilize inference but does not create new cues for truly ambiguous surfaces. Finally, a lightweight post-refinement (e.g., joint bilateral filtering guided by ERP RGB or a cross-patch CRF) can suppress small residual quilting in low-texture areas.

5.1. Failure Cases

Despite its robustness and scalability, FRED still faces limitations under challenging visual conditions. As illustrated in the supplementary material, errors mainly arise under extreme lighting, on specular or transparent surfaces, and with strongly repetitive textures, where RGB cues provide weak geometric signals and the coarse prior may drift in scale. In some cases, prompt-guided fusion can also amplify minor inconsistencies across overlapping patches or introduce faint seams at high latitudes under severe distortions. These failures suggest that incorporating semantic or normal-based priors could further enhance robustness and surface continuity.

6. Conclusions

We introduced FRED, a training-free framework for full-resolution panoramic depth estimation that unifies global 360° reasoning with local high-fidelity refinement. By decomposing the panorama into uniformly sampled perspective patches guided by a coarse spherical prior and recomposing them via low-cost geometry-consistent fusion, FRED achieves native 8K–60MP inference with accurate, seamless depth reconstruction. Experiments on synthetic and real benchmarks show consistent improvements over state-of-the-art methods in both accuracy and perceptual quality. In conclusion, FRED demonstrates that scalable, high-resolution 360° geometry recovery is feasible without retraining. Future work will explore integration with foundation models for fully consistent 3D scene understanding and neural rendering systems.

Acknowledgments This work was supported by NPRP-S 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of

Qatar Foundation). MA also acknowledges support from HBKU Vice Presidence of Research under the Thematic Research Grant HBKU-INT-VPR-TG-03-03-3G: Gaze, Guide, and Grasp: LLM-Augmented Immersive Experiences for Qatar's Heritage Site, while GP and EG acknowledge support from Sardinian Regional Authorities under the XDATA project (Art9 LR 20/2015). ChatGPT was used to improve the manuscript's language, grammar, and flow. We finally acknowledge Doha Design District for providing access to BLAUUCASA showroom in Msheireb Downtown Doha for high resolution image acquisitions. The findings reflect the work and are solely the authors' responsibility.

References

[ACC*23] AI H., CAO Z., CAO Y.-P., SHAN Y., WANG L.: HRDFuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proc. CVPR* (2023), pp. 13273–13282. doi:10.1109/CVPR52729.2023.01275. 3

[ACW25] AI H., CAO Z., WANG L.: A survey of representation learning, optimization strategies, and applications for omnidirectional vision. *Int. J. Comput. Vision 133*, 8 (2025), 4973–5012. doi:10.1007/s11263-025-02391-w. 1, 2, 3

[AW24] AI H., WANG L.: Elite360D: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *Proc. CVPR* (2024), pp. 9926–9935. doi:10.1109/CVPR52733.2024.00947. 3

[CAVW24] CAO Z., AI H., VASILAKOS A. V., WANG L.: 360° high-resolution depth estimation via uncertainty-aware structural knowledge transfer. *IEEE Trans AI 5*, 11 (2024), 5392–5402. doi:10.1109/TAI.2024.3427068. 3

[CCD*18] CHENG H.-T., CHAO C.-H., DONG J.-D., WEN H.-K., LIU T.-L., SUN M.: Cube padding for weakly-supervised saliency prediction in 360° videos. In *Proc. CVPR* (2018), pp. 1420–1429. doi:10.1109/CVPR.2018.00154. 3

[CGZ*25] CHEN S., GUO H., ZHU S., ZHANG F., HUANG Z., FENG J., KANG B.: Video Depth Anything: Consistent depth estimation for super-long videos. In *Proc. CVPR* (2025), pp. 22831–22840. doi:10.1109/CVPR52734.2025.02126. 3

[CZA*25] CAO Z., ZHU J., AI H., JIANG L., LYU Y., XIONG H.: ST^{2}360D: Spatial-to-temporal consistency for training-free 360 monocular depth estimation. In *Proc. NeurIPS* (2025). 3, 5

[CZZ*25] CAO Z., ZHU J., ZHANG W., AI H., BAI H., ZHAO H., WANG L.: PanDA: Towards panoramic Depth Anything with unlabeled panoramas and mobius spatial augmentation. In *Proc. CVPR* (2025), pp. 982–992. doi:10.1109/CVPR52734.2025.00100. 2, 3, 5, 6

[dSPMLJ22] DA SILVEIRA T. L., PINTO P. G., MURRUGARRA-LLERENA J., JUNG C. R.: 3D scene geometry estimation from 360° imagery: A survey. *ACM Computing Surveys 55*, 4 (2022), 1–39. 2

[ESLF20] EDER M., SHVETS M., LIM J., FRAHM J.-M.: Tangent images for mitigating spherical distortion. In *Proc. CVPR* (2020). doi:10.1109/CVPR42600.2020.01244. 3

[GPA24] GOBBETTI E., PINTORE G., AGUS M.: Automatic 3D modeling and exploration of indoor structures from panoramic imagery. In *SIGGRAPH Asia Courses* (2024), pp. 1:1–1:9. doi:10.1145/3680532.3689580. 1, 2

[HCCJ17] HUANG J., CHEN Z., CEYLAN D., JIN H.: 6-DOF VR videos with a single 360-camera. In *Proc. IEEE VR* (2017), pp. 37–44. doi:10.1109/VR.2017.7892229. 2

[HSC*25] HUANG C., SHAO F., CHEN H., MU B., XU L.: GADFNet: Geometric priors assisted dual-projection fusion network for monocular panoramic depth estimation. *IEEE Trans. Circuits and Systems for Video Technology 35*, 9 (2025), 9060–9074. doi:10.1109/TCSVT.2025.3553472. 3

[IHR*16] IM S., HA H., RAMEAU F., JEON H.-G., CHOE G., KWEON I. S.: All-around depth from small motion with a spherical panoramic camera. In *Proc. ECCV* (2016), pp. 156–172. doi:10.1007/978-3-319-46487-9_10. 2

[JOV19] JOKELA T., OJALA J., VÄÄNÄNEN K.: How people use 360-degree cameras. In *Proc. Int. Conf. Mobile and Ubiquitous Multimedia* (2019), pp. 1–10. doi:10.1145/3365610.3365645. 2

[JSZ*21] JIANG H., SHENG Z., ZHU S., DONG Z., HUANG R.: UniFuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters 6*, 2 (2021), 1519–1526. doi:10.1109/LRA.2021.3058957. 3, 7

[LGY*22] LI Y., GUO Y., YAN Z., HUANG X., DUAN Y., REN L.: Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proc. CVPR* (2022), pp. 2801–2810. doi:10.1109/CVPR52688.2022.00282. 3

[LHS*20] LUO X., HUANG J.-B., SZELISKI R., MATZEN K., KOPF J.: Consistent video depth estimation. *ACM Trans. Graph. 39*, 4 (2020). doi:10.1145/3386569.3392377. 3

[LPC*25] 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 *Proc. CVPR* (2025), pp. 17070–17080. doi:10.1109/CVPR52734.2025.01591. 2, 3, 4, 5

[ODM*24] OQUAB M., DARCET T., MOUTAKANNI T., VO H. V., SZAFRANIEC M., KHALIDOV V., FERNANDEZ P., HAZIZA D., MASSA F., EL-NOUBY A., ASSRAN M., BALLAS N., GALUBA W., HOWES R., HUANG P.-Y., LI S.-W., MISRA I., RABBAT M., SHARMA V., SYNNAEVE G., XU H., JEGOU H., MAIRAL J., LABATUT P., JOULIN A., BOJANOWSKI P.: DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024). Featured Certification. URL: https://openreview.net/forum?id=a68SUt6zFt. 7

[PAA*21] PINTORE G., AGUS M., ALMANSA E., SCHNEIDER J., GOBBETTI E.: SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. CVPR* (2021), pp. 11536–11545. doi:10.1109/CVPR46437.2021.01137. 3, 7

[PASG25] PINTORE G., AGUS M., SIGNORONI A., GOBBETTI E.: DDD++: Exploiting density map consistency for deep depth estimation in indoor environments. *Graphical Models 140* (2025), 101281. doi:10.1016/j.gmod.2025.101281. 3

[PASG26] PINTORE G., AGUS M., SCHNEIDER J., GOBBETTI E.: State-of-the-art in deep learning approaches for automatic single-panorama indoor modeling and exploration. *Computer Graphics Forum 45*, 2 (2026). doi:10.1111/cgf.70396. 1, 2, 3

[PJAG25] PINTORE G., JASHARI S., AGUS M., GOBBETTI E.: PanoFloor: reconstruction and immersive exploration of large multi-room scenes from a minimal set of registered panoramic images using denoised density maps. In *Proc. IEEE ISMAR* (2025), pp. 414–424. doi:10.1109/ISMAR67309.2025.00052. 1

[PJVH*24] PINTORE G., JASPE-VILLANUEVA A., HADWIGER M., SCHNEIDER J., AGUS M., MARTON F., BETTIO F., GOBBETTI E.: Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image. *Computers & Graphics 119* (2024), 103907. doi:10.1016/j.cag.2024.103907. 1, 3

[PMG*20] PINTORE G., MURA C., GANOVELLI F., FUENTES-PEREZ L., PAJAROLA R., GOBBETTI E.: State-of-the-art in automatic 3D reconstruction of structured indoor environments. *Comput. Graph. Forum 39*, 2 (2020), 667–699. doi:10.1111/cgf.14021. 1, 2

[PZ23] PENG C.-H., ZHANG J.: High-resolution depth estimation for 360° panoramas through perspective and panoramic depth images registration. In *Proc. WACV* (2023), pp. 3115–3124. doi:10.1109/WACV56688.2023.00313. 3, 5

[RAYR22] REY-AREA M., YUAN M., RICHARDT C.: 360MonoDepth: High-resolution 360° monocular depth estimation. In *Proc. CVPR* (2022), pp. 3762–3772. doi:10.1109/CVPR52688.2022.00374. 2, 3, 5, 6, 7

[SG17] SU Y.-C., GRAUMAN K.: Learning spherical convolution for fast features from 360 imagery. In *Proc. NeurIPS* (2017), pp. 529–539. URL: https://dl.acm.org/doi/abs/10.5555/3294771.3294822. 3

[SG19] SU Y., GRAUMAN K.: Kernel transformer networks for compact spherical convolution. In *Proc. CVPR* (2019), pp. 9434–9443. doi:10.1109/CVPR.2019.00967. 3

[SJT*25] SHAH U., JASHARI S., TUKUR M., HOUSEH M., SCHNEIDER J., PINTORE G., GOBBETTI E., AGUS M.: Virtual staging of indoor panoramic images via multi-task learning and inverse rendering. *IEEE CG&A* (2025), 1–14. doi:10.1109/MCG.2025.3605806. 1

[SLL*22] SHEN Z., LIN C., LIAO K., NIE L., ZHENG Z., ZHAO Y.: PanoFormer: Panorama transformer for indoor 360 depth estimation. In *Proc. ECCV* (2022), pp. 195–211. doi:10.1007/978-3-031-19769-7_12. 3, 5

[SSC21] SUN C., SUN M., CHEN H.-T.: HoHoNet: 360° indoor holistic understanding with latent horizontal features. In *Proc. CVPR* (2021), pp. 2573–2582. doi:10.1109/CVPR46437.2021.00260. 3

[STA*24] SHAH U., TUKUR M., ALZUBAIDI M., PINTORE G., GOBBETTI E., HOUSEH M., SCHNEIDER J., AGUS M.: MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image. In *Proc. CVPRW - OmniCV* (2024), pp. 1311–1321. doi:10.1109/CVPRW63382.2024.00138. 2, 3, 4, 5

[SWM*19] STRAUB J., WHELAN T., MA L., CHEN Y., WIJMANS E., GREEN S., ENGEL J. J., MUR-ARTAL R., REN C., VERMA S., CLARKSON A., YAN M., BUDGE B., YAN Y., PAN X., YON J., ZOU Y., LEON K., CARTER N., BRIALES J., GILLINGHAM T., MUEGGLER E., PESQUEIRA L., SAVVA M., BATRA D., STRASDAT H. M., NARDI R. D., GOESELE M., LOVEGROVE S., NEWCOMBE R.: The Replica Dataset, 2019. [Online; accessed 2025-09-17]. URL: https://github.com/facebookresearch/Replica-Dataset. 6

[SZL*23] SHEN Z., ZHENG Z., LIN C., NIE L., LIAO K., ZHENG S., ZHAO Y.: Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In *Proc. CVPR* (2023), pp. 17337–17345. doi:10.1109/CVPR52729.2023.01663. 2, 3

[TBJ*24] TUKUR M., BORAEY Y., JASHARI S., VILLANUEVA A. J., SHAH U., AL-ZUBAIDI M., PINTORE G., GOBBETTI E., SCHNEIDER J., AGUS M., FETAIS N.: Virtual staging technologies for the metaverse. In *Proc. iMETA* (2024), pp. 206–213. doi:10.1109/iMETA62882.2024.10807986. 2

[TNT18] TATENO K., NAVAB N., TOMBARI F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proc. ECCV* (2018), pp. 732–750. doi:10.1007/978-3-030-01270-0_43. 3

[TUP*23] TUKUR M., UR REHMAN A., PINTORE G., GOBBETTI E., SCHNEIDER J., AGUS M.: PanoStyle: Semantic, geometry-aware and shading independent photorealistic style transfer for indoor panoramic scenes. In *Proc. ICCVW* (2023), pp. 1553–1564. doi:10.1109/ICCVW60793.2023.00170. 1

[WHC*18] WANG F.-E., HU H.-N., CHENG H.-T., LIN J.-T., YANG S.-T., SHIH M.-L., CHU H.-K., SUN M.: Self-supervised learning of depth and camera motion from 360 videos. In *Proc. ACCV* (2018), pp. 53–68. doi:10.1007/978-3-030-20873-8_4. 3

[WHZ*25] WANG X., HE Z., ZHANG Q., YANG Y., ZHAO T., JIANG J.: Geometry-aware self-supervised indoor 360° depth estimation via asymmetric dual-domain collaborative learning. *IEEE Trans. Multimedia 27* (2025), 3224–3237. doi:10.1109/TMM.2025.3535340. 3

[WL24] WANG N.-H. A., LIU Y.-L.: Depth Anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *Proc. NeurIPS 37* (2024), 127739–127764. URL: https://dl.acm.org/doi/10.5555/3737916.3741972. 3, 5

[WYS*20] WANG F.-E., YEH Y.-H., SUN M., CHIU W.-C., TSAI Y.-H.: BiFuse: Monocular 360 depth estimation via bi-projection fusion. In *Proc. CVPR* (2020), pp. 462–471. doi:10.1109/CVPR42600.2020.00054. 3

[WYT*22] WANG F.-E., YEH Y.-H., TSAI Y.-H., CHIU W.-C., SUN M.: BiFuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. *IEEE TPAMI 45*, 5 (2022), 5448–5460. doi:10.1109/TPAMI.2022.3203516. 3

[YJL*18] YANG Y., JIN S., LIU R., , YU J.: Automatic 3D indoor scene modeling from single panorama. In *Proc. CVPR* (2018), pp. 3926–3934. doi:10.1109/CVPR.2018.00413. 2

[YWZ*25] YAN Q., WANG Q., ZHAO K., CHEN J., LI B., CHU X., DENG F.: Spherefusion: Efficient panorama depth estimation via gated fusion. In *Proc. 3DV* (2025), pp. 855–865. doi:10.1109/3DV66043.2025.00084. 2, 3

[ZKZ*19] ZIOULIS N., KARAKOTTAS A., ZARPALAS D., ALVAREZ F., DARAS P.: Spherical view synthesis for self-supervised 360° depth estimation. In *Proc. 3DV* (2019). doi:10.1109/3DV.2019.00081. 3

[ZKZD18] ZIOULIS N., KARAKOTTAS A., ZARPALAS D., DARAS P.: OmniDepth: Dense depth estimation for indoors spherical panoramas. In *Proc. ECCV* (2018), pp. 453–471. doi:10.1007/978-3-030-01231-1_28. 3