Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Training-Free Monocular-Stereo Depth Fusion with Sparse Geometric Anchors for Robust Depth Perception

Shuaijia Chen
Jilin University
Xiucai Zhang
Jilin University
Yang Liu
liu_yang@jlu.edu.cn
Jilin University

Research Article

Keywords: Depth perception, cross-modal visual cue fusion, monocular-stereo fusion, sparse geometric constraints, edge-aware propagation, zero-shot generalization

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10678705/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Training-Free Monocular-Stereo Depth Fusion with Sparse Geometric Anchors for Robust Depth Perception

Shuaijia Chen[1], Xiucai Zhang[1], Yang Liu[1*]

[1]College of Instrumentation and Electrical Engineering, Jilin University, Changchun, 130012, Jilin, China.

*Corresponding author(s). E-mail(s): liu_yang@jlu.edu.cn;

Abstract

Depth perception is a fundamental visual computing task that requires combining dense scene structure with reliable metric geometry. Monocular depth estimation models provide dense structural cues but lack calibrated scale, whereas stereo matching supplies metric geometric constraints that may be sparse or unreliable in low-texture and occluded regions. We propose a training-free monocular-stereo visual cue fusion framework that combines these complementary cues without task-specific training, fine-tuning, or iterative optimization. An off-the-shelf monocular depth estimation model first predicts a dense relative depth prior. In parallel, Census-based stereo matching followed by left-right consistency checking produces sparse, high-confidence disparity anchors. Cross-modal scale alignment maps the monocular structure into the disparity domain through affine regression, and an edge-aware propagation mechanism refines the aligned prior under the guidance of the left view. On the combined Middlebury 2014 and ETH3D benchmarks, the proposed method achieves competitive edge-region accuracy and outperforms PSMNet and GraftNet in edge-region errors under zero-shot evaluation. Although RAFT-Stereo remains the strongest overall pre-trained cross-domain baseline, our results show that fusing monocular structure with sparse stereo geometry enables robust depth perception without target-domain training.

Keywords: Depth perception, cross-modal visual cue fusion, monocular-stereo fusion, sparse geometric constraints, edge-aware propagation, zero-shot generalization

1 Introduction

Depth perception is a fundamental capability in visual computing, supporting applications such as autonomous systems, 3D reconstruction, and scene understanding. A practical depth perception system should recover reliable metric scene geometry while preserving object boundaries and remaining robust to low-texture regions, occlusion, and unknown deployment conditions.

Stereo matching provides metric geometric cues from binocular observations, but its reliability degrades in low-texture, occluded, and repetitive-texture regions, resulting in sparse or noisy disparity estimates. Monocular depth estimation provides dense structural cues from a single image, but its output is relative and scale-ambiguous. These two visual cues are therefore complementary: stereo provides calibrated geometry, while monocular estimation offers dense scene structure. Learning-based depth models [1–3] can improve individual estimates, but they still suffer from scale shifts and domain gaps when deployed to new devices or environments.

Existing fusion methods typically require large-scale training data, target-specific fine-tuning, expensive optimization, or sensor-specific supervision. Such requirements limit deployment in new scenes and hinder the integration of off-the-shelf visual models into general visual computing pipelines. A training-free approach that explicitly fuses monocular structure with sparse stereo geometry remains underexplored.

To address the above issues, we propose a training-free fusion framework that combines a dense monocular structural prior with sparse metric stereo anchors. The monocular prior is transferred into the calibrated disparity domain through cross-modal affine alignment and further refined by an edge-aware propagation mechanism guided by the left view. The resulting dense disparity can be converted to metric depth through camera calibration, without task-specific training or iterative optimization.

Extensive zero-shot experiments on Middlebury 2014 and ETH3D demonstrate that the proposed framework achieves competitive edge-region accuracy and robust generalization without target-domain training.

The main contributions of this paper are summarized as follows:

- We propose a training-free monocular-stereo visual cue fusion framework that combines dense monocular structure with sparse metric stereo geometry, requiring no task-specific training, fine-tuning, or iterative optimization.
- We introduce cross-modal scale alignment based on sparse stereo anchors, which transfers monocular structural priors into the calibrated disparity domain and establishes consistency between the two visual cues.
- We integrate an edge-aware geometric propagation module based on guided filtering, which propagates reliable sparse geometry along scene structure while preserving depth boundaries, and we provide systematic analyses of sensitivity to sparse geometric anchors and generalization across monocular structural priors.

2 Related Work

2.1 Neural Network-Based Depth Estimation

Deep-learning-based monocular depth estimation (MDE) has become an active research area in computer vision [4]. Unlike traditional multi-view geometry methods, MDE estimates a dense depth map from a single two-dimensional image [3], using data priors to compensate for the inherent ambiguity of the problem.

Early supervised approaches such as Eigen et al. introduced multi-scale deep architectures for single-image depth prediction [5], but they require pixel-level depth labels and have limited deployment flexibility. Although recent metric depth models such as Depth Pro [6] and UniDepth [7] improve accuracy, they still rely on large-scale training data and specific camera-intrinsic assumptions, limiting their absolute-scale reliability across devices and scenes.

To avoid fine-tuning and reduce deployment cost, recent test-time fusion methods such as Marigold-DC [8] use sparse observations as guidance without additional training, but rely on iterative diffusion sampling and gradient optimization. In contrast, the proposed method performs a single non-iterative propagation pass and combines monocular structure with calibrated stereo cues without task-specific training.

2.2 Stereo Vision and Depth Computation

Stereo vision can provide absolute depth with physical scale, but its performance strongly depends on image texture. In texture-sparse regions such as white walls, sky, and smooth floors, the pixel intensity variation within local windows is small, which makes matching difficult and easily produces incorrect matches or missing disparity. Occlusion, repetitive texture, and illumination changes further degrade matching quality.

To improve robustness, classical stereo methods range from local window matching to semi-global matching (SGM) [9], but they still rely on texture cues. Learning-based methods such as PSMNet [10] and GA-Net [11] improve matching robustness, yet they require high-quality annotated data and often generalize poorly to out-of-domain scenes. In contrast, the proposed framework uses stereo matching only to extract sparse, high-confidence disparity anchors, which are then fused with a dense monocular prior.

2.3 Monocular-Stereo Fusion and Depth Fusion

Monocular depth estimation provides a dense but usually relative depth prior, whereas binocular or sparse depth provides physically scaled absolute observations. Fusing these two cues has therefore become an important problem in monocular-stereo fusion and depth estimation. An early representative work, Self-Supervised Sparse-to-Dense [12], uses a self-supervised approach to combine monocular images with sparse LiDAR depth for dense depth prediction, demonstrating the complementarity of dense relative structure and monocular image information. However, it still requires training data and in-domain supervision and cannot be directly transferred to new sensors or scenes.

In recent years, several works have begun to explicitly use monocular structural priors to guide sparse depth or multi-view stereo tasks. Methods such as BridgeDepth [13] and MonoMVSNet [14] introduce monocular depth estimation as an auxiliary cue into matching or fusion processes, but remain end-to-end learning approaches that rely on training data and camera parameters.

Unlike these approaches, the proposed method does not train an additional network or perform iterative test-time optimization; it uses a pre-trained monocular model only as a structural prior and achieves dense depth fusion, scale transfer, and edge preservation with a single edge-aware propagation pass.

3 Method

The proposed method is a training-free monocular-stereo visual cue fusion framework. The overall pipeline is illustrated in Fig. 1.

3.1 Monocular Depth Prior Acquisition

We use Depth Anything V2 with the ViT-Base encoder as the backbone network for monocular depth estimation [15]. The model is pre-trained on large-scale heterogeneous data and provides dense depth prediction in diverse scenes, maintaining accurate structural consistency in weak-texture regions and at object boundaries. Its output is a relative depth prior. Since the proposed method only requires the relative depth relationship as a structural prior and does not require exact absolute scale, the relative output is used directly. Numerically, the depth prior only reflects the relative depth ordering of scene points and does not correspond to true physical distances.

The raw monocular depth priors produced by Depth Anything V2 vary in numerical range across scenes and may contain local noise and uncertainty. To improve the stability of subsequent fusion, the raw depth prior is first normalized per image to $[0, 1]$ to eliminate numerical range differences across scenes while preserving relative structure information, including object occlusion relationships and depth-step positions. Although Depth Anything V2 itself has some robustness to noise, local depth jitter can still occur in low-texture or complex illumination regions. To suppress such noise while preserving depth edges, bilateral filtering [16] is applied to the normalized depth prior. The advantage of bilateral filtering is that it smooths continuous flat regions while maintaining sharp transitions at depth edges, thereby better preserving the geometric structure of the scene.

3.2 Sparse Stereo Anchor Generation

To obtain sparse metric geometric constraints, we extract reliable stereo anchors from binocular images. We adopt a local window matching method based on the Census transform, which is robust to illumination changes and radiometric distortion [17] and is suitable for diverse Middlebury scenes. Specifically, for each pixel in the left image, the best matching point is searched along the horizontal direction in the right image, and the matching cost is the Hamming distance of the Census-transformed descriptors.

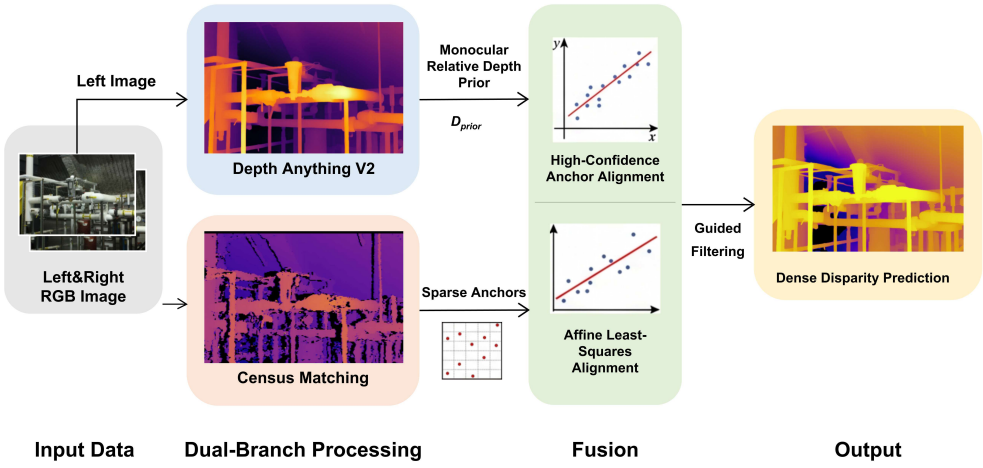

Fig. 1 Overview of the proposed training-free monocular-stereo visual cue fusion framework. The framework combines dense monocular structural priors with sparse stereo disparity anchors through cross-modal affine alignment and edge-aware guided refinement, producing a dense disparity prediction without additional training

A winner-takes-all strategy selects the disparity with the minimum cost to obtain the initial disparity map.

Because local matching methods are prone to incorrect matches in occluded, repetitive-texture, and weak-texture regions, directly using the initial disparity map will introduce significant depth errors. To filter out unreliable matching points, we adopt left-right consistency checking as the main quality control method. The left-right and right-left disparity maps are computed separately. Pixels satisfying $|D_L - D_R| \leq 1$ are retained, and connected components smaller than 50 pixels are removed. A joint bilateral consistency check is then applied with radius 6, $\sigma_s = 4.0$, $\sigma_r = 0.03$, and a residual threshold of 1.2 pixels. After these steps, only high-confidence disparity points are retained as sparse calibrated stereo disparity anchors.

3.3 Cross-Modal Visual Cue Fusion

The proposed fusion framework combines the calibrated geometric information provided by binocular stereo with the dense structural prior provided by monocular depth estimation. Guided filtering is used as an edge-aware propagation mechanism, rather than as a novel filtering operator; its role is to refine the aligned monocular prior under structural guidance from the left view. The core idea involves treating the relative depth prior produced by the neural network as a monocular structural cue, using sparse calibrated stereo disparity values as metric constraints, aligning the monocular structure through cross-modal affine regression, and propagating reliable geometric information under the guidance of the left view.

Unlike Marigold-DC, which performs depth fusion through iterative diffusion sampling and gradient optimization, the proposed method uses forward guided filtering without iterative computation, avoiding iterative optimization while maintaining accuracy.

3.3.1 Basic Principles of Edge-Aware Propagation

In this framework, guided filtering is the implementation of the proposed edge-aware propagation mechanism. It is a local linear filtering model that assumes the output image and the guidance image satisfy a linear relationship within a local window:

$$Q_i = a_k G_i + b_k, \quad \forall i \in \omega_k, \tag{1}$$

where the coefficients a_k and b_k are obtained by minimizing the difference between the output Q and the input target image T within each window

$$E(a_k, b_k) = \sum_{i \in \omega_k} \left((a_k G_i + b_k - T_i)^2 + \varepsilon a_k^2 \right), \tag{2}$$

where ε is a regularization parameter.

Under this local linear model, strong gradients in the guidance image are preserved in the output. The filtering process smooths flat regions and enables reliable anchor values to propagate to neighboring pixels. This enables edge-aware propagation without explicit edge detection or learnable parameters [18], as illustrated in Fig. 2.

3.3.2 Input and Output Design of the Fusion Framework

In the proposed method, the three core signals are defined as follows:

- **Guidance image** G: the grayscale version of the original left view. The left view contains rich texture and edge information of the scene. Using it as the guidance image ensures that the final depth map maintains sharp transitions at object contours.
- **Target image** T: the scale-aligned monocular structural prior. This map provides a dense structural prior containing only relative depth information and is the main input to guided filtering.
- **Sparse anchors** S: high-confidence binocular disparity points retained after left-right consistency checking. These points serve as metric geometric constraints with calibrated stereo disparity scale (in pixels), but they are sparse and occupy only a small fraction of the pixels in the image.

The output Q of edge-aware propagation is an edge-preserving refined version of T. Because T is initialized from the scale-aligned monocular prior and the sparse calibrated disparity anchors, Q remains in the disparity domain and follows the anchor-derived scale during propagation. Note that guided filtering is a local linear smoothing operator rather than a hard-constraint operator, so Q is not required to equal the input disparities exactly at anchor positions; the anchors act as reliable metric-scale initialization and local guidance rather than hard equality constraints.

3.3.3 Cross-modal Alignment and Sparse Anchor Injection

Because the monocular structural prior is a relative depth prior normalized per image to $[0, 1]$ while the binocular disparity points are calibrated stereo disparity values in pixels, we first estimate an affine mapping from valid anchors:

$$(\alpha, \beta) = \arg\min_{\alpha,\beta} \sum_{p \in S} \left(\alpha D_{\text{prior}}(p) + \beta - D_{\text{stereo}}(p)\right)^2, \tag{3}$$

where $D_{\text{stereo}}(p)$ denotes the stereo disparity at anchor p. The fitted mapping is then applied to the full image and clipped to the valid disparity range:

$$D_{\text{aff}}(x, y) = \alpha D_{\text{prior}}(x, y) + \beta, \quad D_{\text{align}}(x, y) = \text{clip}\left(D_{\text{aff}}(x, y),\, 0,\, d_{\max}^{Q}\right), \tag{4}$$

where $d_{\max}^{Q}$ denotes the maximum disparity after resizing to Q-size. The target image T is initialized as D_{align}, and anchor pixels are replaced by their corresponding stereo disparities: $T(p) = D_{\text{stereo}}(p)$ for $p \in S$. This anchor-injected initialization provides reliable metric-scale guidance to guided filtering while preserving the dense structure of the monocular prior.

3.3.4 Edge-Aware Geometric Propagation

In this framework, edge-aware propagation is implemented by guided filtering. Using the grayscale original left view G as the guidance image, this operation is performed on the target image T:

$$Q = \text{GuidedFilter}(G, T, r, \varepsilon), \tag{5}$$

where $\text{GuidedFilter}(\cdot)$ denotes the guided filtering operation.

Here, the filtering operation is implemented with `imguidedfilter` using NeighborhoodSize 3×3 and DegreeOfSmoothing 0.001, with the grayscale original left view as the guidance image. Under the edge-preservation constraint, these settings mainly prevent cross-boundary diffusion; local jitter suppression is mainly handled by the bilateral filtering described in Section 3.1.

The propagated output Q is the final dense disparity map. Its edge-preserving effect originates from the local linear model of the guidance-based propagation. When the guidance image G has a strong gradient at an object edge, the model must maintain a step-like transition in the output Q at the corresponding position to preserve fitting accuracy, rather than smoothing the edge out. At the same time, in weak-texture regions (e.g., white walls and floors), the guidance image lacks strong gradients, so the model degenerates into ordinary smoothing that removes small jitter in the monocular structural prior and recovers continuous depth planes.

Because guided filtering is a local linear smoothing operator, the values at anchor positions are not enforced to remain exactly unchanged in Q. We therefore treat the sparse anchors as reliable metric-scale initialization and soft guidance rather than hard equality constraints. The affine alignment already transfers the calibrated disparity scale globally before propagation, and this soft-constraint design avoids injecting isolated stereo outliers into the final disparity map.

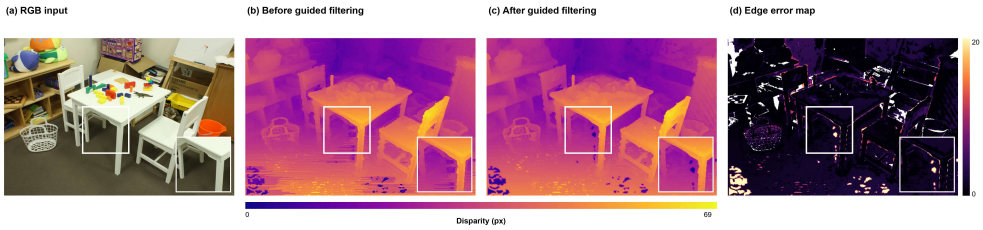

Fig. 2 Visual effect of guided filtering on edge-aware disparity refinement. Guided filtering suppresses local noise in smooth regions while preserving disparity discontinuities around object boundaries

The entire propagation process does not involve any learnable parameters. The edge-preserving behavior is fully driven by the structure of the guidance image and does not require scene-specific parameter tuning, improving portability across different scenes.

3.4 Summary of the Algorithm Pipeline

The complete pipeline of the proposed method can be summarized as follows:

1. Monocular depth prior preprocessing: normalize to $[0, 1]$ and apply bilateral filtering to suppress local jitter.
2. Sparse anchor extraction: use Census matching and left-right consistency checking to obtain high-confidence sparse disparity points.
3. Global scale alignment: perform linear regression on the monocular structural prior using the sparse anchors so that its numerical range is aligned with the disparity scale.
4. Anchor-initialized disparity construction: use the scale-aligned monocular structural prior as the initial target image T and inject the corresponding disparity values at anchor positions.
5. Edge-aware propagation: use the grayscale original left view as the guidance image G, perform edge-aware propagation on the target image T, and output the final disparity map.
6. Depth conversion: convert the disparity map into a physical depth map according to the binocular camera baseline, focal length, and other parameters.

The fusion stage requires no additional task-specific training or fine-tuning; it needs only a single edge-aware propagation pass, aside from monocular prior generation and sparse stereo matching, thus avoiding iterative test-time optimization.

4 Experiments and Evaluation

4.1 Experimental Setup

We evaluate the proposed method on the 15 public ground-truth scenes of Middlebury 2014 [19] and the 27 public ground-truth scenes of ETH3D [20]. All methods are evaluated zero-shot without fine-tuning.

Global MAE and Global RMSE measure the overall disparity-error magnitude, where lower values indicate better global accuracy. Edge-MAE and Edge-RMSE quantify disparity errors within edge regions and therefore indicate how accurately object-boundary transitions are preserved, where lower values are better. Bad 2.0 denotes the percentage of pixels with disparity error larger than 2 pixels, where lower values are better. Ratio-MAE and Ratio-RMSE reflect scale-consistency errors, where lower values indicate closer agreement with the ground-truth disparity scale. GRR measures the ratio between predicted and ground-truth gradient magnitudes in edge regions; because all reported values exceed 1, lower GRR values indicate less edge over-sharpening. F1 summarizes the overlap between predicted and ground-truth disparity edges, where higher values indicate better edge localization.

Implementation details. The monocular backbone is Depth Anything V2 with the ViT-Base encoder (`depth_anything_v2_vitb.pth`). All Middlebury images were evaluated at Q-size. For each scene, the full-resolution disparity bound was divided by four, and the disparity search range was $[1, \mathrm{round}(\mathrm{ndisp_orig}/4)]$. Census matching used a 9×7 Census transform and 7×7 cost aggregation (window radius 3). Left-right consistency retained pixels with $|D_L - D_R| \leq 1$, and connected components smaller than 50 pixels were removed. A joint bilateral consistency check was then applied with radius 6, $\sigma_s = 4.0$, $\sigma_r = 0.03$, and a residual threshold of 1.2 pixels. The monocular prior was normalized per image to $[0, 1]$ and aligned to the disparity domain by ordinary least-squares affine regression using valid anchors (at least 100 points), and the aligned disparity was clipped to $[0, d^Q_{\max}]$. Guided filtering was implemented with `imguidedfilter` using NeighborhoodSize 3×3 and DegreeOfSmoothing 0.001, with the grayscale left view as guidance. Evaluation was performed on the Q-size disparity maps, excluding ground-truth pixels that are zero or infinite.

4.2 Zero-Shot Depth Perception Evaluation

To investigate whether a training-free monocular-stereo fusion framework can achieve robust depth perception across indoor and outdoor scenes under zero-shot conditions, we select three representative methods for comparison: PSMNet as a classical end-to-end supervised stereo matching method, RAFT-Stereo [21] as a strong pre-trained cross-domain iterative stereo baseline, and GraftNet [22] as a domain-generalization method specifically designed for stereo matching. All compared methods use their official pre-trained weights and are directly evaluated on Middlebury 2014 and ETH3D without fine-tuning, consistent with the zero-shot deployment setting of the proposed method.

PSMNet [10] uses the official KITTI 2015 pre-trained weights (pre-trained_model_KITTI2015.tar). GraftNet [22] uses the author-provided checkpoint_final_10epoch.tar with a frozen VGG16 backbone (vgg16-397923af.pth from the PyTorch model zoo). RAFT-Stereo [21] uses the official realtime variant (raftstereo-realtime.pth, release v1.0) with a shared backbone, two GRU layers, and $3\times$ downsampling. All baseline models are used exactly as provided by the respective authors, with no fine-tuning or domain adaptation.

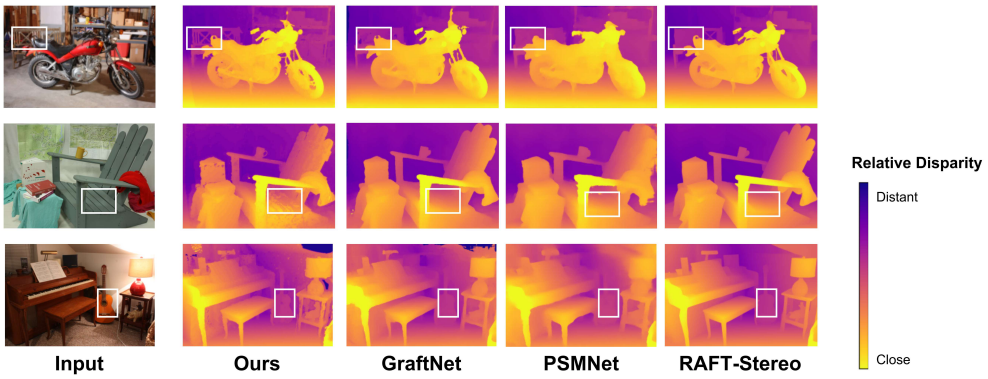

Fig. 3 Qualitative comparison of disparity maps on representative scenes of the Middlebury 2014 and ETH3D benchmarks. ROI boxes highlight local regions that contain both rich texture and clear object boundaries, where edge-preservation differences are most visible across methods

Table 1 Global accuracy comparison on the combined Middlebury 2014 and ETH3D benchmarks

Method	Bad 2.0	Global RMSE	Global MAE
Proposed	14.45%	2.62	1.32
PSMNet	41.05%	5.39	3.03
GraftNet	9.62%	6.45	2.12
RAFT-Stereo	7.15%	2.17	0.87

Table 2 Edge-preservation comparison on the combined Middlebury 2014 and ETH3D benchmarks. Edge-RMSE and Edge-MAE are computed within edge regions; F1 summarizes the agreement between predicted and ground-truth edges

Method	Edge-RMSE	Edge-MAE	Ratio-RMSE	Ratio-MAE	GRR	F1
Proposed	2.72	2.04	1.37	1.55	15.69	0.72
PSMNet	5.66	3.58	1.10	1.23	17.67	0.68
GraftNet	6.25	2.66	1.02	1.30	26.85	0.73
RAFT-Stereo	2.83	1.47	1.53	1.97	5.44	0.77

All experiments were run on Windows 11 with an Intel Core i7-14650HX, 16 GB RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU (8 GB), using Python 3.12.7, PyTorch 2.5.1, and CUDA 12.1.

Fig. 3 shows representative qualitative comparisons.

On the combined benchmarks, the proposed method achieves an Edge-RMSE of 2.72 and an Edge-MAE of 2.04, outperforming PSMNet (5.66/3.58) and GraftNet (6.25/2.66) in edge-region errors, while RAFT-Stereo remains the strongest overall pre-trained cross-domain baseline with lower global errors. GRR measures the gradient-response ratio; values closer to 1 are ideal. GRR is mainly used to analyze gradient amplification behavior rather than serving as an independent accuracy metric. Since all reported values are larger than 1, lower values indicate less oversharpening. Our

GRR (15.69) is lower than those of PSMNet (17.67) and GraftNet (26.85). Compared with the iterative RAFT-Stereo baseline (GRR 5.44), the proposed method still has a substantial gap, reflecting the non-iterative nature of a single edge-aware propagation pass.

From an engineering perspective, although RAFT-Stereo remains the strongest overall baseline with lower global errors and a lower Edge-MAE, it relies on a supervised deep stereo network. GraftNet has a relatively low outlier ratio but larger edge errors and higher GRR, indicating stronger overshoot in the evaluated settings. The proposed method offers a more balanced trade-off under the evaluated settings: it does not require target-specific training, its edge errors are better than those of PSMNet and GraftNet, and its GRR remains lower than those of these non-iterative baselines, validating a framework combining sparse stereo anchors, a monocular structural prior, and guided filtering.

4.3 Ablation Study

To analyze how monocular structure, sparse stereo geometry, and edge-aware propagation jointly contribute to depth perception, we conduct systematic ablation experiments on all ground-truth scenes from the above datasets (i.e., Middlebury 2014 and ETH3D). Five variants are designed to isolate the contribution of each visual cue and propagation strategy:

- **G1 (baseline)**: A sparse disparity map is generated using Census stereo matching, followed by left-right consistency checking and removal of small invalid regions.
- **G2 (traditional propagation)**: Joint bilateral filtering is applied to the G1 result for consistency refinement, followed by nearest-neighbor initialization of low-confidence disparity regions.
- **G3 (monocular prior propagation)**: Based on the refined G1 result, low-confidence disparity regions are initialized using the scale-aligned monocular structural prior.
- **G4 (adaptive fusion)**: Texture-aware fusion weights are computed from a 9×9 local standard-deviation map as $\mathrm{clamp}((\mathrm{texture}-3)/5, 0, 1)$. Stereo disparity results from G2 are assigned higher weights in high-texture regions, while monocular structural priors receive higher weights in low-texture regions; low-confidence regions are initialized as $0.2 \cdot D_{\mathrm{G2}} + 0.8 \cdot$ aligned prior.
- **G5 (full method)**: Based on G3, guided filtering is applied to the dense scale-aligned disparity prior, producing the final dense disparity result of the proposed framework.

All variants are evaluated on the same scenes using identical parameters.

The quantitative results are listed in Table 3, and Fig. 4 visualizes the error trends across variants.

Key findings of the ablation study:

- The edge-preservation capability of G2 (traditional propagation) clearly deteriorates compared with G1: Edge-MAE increases from 2.65 to 3.06, GRR increases from 16.02 to 35.44, and F1 drops from 0.75 to 0.68. Boundary error increases, gradient

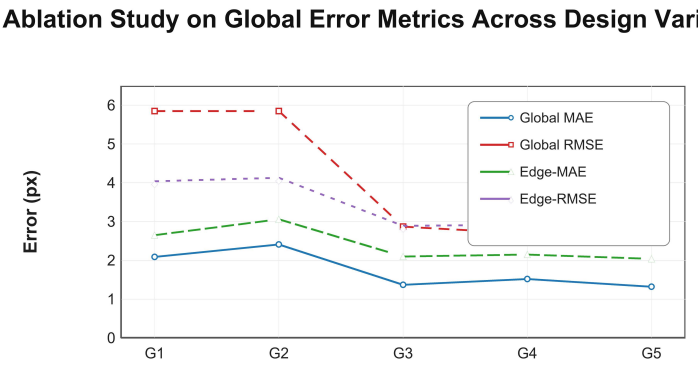

Fig. 4 Ablation study on global error metrics across design variants

Table 3 Ablation study on the combined Middlebury 2014 and ETH3D benchmarks (G1–G5)

Variant	Global MAE	Global RMSE	Bad 2.0	Edge-MAE	Edge-RMSE	GRR	F1
G1	2.09	5.85	27.49%	2.65	4.04	16.02	0.75
G2	2.41	5.85	17.00%	3.06	4.13	35.44	0.68
G3	1.37	2.87	14.22%	2.10	2.89	18.97	0.71
G4	1.52	2.69	18.86%	2.15	2.92	18.95	0.71
G5	1.32	2.62	14.45%	2.04	2.72	15.69	0.72

retention exhibits noticeable overshoot, and edge structure is damaged. This confirms that propagation without a structural prior introduces erroneous depth values at object boundaries and damages edge structure, which is an inherent limitation of traditional post-processing methods.

- G3 (monocular prior propagation) provides significant recovery: Edge-MAE drops to 2.10, Edge-RMSE substantially decreases to 2.89, and F1 recovers to 0.71. The monocular structural prior provides structurally consistent disparity initialization and maintains plausible depth transitions at object boundaries.
- G4 (adaptive fusion) yields mixed results: global RMSE improves from 2.87 to 2.69, whereas MAE, Bad 2.0, and both edge errors worsen slightly. Although adaptive weighting improves local flexibility, it introduces inconsistent boundary behavior compared with guided filtering, indicating that texture-variance-based linear weighting cannot fully integrate the two modalities.
- G5 (full method) achieves the best overall performance: an Edge-MAE of 2.04 and an Edge-RMSE of 2.72 are the lowest among the five variants, and G5 achieves the second-highest F1 score after G1. G5 achieves the best balance among global accuracy, edge preservation, and error-distribution concentration.

The evolution from G1 to G2 to G3 to G5 reveals a core insight: traditional propagation (G2) damages edge structure, monocular prior propagation (G3) restores structural completeness, and edge-aware propagation (G5) further refines edge accuracy. The complete pipeline reduces Edge-RMSE by 32.7% from G1 to G5 (4.04 to 2.72). The isolated G3-to-G5 change attributable to guided filtering is 5.9% (2.89 to 2.72), demonstrating that guided filtering further improves edge quality after anchor-constrained initialization.

In summary, G5 achieves a better balance among global accuracy, edge preservation, and error-distribution concentration, validating the effectiveness of the collaborative modules. The mixed result of G4 further highlights the advantage of guided filtering over texture-variance-based weighting, as it serves as an edge-aware local linear refinement operator: the former is adaptively driven by the structure of the guidance image and requires no predefined global threshold, whereas the latter is limited by fixed weighting strategies and cannot handle diverse texture distributions.

4.4 Generalization across Monocular Structural Priors

To examine whether the proposed fusion framework generalizes across different monocular structural cues, we replace the monocular structural prior while keeping stereo anchors, cross-modal scale alignment, edge-aware propagation, and the evaluation protocol fixed. Depth Anything V2 achieves the best overall accuracy. MiDaS [3] and ZoeDepth [23] can also be integrated into the proposed pipeline without retraining, but their higher global and edge-region errors indicate that the final accuracy still depends strongly on the quality of the monocular structural prior. These results support the modularity of the framework while revealing its dependence on the selected monocular model.

Table 4 Different monocular structural priors on the combined Middlebury 2014 and ETH3D benchmarks

Method	Global MAE	Global RMSE	Bad 2.0	Edge-MAE	Edge-RMSE	F1
Depth Anything V2	1.32	2.62	14.45%	2.04	2.72	0.72
MiDaS	2.13	3.90	23.53%	3.22	5.14	0.64
ZoeDepth	2.05	3.77	23.54%	3.19	5.15	0.63

4.5 Sensitivity to Sparse Geometric Anchors

To analyze the sensitivity of depth perception to sparse geometric anchors, we randomly retain 100%, 75%, 50%, 25%, and 10% of the valid anchors, and repeat each setting five times; Table 5 reports the numerical means over five random runs. The 100% setting corresponds to the full method and is consistent with the main comparison results in Tables 1 and 2: an Edge-RMSE of 2.72, an Edge-MAE of 2.04, a Global MAE of 1.32, and a Global RMSE of 2.62. When the ratio decreases to 75%, Edge-RMSE increases to 5.86 and Edge-MAE to 3.93; at 50%, the edge-region errors increase moderately (Edge-RMSE 5.97, Edge-MAE 4.54), and further subsampling to 25% and 10% increases Edge-RMSE to 6.39 and 6.74, respectively. In contrast, Global RMSE rises sharply from 2.62 at 100% to about 15.14 at 75% and then remains in a narrow range of 15.14 to 15.57. These results indicate that retaining the full anchor set is important for accurate scale estimation, while edge-region error degrades more gradually as anchors are further sparsified. Each sampling ratio retained at least 100 valid anchors in our experiments. Although F1 changes modestly under reduced anchor density, this metric only reflects the overlap of detected edge locations and does not measure disparity-error magnitude. The simultaneous increase in Edge-MAE and Edge-RMSE indicates that sparse anchors still degrade edge-region accuracy despite modest F1 changes.

Table 5 Sparse geometric anchor density analysis on the combined Middlebury 2014 and ETH3D benchmarks. Values are means over five random runs

Anchor density (%)	Edge-RMSE	Edge-MAE	Global MAE	Global RMSE	F1
100	2.72	2.04	1.32	2.62	0.72
75	5.86	3.93	10.54	15.14	0.64
50	5.97	4.54	11.03	15.21	0.64
25	6.39	5.22	11.57	15.41	0.66
10	6.74	5.63	11.88	15.57	0.68

4.6 Affine Alignment Ablation

To examine whether cross-modal scale consistency is necessary for fusing monocular structure with stereo geometry, we compare A1 (no alignment), A2 (linear scaling only), and A3 (full affine alignment) while keeping stereo anchors, edge-aware propagation, and the evaluation protocol fixed. As summarized in Table 6, full affine alignment is consistently superior under zero-shot evaluation on ETH3D: Edge-MAE decreases from 3.99 (A1) to 1.68 (A3), Edge-RMSE from 7.22 to 3.76, and F1 improves from 0.64 to 0.78. Linear scaling alone provides only limited improvement (Edge-RMSE 6.71 and Edge-MAE 3.58). The global errors show a consistent trend: A3 achieves a Global MAE of 5.96 and a Global RMSE of 8.88, substantially lower than A1 (7.91/10.39) and A2 (7.46/10.21). Therefore, the complete scale-and-offset affine alignment is not only necessary for converting the monocular prior into the disparity domain but also significantly affects edge-preservation performance, making it an indispensable component of the fusion pipeline.

Fig. 5 visualizes the fitted affine mapping between the monocular structural prior and calibrated stereo disparity anchors.

Fig. 6 illustrates a failure case caused by sparse and unreliable stereo anchors.

Table 6 Affine alignment ablation on ETH3D

Variant	Edge-RMSE	Edge-MAE	Global MAE	Global RMSE	F1
A1 (no alignment)	7.22	3.99	7.91	10.39	0.64
A2 (linear scaling)	6.71	3.58	7.46	10.21	0.66
A3 (full affine)	3.76	1.68	5.96	8.88	0.78

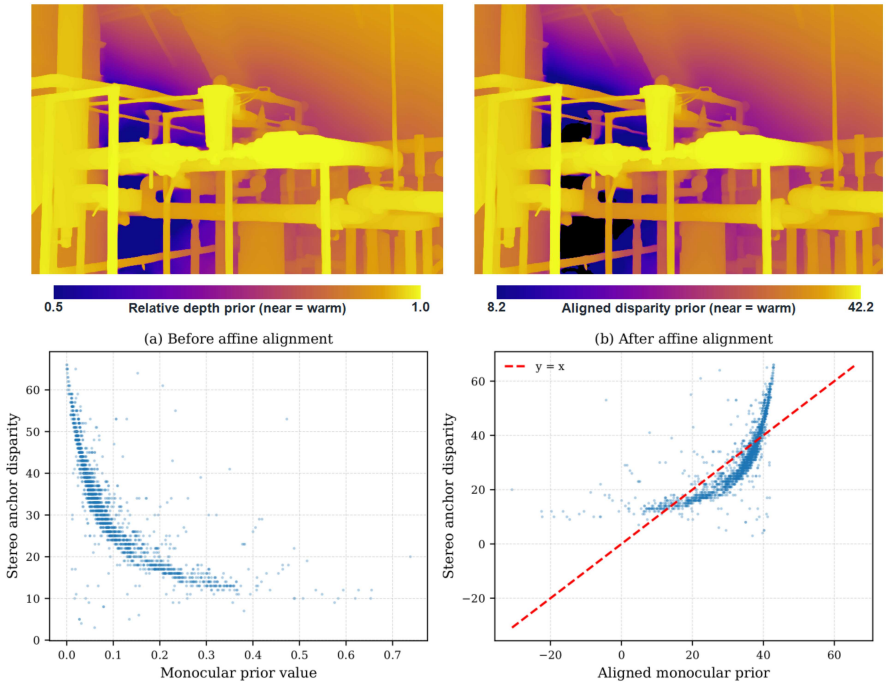

Fig. 5 Affine scale alignment between the monocular structural prior and calibrated stereo disparity anchors. (a) Relative monocular prior before alignment. (b) Aligned disparity prior after affine alignment. (c) Relationship between raw monocular prior values and stereo anchor disparities. (d) Relationship between aligned monocular prior values and stereo anchor disparities. The red dashed line in (d) represents the ideal $y = x$ consistency relationship. The scatter plots show the affine outputs before clipping to visualize the fitted relationship

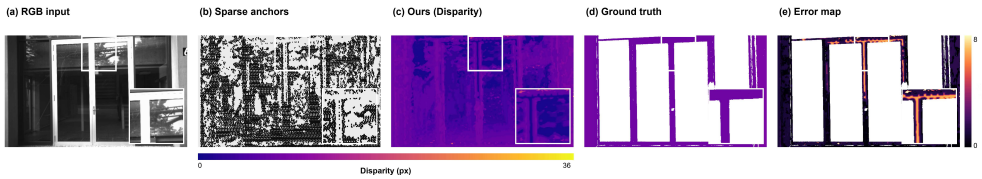

Fig. 6 Failure case caused by sparse and unreliable stereo anchors. From left to right: RGB input, sparse stereo anchors, predicted disparity, ground truth disparity, and disparity error map. White boxes highlight a representative failure region

4.7 Discussion

On the combined benchmarks, our method achieves competitive edge preservation and stable zero-shot generalization compared with non-iterative baselines. Guided filtering further improves edge quality, while RAFT-Stereo remains the strongest overall pre-trained cross-domain baseline. The monocular-prior study further shows that MiDaS and ZoeDepth can be integrated without retraining, while the final accuracy still depends on the quality of the monocular prior. The sparse-anchor analysis shows that the full anchor set is important for scale accuracy, while edge-region performance degrades gradually under further sparsification. The affine-alignment ablation further confirms that the scale-and-offset mapping is essential for transferring monocular priors into the calibrated disparity domain. In summary, the method provides a practical balance among generalization, accuracy, edge preservation, and deployment complexity without task-specific training.

First, we evaluate edge-preserving accuracy on the combined benchmarks. The proposed method achieves an Edge-RMSE of 2.72 and an Edge-MAE of 2.04, outperforming PSMNet (5.66/3.58) and GraftNet (6.25/2.66) in edge-region errors, while RAFT-Stereo remains the strongest overall pre-trained cross-domain baseline. Its GRR (15.69) is lower than those of PSMNet (17.67) and GraftNet (26.85), and its F1 (0.72) is higher than that of PSMNet (0.68) while remaining close to RAFT-Stereo (0.77). These results indicate that the proposed training-free fusion pipeline improves edge preservation over non-iterative baselines.

Second, we assess zero-shot cross-domain generalization. Under the evaluated zero-shot settings, the proposed training-free fusion pipeline remains more stable than PSMNet, which shows severe performance degradation on ETH3D, and achieves lower edge-region errors than GraftNet and PSMNet on the combined benchmarks. These results support the robustness of the proposed method without target-specific training or fine-tuning.

Third, we examine the edge-preservation advantage over non-iterative baselines. In the combined edge-region evaluation, the proposed method achieves lower Edge-RMSE and Edge-MAE than GraftNet and PSMNet, and its GRR is lower than those of both non-iterative baselines. The ablation study further shows that the full method G5 achieves the best Edge-RMSE and Edge-MAE among the five variants, reducing them by approximately 5.9% and 2.9% compared with G3, confirming that guided filtering helps maintain object-boundary depth discontinuities while performing edge-aware refinement.

5 Conclusion

We propose an explicit training-free fusion framework that combines relative monocular structural priors, sparse calibrated stereo disparity anchors, affine scale alignment, and edge-preserving guided filtering. The fusion stage requires no task-specific training, fine-tuning, or iterative optimization, and it can be integrated with different off-the-shelf monocular structural priors. Extensive experiments on indoor and outdoor

benchmarks demonstrate competitive edge preservation, stable zero-shot generalization, and sensitivity to sparse anchors and alignment configurations, providing a practical solution for robust depth perception.

The proposed framework offers a practical trade-off among robustness, accuracy, and deployment complexity. Because it requires no task-specific training or iterative optimization, it can be integrated with off-the-shelf monocular models and conventional stereo pipelines, which is valuable for applications such as autonomous systems, 3D reconstruction, and scene understanding.

Limitations include its dependence on the monocular prior and the limited expressiveness of a single global affine mapping, especially when the scene has a large depth range or the prior exhibits nonlinear scale drift. Additionally, the method is sensitive to very sparse or poorly distributed anchors, and the GRR still indicates some degree of overshoot. Spatially partitioned or sparse local affine regression may be explored in future work.

Declarations

Funding. This document is the result of the research project funded by the Jilin Province Major Scientific and Technological Project (Grant No. 20240301013ZD): Research on Multi-modal Fusion Optical-Mechanical Systems and Their Brain-like Perception Technologies.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Data availability. The current study did not generate any new datasets. All datasets analysed during the study are publicly available: the Middlebury 2014 stereo dataset at https://vision.middlebury.edu/stereo/data/scenes2014/ and the ETH3D benchmark at https://www.eth3d.net/datasets.

References

[1] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 4009–4018.

[2] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 3828–3838.

[3] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, 2022.

[4] V. Arampatzakis, G. Pavlidis, N. Mitianoudis, and N. Papamarkos, "Monocular depth estimation: A thorough review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2397–2416, Apr. 2024.

[5] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2650–2658.

[6] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth Pro: Sharp monocular metric depth in less than a second," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[7] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "UniDepth: Universal monocular metric depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 10106–10116.

[8] M. Viola, K. Qu, N. Metzger, et al., "Marigold-DC: Zero-shot monocular depth completion with guided diffusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 5359–5370.

[9] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[10] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5410–5418.

[11] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 185–194.

[12] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from LiDAR and monocular camera," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, Montreal, QC, Canada, May 2019, pp. 3288–3295.

[13] T. Guan, J. Guo, C. Wang, and Y.-H. Liu, "BridgeDepth: Bridging monocular and stereo reasoning with latent alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 27681–27691.

[14] J. Jiang, Q. Liu, H. Yu, H. Liu, L. Wang, J. Chen, and H. Ma, "MonoMVSNet: Monocular priors guided multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 27806–27816.

[15] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," *arXiv preprint arXiv:2406.09414*, 2024.

[16] M. Elad, "On the origin of the bilateral filter and ways to improve it," *IEEE Trans. Image Process.*, vol. 11, no. 10, pp. 1141–1151, Oct. 2002.

[17] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 1994, pp. 151–158.

[18] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, 2013.

[19] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition (GCPR)*, Münster, Germany, Sep. 2014, pp. 31–42.

[20] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2538–2547.

[21] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2021, pp. 218–227.

[22] B. Liu, H. Yu, and G. Qi, "GraftNet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 13012–13021.

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