

CogStereo: Neural Stereo Matching with Implicit Spatial Cognition Embedding

Lihuang Fang¹, Xiao Hu², Yuchen Zou³, Hong Zhang^{1,*}, *Life Fellow, IEEE*

Abstract—Deep stereo matching has advanced significantly on benchmark datasets through fine-tuning but falls short of the zero-shot generalization seen in foundation models in other vision tasks. We introduce CogStereo, a novel framework that addresses challenging regions, such as occlusions or weak textures, without relying on dataset-specific priors. CogStereo embeds implicit spatial cognition into the refinement process by using monocular depth features as priors, capturing holistic scene understanding beyond local correspondences. This approach ensures structurally coherent disparity estimation, even in areas where geometry alone is inadequate. CogStereo employs a dual-conditional refinement mechanism that combines pixel-wise uncertainty with cognition-guided features for consistent global correction of mismatches. Extensive experiments on Scene Flow, KITTI, Middlebury, ETH3D, EuRoc, and real-world demonstrate that CogStereo not only achieves state-of-the-art results but also excels in cross-domain generalization, shifting stereo vision towards a cognition-driven approach. More details are available at <https://github.com/lhfang228/CogStereo>.

I. INTRODUCTION

Stereo matching, essential for estimating depth from binocular images, remains a core challenge in robotics and computer vision [1]–[5]. Despite advancements with synthetic datasets and powerful neural architectures, robust performance in varied real-world environments is elusive. Challenges arise in *ill-posed regions* like occlusions and weak textures, where pixel correspondence is unreliable. Current state-of-the-art (SOTA) methods often rely on domain-specific tuning with techniques such as cost volumes [5], recurrent refinements [6], [7], and transformer-based reasoning [8], limiting their generalization to ill-posed regions.

In contrast, foundation models for other vision tasks have shown strong *zero-shot generalization* [9]. Models pretrained for classification [10], [11], segmentation [12], and monocular depth estimation [13] have demonstrated robust

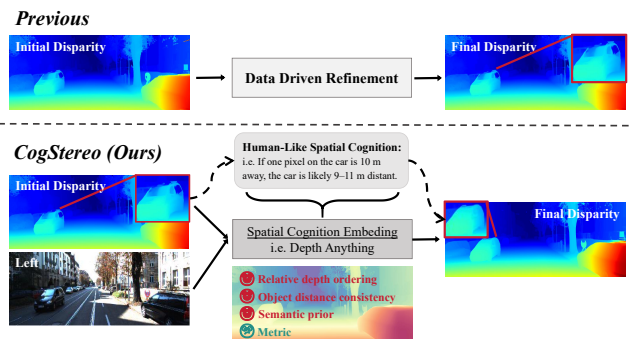


Fig. 1. Illustration of the motivation behind CogStereo. While modern stereo networks achieve highly accurate disparity estimation in most regions, they remain vulnerable to mismatches in textureless, reflective, or occluded areas. CogStereo addresses these challenges by embedding implicit SC, enabling globally consistent and robust disparity estimation.

performance on diverse data without domain adaptation. This raises the question: *Can stereo matching adopt a foundation-model approach for improved generalization to ill-posed regions?* This is crucial for applications like autonomous driving [14] and robotics [1], where consistent performance over diverse regions is vital.

A key shortcoming in stereo systems is their reliance on local geometric correspondence, often failing in difficult regions. In contrast, the depth foundation model captures object-level geometry and global scene understanding [13]—a capability we term *spatial cognition (SC)*, similar to human perception (as shown in Fig. 1). This understanding encompasses depth ordering, shape consistency, and semantic priors (e.g., flat roads, upright humans), allowing monocular models to maintain coherence. Previous attempts to merge monocular cues with stereo [15]–[17] have *explicit depth maps*, which are limited by local errors and lack global consistency.

To address this, we propose **CogStereo**, a framework transcending geometric correspondences by embedding implicit SC. Instead of explicit depth predictions, CogStereo utilizes feature representations from depth foundation model as SC priors. To refine disparities, it introduces a Dual-Condition Refinement mechanism, which adjusts disparity based on (i) pixel-wise uncertainty identifiers for unreliable regions and (ii) spatial cognition features for semantic and geometric consistency. This ensures globally coherent disparity maps, even in challenging conditions, as Fig. 1 shows how SC aids in achieving robust disparities. Overall, our contributions are summarized as follows:

- We propose **CogStereo**, embedding **implicit spatial**

*Corresponding author: Hong Zhang

¹L.Fang and H.Zhang are with Shenzhen Key Laboratory of Robotics and Computer Vision and the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, 518055, China. l.h.fang228@gmail.com; hzhang@sustech.edu.cn

²X.Hu is with the Lower Airspace Economy Research Institute, International Digital Economy Academy, ShenZhen, 5100085, China. huxiaol@idea.edu.cn

³Yuchen Zou with the School of Automation Science and Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi, 710049, China. yuchenzou@stu.xjtu.edu.cn

This work was supported in part by Shenzhen Science and Technology Program (No. SGDXX20240115111759002), in part by Meituan Academy of Robotics Shenzhen, in part by the Shenzhen Association for Science and Technology (No. XHXS2025-003), and in part by High level of special funds (G03034K003) from Southern University of Science and Technology, Shenzhen, China.

cognition into stereo matching, leveraging monocular depth features as priors for improved understanding and accuracy.

- We introduce a **Dual-Condition Refinement** mechanism to integrate uncertainty priors with implicit SC features, enhancing disparity correction in ambiguous regions and preventing metric drift caused by implicit optimization.
- Extensive benchmarks and real-world experiments demonstrate CogStereo’s SOTA results and robust zero-shot generalization, marking a shift toward cognitively informed stereo matching.

II. RELATED WORK

A. Learning based Stereo Matching

Early deep stereo methods [18]–[20] built 3D cost volumes from binocular features and used 3D convolutions for disparity regression. Later works introduced recurrent refinement [18], [21] and transformer-based methods [22] for long-range pixel correspondences, while NMRF-Stereo [23] used neural MRFs for improved accuracy. Despite these advances, all rely on geometric matching and still exhibit poor generalization in occlusions, weak textures, or repetitive patterns. Recursive stereo methods like RAFT-Stereo [18] and its variants improve structure and information integration [5], [24], but still struggle with weak texture or reflective regions. Even large-scale pre-training (FoundationStereo [25]) shows high zero-shot errors in complex scenes (Fig. 2), revealing a cognitive bottleneck in pure stereo paradigms. In contrast, CogStereo incorporates spatial cognition from depth foundation model, directly addressing the matching challenge.

B. Monocular Depth Estimation

Monocular depth estimation has advanced rapidly with deep learning [26]–[28]. Early CNN-based regression methods suffered from scale ambiguity and domain gaps, while MiDaS [29] reframed the task as relative scene understanding, enabling zero-shot inference. Depth Anything (DA) [30] and DAV2 [13] further improved robustness through semantic alignment and pseudo-label distillation, and diffusion-based approaches [31], [32] leverage generative priors to enforce structural consistency. With large-scale pretraining and foundation models such as DINOv2 [33], modern monocular depth models can capture relative depth ordering, object geometry, and global layout. Crucially, they preserve geometric and semantic consistency even in textureless or occluded regions, a property we call **SC**, reflecting holistic scene understanding and strong zero-shot generalization. This emerging property provides the key inspiration [34] for our CogStereo, which aims to embed SC into stereo matching for robust and globally consistent disparity estimation.

III. COGSTEREO

A. Preliminaries

Given a rectified image pair I_l and I_r , stereo matching estimates a dense horizontal displacement field f_h , mapping each pixel (u, v) in I_l to $(u + f_h, v)$ in I_r . A method like

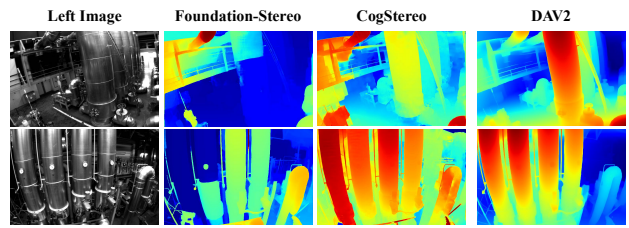


Fig. 2. Zero-shot prediction on EuRoC, demonstrating CogStereo’s generalization to practical scenarios with challenging properties like weak texture, appearance ambiguities, reflectance, translucency and occlusion.

RAFT-Stereo [18] constructs a cost volume by correlating features from the left and right images:

$$C(i, j, k) = \frac{F(I_l) \cdot F(I_r)}{\sqrt{D}}, \quad (1)$$

where $F(\cdot)$ denotes feature maps, D is the feature dimension, and $C(i, j, k)$ measures the similarity between pixel (i, j) in I_l and displacement k in I_r . Disparity is then obtained by iterative refinement over the cost volume. While effective in textured regions, these matching and refinement methods often fail in areas such as textureless surfaces, reflections, repetitive patterns, and occlusions where correspondences are ambiguous. The underlying limitation is the absence of SC: reasoning is restricted to local feature similarity without object- or scene-level awareness.

B. Framework Overview

As shown in Fig. 3, CogStereo is a novel neural stereo matching method that significantly enhances matching accuracy and robustness through implicit spatial cognition embedding. The encoder F from Eq. 1 first extracts features from I_l and I_r to construct a 3D cost volume, forming the conventional stereo matching backbone. On top of this baseline, Our framework utilizes the depth foundation model DAV2 [13] to extract features rich in spatial cognition, including relative depth ordering, object shape consistency, and physical priors. Furthermore, to address limitations that metric drift arises when directly applying the DAV2 feature to stereo matching, we introduce several complementary mechanisms. During the correlation optimization phase, Uncertainty Adapter is introduced to predict the log-variance of each pixel, which is jointly optimized with disparity regression, allowing uncertainty information to be directly integrated into feature learning and disparity optimization. In the dual-condition refinement stage via SC, CogStereo integrates pixel-level uncertainty priors with SC features from DAV2. An uncertainty-guided spatial cognitive attention mechanism identifies areas requiring correction while leveraging reliable SC information to achieve globally consistent refinement. Meanwhile, to prevent metric drift during implicit optimization, CogStereo employs a KNN-based scale-and-shift alignment strategy (*LU-KSS*) in low-uncertainty regions. Pixels with uncertainty below the θ -th percentile serve as anchor points, and weighted scale and shift are computed via KNN to align sparse reference disparities.

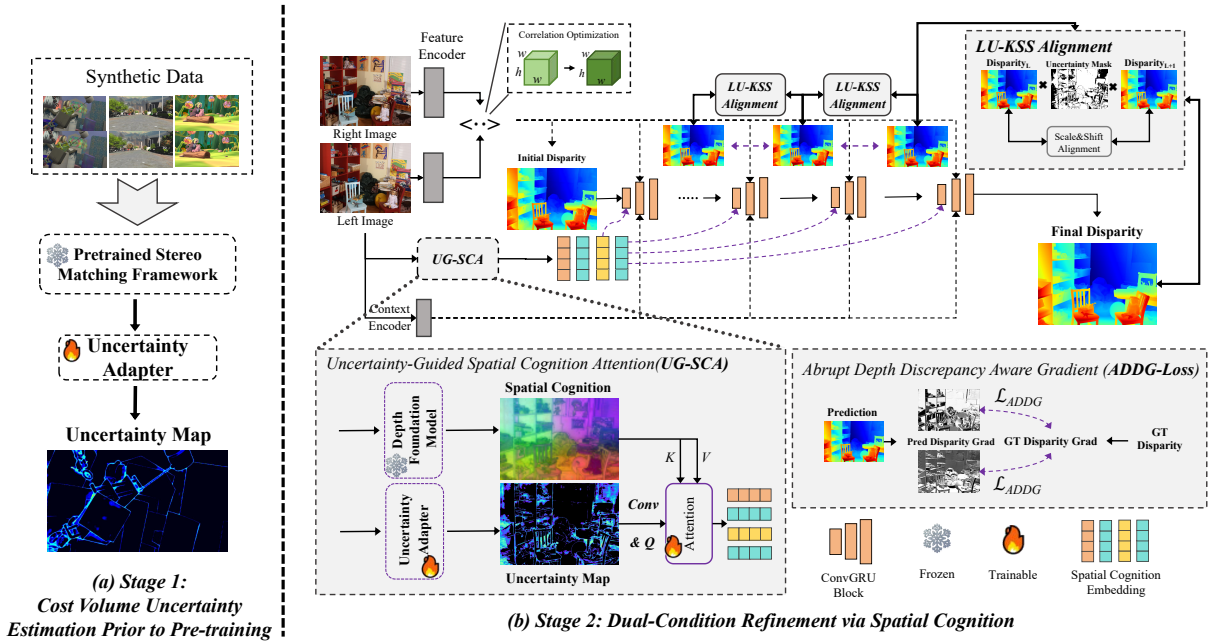


Fig. 3. Overview of Our CogStereo Framework. The CogStereo framework enhances stereo matching accuracy through two stages. Stage 1 pretrains the network to generate an uncertainty map from the cost volume. Stage 2 uses the *UG-SCA* module to refine disparities, leveraging SC to correct errors in high-uncertainty regions. The ConvGRU Block can be selected from an arbitrary neural stereo matching framework.

Low-uncertainty areas are directly anchored, while high-uncertainty regions use the average scale and shift alignment. In addition, the ADDG-Loss is introduced to penalize abrupt disparity variations in challenging areas.

C. Cost Volume Uncertainty Estimation Prior to Pre-training

While modern stereo networks often deliver accurate disparity predictions in most regions, they remain susceptible to errors in textureless surfaces, reflective areas, and occlusions. A key limitation is that existing methods typically estimate uncertainty only after disparity regression, which prevents fully leveraging cost volume information during feature learning and optimization.

To address this, we introduce *Cost Volume Uncertainty Estimation Prior to Pre-training* (as illustrated in Fig. 3 (a)), where uncertainty is explicitly modeled at the cost volume stage. Specifically, we design a lightweight *Uncertainty Adapter* atop the cost volume to jointly predict the log-variance of each pixel alongside disparity regression, thereby embedding confidence information into the optimization process. The adapter consists of two convolutional layers with a ReLU activation in between, outputting a log-variance map that characterizes pixel-wise uncertainty. This uncertainty map is incorporated into the training objective through a residual-adaptive loss formulation:

$$\mathcal{L}_{\text{uncertainty}} = \frac{1}{2} e^{-\log \sigma^2} (d_{\text{pred}} - d_{\text{gt}})^2 + \frac{1}{2} \log \sigma^2, \quad (2)$$

which enforces stronger supervision in confident regions while attenuating gradients in uncertain ones, effectively reducing overfitting and noise propagation.

By making uncertainty an intrinsic and interpretable signal rather than a post-processing byproduct, our approach pro-

duces meaningful uncertainty maps and enables more robust disparity optimization. As shown in Fig. 4, thresholding high-uncertainty pixels yields a significant reduction in EPE, highlighting the effectiveness of the proposed uncertainty prior. After learning uncertainty-aware representations, CogStereo leverages them jointly with spatial cognition priors for disparity refinement.

D. Dual-Condition Refinement via Spatial Cognition

Initial disparity estimates, even from the SOTA stereo or multi-view networks, often contain localized errors due to occlusions, textureless regions, or matching ambiguities. Humans, when assessing depth, instinctively rely on two complementary cues: (i) *error awareness*, signaling unreliable regions, and (ii) *object-level priors*, ensuring consistent depth across the same object. Inspired by this and condition control [35], we propose a **dual-condition refinement mechanism** (as illustrated in Fig. 3 (b)), where disparity correction is guided by (i) a *pixel-wise uncertainty prior*, indicating *where* corrections are needed, and (ii) *DAv2 features*, offering object-level semantic and geometric cues for *how* to correct errors. This design allows the network to focus on high-uncertainty regions while propagating reliable depth information across object surfaces, yielding disparity maps that are both locally accurate and globally coherent.

a) *Uncertainty as a Reliability Condition*: A pre-estimated disparity uncertainty map $\mathbf{U} \in \mathbb{R}^{H \times W}$ is incorporated to indicate *where corrections are required*. High-uncertainty pixels represent potential errors, while low-uncertainty pixels serve as reliable anchors. By directing corrections from reliable to uncertain regions, this design prevents over-optimizing already accurate areas and ensures appropriate depth adjustment in ambiguous regions.

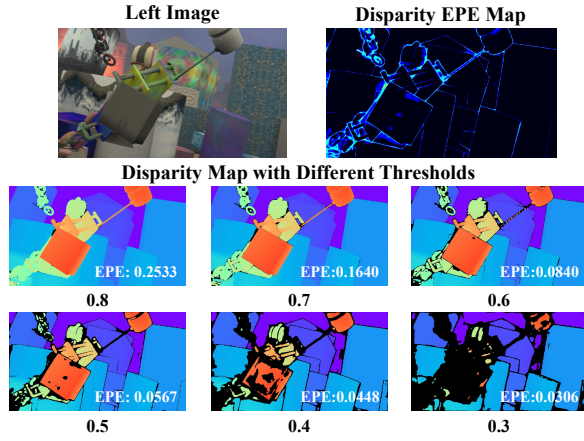


Fig. 4. Effectiveness of Uncertainty Masking on Disparity Estimation Accuracy. The top row shows the left image and disparity EPE map, highlighting high-error regions. Subsequent rows show reduced EPE with high uncertainty regions masked, where lower values mean higher accuracy.

b) *Depth Anything Features as a Spatial Cognition Condition*: We leverage DAv2 [13], a monocular depth model pre-trained on large-scale imagery, to supply object-level semantic and geometric priors. Instead of using raw depth predictions, we employ the intermediate feature representation $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$, which encodes structural and semantic consistency. These features serve as implicit spatial cognition cues, guiding *how corrections should be propagated* from reliable to uncertain regions, enabling globally coherent refinement even in challenging scenarios.

c) *Uncertainty Guided Spatial Cognition Attention (UG-SCA)*: We propose a UG-SCA module that integrates uncertainty information into both spatial cognition and disparity refinement. Specifically, given uncertainty \mathbf{U} and DAv2 features \mathbf{F} , uncertainty-conditioned queries interact with features to produce corrective signals \mathbf{F}_{sc} :

$$\mathbf{F}_{sc} = \text{Softmax}\left(\frac{\phi_q(\text{Conv}(\mathbf{U})) \cdot \phi_k(\mathbf{F})^\top}{\sqrt{d}}\right)\phi_v(\mathbf{F}), \quad (3)$$

where ϕ_q, ϕ_k, ϕ_v are learnable projections and d is the feature dimension.

d) *Low-Uncertainty Area KNN-Based Scale-and-Shift Alignment (LU-KSS)*: To prevent metric drift caused by DAv2 implicit optimization, we propose a KNN-based scale-and-shift alignment strategy guided by low-uncertainty area (LU-KSS), inspired by [36]. First, pixels with uncertainty below the θ -th percentile are selected as reliable anchors:

$$P_{\text{reliable}} = \{(x, y) \mid U(x, y) \leq \tau_\theta\}, \quad \tau_\theta = Q_\theta(U), \quad (4)$$

where $Q_\theta(U)$ denotes the θ -th percentile of the uncertainty distribution. Next, for each low uncertainty pixel, its K nearest neighbors in the low uncertainty regions are identified, and local alignment parameters (s, t) are estimated via inverse-distance weighted least squares:

$$(s, t) = \arg \min_{s, t} \sum_{i=1}^K w_i \|s d_{p, k+1}(x, y) + t - d_{p, k}(x, y)\|^2, \quad (5)$$

where $d_{p, k}(x, y)$ denotes k -th predicted disparity and w_i denotes the normalized inverse-distance weight of the i -th neighbor. The aligned disparity $d_{p, k+1}^a$ for low-uncertainty regions is then given by:

$$d_{p, k+1}^a(x, y) = s \cdot d_{p, k+1}(x, y) + t, \quad (x, y) \in P_{\text{reliable}}, \quad (6)$$

For high-uncertainty regions, they do not directly participate in the KNN fitting. Instead, the alignment parameters (s^*, t^*) for these regions are obtained by averaging the local alignment results from multiple low-uncertainty anchors:

$$(s^*, t^*) = \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} (s_j, t_j), \quad \mathcal{A} \subseteq P_{\text{reliable}} \quad (7)$$

and the aligned disparity for high-uncertainty regions is computed as:

$$d_{p, k+1}^a(x, y) = s^* d_{p, k+1}(x, y) + t^*, \quad (x, y) \in P_{\text{unreliable}}, \quad (8)$$

e) *Abrupt Depth Discrepancy Aware Gradient Loss*:

To improve spatial consistency and reduce abrupt depth changes [37] in disparity estimation, we introduce the *Abrupt Depth Discrepancy Aware Gradient Loss* (ADDG). This loss is specifically designed to penalize sudden changes in disparity, enhancing the smoothness and accuracy of disparity maps. It is particularly effective in regions with low texture, high reflectivity, or occlusions, where traditional stereo methods often struggle. The ADDG loss is defined as:

$$\mathcal{L}_{ADDG} = \left(\left| \frac{\partial(d_{p, k}^a - d_{gt})}{\partial x} \right| + \left| \frac{\partial(d_{p, k}^a - d_{gt})}{\partial y} \right| \right), \quad (9)$$

where $d_{p, k}^a$ and d_{gt} denote the aligned predicted and ground-truth disparity, respectively, and $\frac{\partial(d_{p, k}^a - d_{gt})}{\partial x}$, $\frac{\partial(d_{p, k}^a - d_{gt})}{\partial y}$ are the gradients of their difference along the x and y directions.

E. *Learning Objectives*

The learning objectives for the CogStereo is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{init} + \sum_{k=1}^N \gamma^{N-k} \|d_{p, k}^a - d_{gt}\|_1 + \mathcal{L}_{ADDG} + \mathcal{L}_{uncertainty} \quad (10)$$

where $\mathcal{L}_{init} = L_{Smooth}(d_{p, 0} - d_{gt})$ denotes the smooth \mathcal{L}_1 loss of the initial disparity $d_{p, 0}$, $\|d_{p, k}^a - d_{gt}\|_1$ is the \mathcal{L}_1 loss between the aligned disparity $d_{p, k}^a$ after the k -th update and the ground truth disparity d_{gt} , γ is a decay coefficient, typically set to 0.9, N is the number of iterative updates.

IV. EXPERIMENT

A. *Implementation Details*

CogStereo is implemented in PyTorch and trained on NVIDIA A100 GPUs. Following the training protocol of [38], we optimize with AdamW [39] under a one-cycle learning rate policy. Spatial cognition embeddings are extracted from a frozen DAv2 ViT-L backbone, which retains its zero-shot generalization on real-world imagery. The stereo stream uses the BasicEncoder of IGEV [5] for feature extraction. We pretrain on Scene Flow [40] for 200K steps (batch size 8), with a cosine one-cycle schedule peaking at $2e-4$.

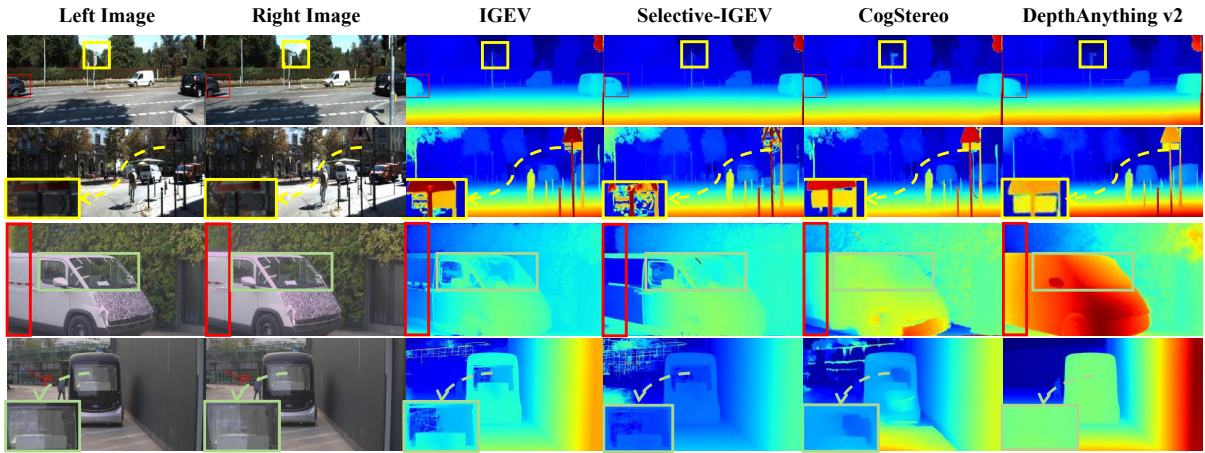


Fig. 5. Qualitative comparison of zero-shot inference on in-the-wild images, including samples from the KITTI training dataset and our self-collected mid-range stereo matching dataset for autonomous driving. The **red boxes** highlight **occlusion regions**, the **green boxes** indicate **reflection regions**, and the **yellow** mark **weak textures**. CogStereo outperforms the baselines, demonstrating its effectiveness in handling such complexities.

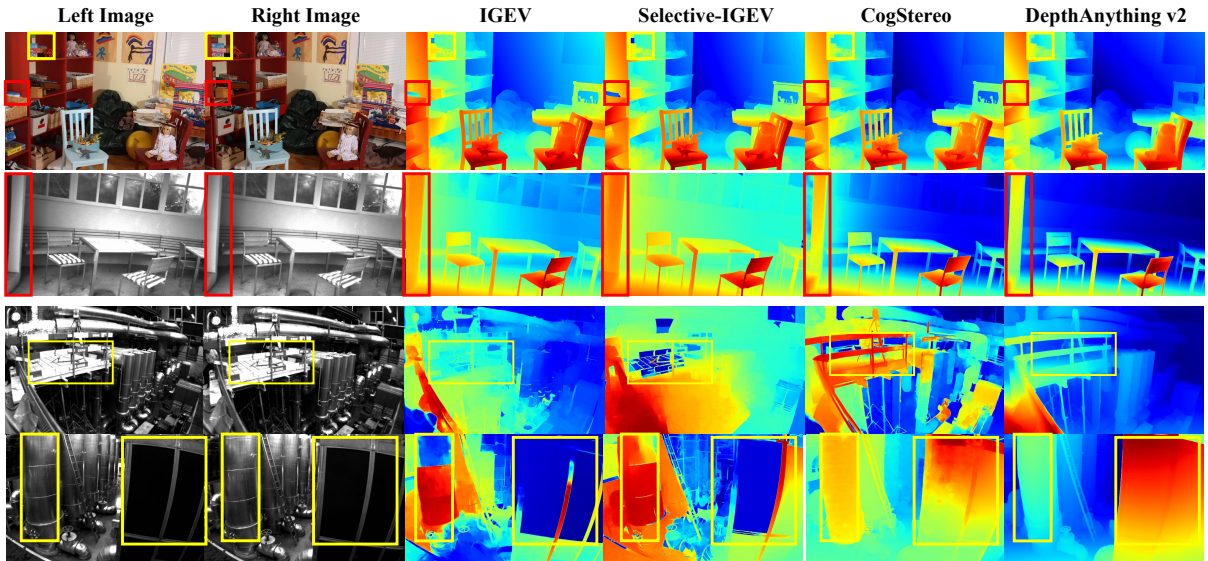


Fig. 6. Comparative zero-shot performance of stereo vision algorithms on the Middlebury (1st row), ETH3D (2nd row), EuRoC (last two rows), specifically tailored for Stereo SLAM applications.

B. Experiment Setting

a) Datasets: Consistent with most existing methods (e.g., [38]), we evaluate on five standard benchmarks spanning synthetic and real-world scenarios. Scene Flow [40] provides synthetic training and testing stereo pairs with dense, accurate disparity maps. For real-world evaluation, KITTI 2012 / 2015 [41], [42] offer outdoor urban scenes with sparse disparity annotations. To assess generalization, we additionally include zero-shot evaluation on Middlebury 2014 [43], featuring high-precision structured-light disparity for indoor scenes, ETH3D [44], containing mixed indoor/outdoor grayscale stereo pairs, EuRoC [45], and our self-collected mid-range dataset for autonomous driving.

b) Competing Methods: We compare CogStereo with 11 state-of-the-art **data-driven** stereo matching methods that span a broad architectural spectrum. These include classical CNN-based pipelines such as ACVNet [19], Mask-

CFNet [46] RAFT-Stereo [18], and PCW-Net [47]; CNN-Transformer hybrid designs like CREStereo++ [6], IGEV [5], IGEV++ [48], Selective-IGE V [24], and NMRF-Stereo [23]; as well as the latest Depth-Anything-v2-powered baselines, FoundationStereo [25] and DEFOMStereo-L [15]. All methods were trained exclusively on Scene Flow, ensuring a strictly zero-shot and fair cross-dataset evaluation.

c) Evaluation Metric: Our central hypothesis is that embedding DA2-derived spatial cognition empowers stereo networks to tackle ill-posed regions (e.g., occlusions) where pure geometric correspondence fails. We verify this with three standard metrics: average end-point error (EPE), Bad Pixel rate (BP- X , fraction of pixels with disparity error $> X$), and D1 (errors > 3 pixels or 5% of ground truth). As large errors concentrate in ill-posed regions like occlusions (OCC), improved BP- X and D1 directly confirm CogStereo’s efficacy in addressing these challenges.

TABLE I

ZERO-SHOT GENERALIZATION ON FOUR PUBLIC BENCHMARKS. FOR EACH DATASET, THE STANDARD EVALUATION METRICS ARE REPORTED. ALL METHODS ARE TRAINED EXCLUSIVELY ON SCENE FLOW FOR A FAIR COMPARISON. * INDICATES METHODS UTILIZING DAV2. NOC, OCC, AND ALL DENOTE NON-OCCLUDED, OCCLUDED, AND ALL PIXELS, RESPECTIVELY. GREEN DENOTES THE BEST RESULT AND YELLOW INDICATES THE SECOND BEST RESULT.

Methods	Middlebury BP-2 ↓			ETH3D BP-1↓			KITTI-12 D1↓			KITTI-15 D1↓		
	NOC	OCC	ALL	NOC	OCC	ALL	NOC	OCC	ALL	NOC	OCC	ALL
ACVNet [19]	22.1	47.4	25.7	8.7	19.6	9.2	12.9	54.5	13.9	11.3	32.9	11.7
Mask-CFNet [46]	-	-	13.7	-	-	5.7	-	-	4.8	-	-	5.8
RAFT-Stereo [18]	9.1	28.0	12.0	2.9	6.0	3.0	4.3	28.4	4.7	5.3	12.7	5.5
PCW-Net [47]	12.2	38.0	15.9	5.3	11.7	5.5	4.1	30.2	4.7	5.5	15.0	5.7
CREStereo++ [6]	-	-	14.8	-	-	4.4	-	-	4.7	-	-	5.2
IGEV [5]	7.3	24.3	9.9	4.1	9.8	4.4	4.9	33.7	5.6	5.6	14.3	5.8
IGEV++ [48]	-	-	7.8	-	-	4.1	-	-	5.1	-	-	5.9
NMRF-Stereo [23]	-	-	7.5	-	-	3.8	-	-	4.2	-	-	5.1
Selective-IGEV [24]	6.7	22.6	9.2	4.1	9.8	4.4	5.1	31.9	5.7	5.7	13.8	5.9
FoundationStereo* [25]	-	-	5.5	-	-	1.8	-	-	3.2	-	-	4.9
DEFOMStereo-L* [15]	4.4	20.6	6.9	2.1	5.1	2.2	3.8	22.0	4.2	4.8	12.6	5.0
CogStereo (Ours)	4.2	10.0	5.1	1.5	4.0	1.6	2.7	16.3	3.0	4.2	9.1	4.3

TABLE II

QUANTITATIVE EVALUATION ON SCENE FLOW TEST SET. **BOLD** DENOTES THE BEST RESULT AND UNDERLINE INDICATES THE RESULT OF THE BASELINE. **RED** DENOTES THE IMPROVEMENT COMPARISON WITH THE BASELINE.

Method	GwcNet [49]	LEAStereo [50]	RAFT-Stereo [18]	IGEV [5]	Selective-IGEV [24]	DEFOMStereo-L [15]	CogStereo
EPE (px)	0.76	0.78	0.67	<u>0.47</u>	0.44	0.42	0.35 -25.53%

C. Zero Shot Qualitative Comparison

As shown in Figs. 5 and 6, we evaluate the zero-shot generalization on real-world scenes with occlusions, reflections, weak textures and fine-grained structures. Despite lacking a metric scale, DAV2 generates structurally and semantically consistent depth maps in these ill-posed regions via pre-trained SC priors, yielding accurate depths for reflective areas, consistent values in repetitive patterns and continuous depths for slender structures. In contrast, the geometry-only IGEV model suffers from noise, discontinuities and detail loss. By embedding DAV2’s structural prior into the IGEV backbone (Section III), CogStereo inherits DAV2’s smooth, coherent structural features while retaining stereo metric accuracy, enabling robust zero-shot disparity estimation.

D. Quantitative Analysis

a) *Zero Shot Generalization Comparison:* As shown in Table I, CogStereo demonstrates strong zero-shot generalization across four public benchmarks. Beyond reducing overall errors (ALL), CogStereo consistently improves performance in both NOC and OCC regions. Notably, in the more challenging OCC regions, it achieves 10.0 on Middlebury, 4.0 on ETH3D, 16.3 on KITTI-12, and 9.1 on KITTI-15, surpassing prior SOTA methods by a clear margin. The larger margin in OCC regions confirms that spatial cognition priors are particularly beneficial in ill-posed settings such as occlusions, weak textures, and reflective surfaces, enabling CogStereo to recover disparities that are often lost by geometry-only methods. These results highlight that CogStereo’s robust improvements in challenging regions are not coincidental, but stem from the integration of spatial cognition, establishing it as a SOTA framework for zero-shot stereo matching.

TABLE III

ABLATION STUDY OF COGSTEREO MODULE. *w/o* DENOTES WITHOUT.

Module	Scene Flow (test) EPE ↓
<i>Baseline</i>	0.47
<i>w / o UG-SCA</i>	0.44
<i>w / o DAV2</i>	0.46
<i>w / o LU-KSS</i>	0.49
<i>w / o \mathcal{L}_{ADDG}</i>	0.36
All	0.35

b) *In-Domain Comparison:* Table II shows a quantitative comparison on Scene Flow. CogStereo surpasses other methods, reducing the best previous EPE from 0.47 to 0.35. Although in-domain training is not the main focus, these results highlight the effectiveness of our model design.

E. Ablation Study

a) *Ablation Study of CogStereo Module:* We conducted an ablation study on the Scene Flow test set to assess each component’s effectiveness, summarized in Table V. The baseline achieves an EPE of 0.47. Removing UG-SCA results in an EPE of 0.44; although slightly improved, the absence of uncertainty guidance causes over-optimization in low-uncertainty regions and under-correction in high-uncertainty areas. Removing LU-KSS worsens the EPE to 0.49, worse than the baseline, due to metric drift without scale-shift alignment. Excluding \mathcal{L}_{ADDG} worsens the EPE to 0.36, showing that penalizing abrupt depth discrepancies enhances robustness in challenging areas. Incorporating all components yields the best EPE of 0.35, confirming that each module contributes to performance improvement.

b) *Comparison of Different Monocular Depth Estimation Models:* As shown in Table IV, we compare CogStereo

with different monocular depth models. DINOv2 and DINOv3 (ViT-L) yield a fixed EPE of 0.46, showing their limited spatial reasoning despite strong classification and segmentation performance. UniDepth achieves a lower EPE of 0.38 with superior metric depth estimation but still falls short of the best result. In contrast, DAV2 outperforms others across all encoder scales, with the ViT-L variant attaining the optimal 0.35 EPE. These results confirm that depth-specialized monocular models trained on large-scale multi-scene data with semantic alignment provide stronger spatial cognition priors for CogStereo. Moreover, the performance improvement from ViT-S to ViT-L underscores the critical role of model capacity in boosting spatial reasoning.

TABLE IV
COMPARISON OF OUR COGSTEREO EQUIPPED WITH DIFFERENT MONOCULAR DEPTH ESTIMATION MODELS

Model	Encoder	Scene Flow (test) EPE ↓
DINOv2 [33]	ViT-L	0.46
DINOv3 [51]	ViT-L	0.46
UniDepth [52]	ViT-L	0.38
DAv2 [13]	ViT-S	0.42
	ViT-B	0.37
	ViT-L	0.35

c) *Extension to Other Stereo Matching Framework:*

To validate the generality of the proposed Spatial Cognition (SC), we integrate it into three representative stereo frameworks: RAFT-Stereo, Selective-Stereo, and IGEV. As shown in Table V, adding SC consistently improves performance: RAFT-Stereo’s EPE drops from 0.63 to 0.44, Selective-Stereo from 0.44 to 0.35, and IGEV from 0.47 to 0.35. These results demonstrate that SC is a versatile prior, effective across different stereo architectures and confirming its broad applicability and robustness.

TABLE V
ABLATION STUDY OF THE UNIVERSALITY OF PROPOSED SPATIAL COGNITION (SC).

Model	Module	Scene Flow (test) EPE ↓
RAFT-Stereo	w / o SC	0.63
	w SC	0.44
Selective-Stereo	w / o SC	0.44
	w SC	0.35
IGEV	w / o SC	0.47
	w SC	0.35

F. *Downstream Benchmark Performance*

Instead of evaluating on additional benchmarks, we complement the evaluation with downstream task scenarios. As shown in the last two rows of Fig. 6, in zero-shot evaluations on the EuRoC dataset for Stereo SLAM, CogStereo outperforms IGEV and Selective-IGEV by generating more coherent and accurate depth maps, especially in challenging regions with weak texture, occlusions, or reflective surfaces. Notably, DAV2 produces structurally complete and semantically plausible depth maps, which underscores the feasibility

of CogStereo’s assumption. By combining stereo matching precision with monocular depth estimation consistency, CogStereo enhances depth quality without fine-tuning. Single-frame inference takes ~ 0.3 s on a single RTX 3090 GPU with ~ 19 GB memory, confirming its practical utility for autonomous driving and robotic navigation.

V. CONCLUSION

This paper introduces CogStereo to neural stereo matching that addresses ill-posed regions such as occlusions, weak textures, and appearance ambiguities by embedding implicit spatial cognition. CogStereo leverages DAV2 feature as spatial cognition, introducing an understanding of scene layout akin to human perception, thereby enhancing the global consistency and accuracy of matching. Comprehensive experiments across numerous standard benchmarks have shown that CogStereo has achieved top-tier performance and demonstrated remarkable generalization to ill-posed regions across diverse datasets. The success of CogStereo illustrates the potential of integrating geometric reasoning with spatial cognition to elevate stereo matching technology beyond basic geometric reasoning to a more sophisticated level of cognitive understanding.

REFERENCES

- [1] X. Liu, S. Wen, and H. Zhang, “A real-time stereo visual-inertial slam system based on point-and-line features,” *IEEE Transactions on Vehicular Technology*, vol. 72, no. 5, pp. 5747–5758, 2023. 1
- [2] F. Tang, H. Li, and Y. Wu, “Fmd stereo slam: Fusing mvng and direct formulation towards accurate and fast stereo slam,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 133–139. 1
- [3] X. Zuo, P. Geneva, Y. Yang, W. Ye, Y. Liu, and G. Huang, “Visual-inertial localization with prior lidar map constraints,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3394–3401, 2019. 1
- [4] Y. Qiu, Y. Chen, Z. Zhang, W. Wang, and S. Scherer, “Mac-vo: Metrics-aware covariance for learning-based stereo visual odometry,” *arXiv preprint arXiv:2409.09479*, 2024. 1
- [5] G. Xu, X. Wang, X. Ding, and X. Yang, “Iterative geometry encoding volume for stereo matching,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 919–21 928. 1, 2, 4, 5, 6
- [6] J. Jing, J. Li, P. Xiong, J. Liu, S. Liu, Y. Guo, X. Deng, M. Xu, L. Jiang, and L. Sigal, “Uncertainty guided adaptive warping for robust and efficient stereo matching,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3318–3327. 1, 5, 6
- [7] Y. Gao and L. Shen, “Iterative volume fusion for asymmetric stereo matching,” *arXiv preprint arXiv:2508.09543*, 2025. 1
- [8] Z. Liu, Y. Li, and M. Okutomi, “Global occlusion-aware transformer for robust stereo matching,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3535–3544. 1
- [9] J. Hong, R. Choi, and J. J. Leonard, “Learning from feedback: Semantic enhancement for object slam using foundation models,” *arXiv preprint arXiv:2411.06752*, 2024. 1
- [10] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” in *European conference on computer vision*. Springer, 2024, pp. 38–55. 1
- [11] Y. Yang, Y. Hu, M. Ye, Z. Zhang, Z. Lu, Y. Xu, U. Topcu, and B. Snyder, “Uncertainty-guided enhancement on driving perception system via foundation models,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8752–8758. 1
- [12] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, “Grounded sam: Assembling open-world models for diverse visual tasks,” *arXiv preprint arXiv:2401.14159*, 2024. 1

- [13] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2024. 1, 2, 4, 7
- [14] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "Driving-stereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1
- [15] H. Jiang, Z. Lou, L. Ding, R. Xu, M. Tan, W. Jiang, and R. Huang, "Defom-stereo: Depth foundation model based stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 857–21 867. 1, 5, 6
- [16] J. Cheng, L. Liu, G. Xu, X. Wang, Z. Zhang, Y. Deng, J. Zang, Y. Chen, Z. Cai, and X. Yang, "Monster: Marry monodepth to stereo unleashes power," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 6273–6282. 1
- [17] L. Bartolomei, F. Tosi, M. Poggi, and S. Mattoccia, "Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 1013–1027. 1
- [18] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227. 2, 5, 6
- [19] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 981–12 990. 2, 5, 6
- [20] F. Tosi, L. Bartolomei, and M. Poggi, "A survey on deep stereo matching in the twenties," *International Journal of Computer Vision*, 2025. 2
- [21] X. Guo, C. Zhang, Y. Zhang, W. Zheng, D. Nie, M. Poggi, and L. Chen, "Lightstereo: Channel boost is all you need for efficient 2d cost aggregation," *arXiv preprint arXiv:2406.19833*, 2024. 2
- [22] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6197–6206. 2
- [23] T. Guan, C. Wang, and Y.-H. Liu, "Neural markov random field for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5459–5469. 2, 5, 6
- [24] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 701–19 710. 2, 5, 6
- [25] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5249–5260. 2, 5, 6
- [26] 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," *arXiv preprint arXiv:2410.02073*, 2024. 2
- [27] R. Wei, B. Li, F. Zhong, H. Mo, Q. Dou, Y.-H. Liu, and D. Sun, "Absolute monocular depth estimation on robotic visual and kinematics data via self-supervised learning," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 4269–4282, 2025. 2
- [28] W. Yin, C. Zhang, H. Chen, Z. Cai, G. Yu, K. Wang, X. Chen, and C. Shen, "Metric3d: Towards zero-shot metric 3d prediction from a single image," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 9043–9053. 2
- [29] R. Birkel, D. Wofk, and M. Müller, "Midas v3. 1—a model zoo for robust monocular relative depth estimation," *arXiv preprint arXiv:2307.14460*, 2023. 2
- [30] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024. 2
- [31] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 9492–9502. 2
- [32] B. Ke, K. Qu, T. Wang, N. Metzger, S. Huang, B. Li, A. Obukhov, and K. Schindler, "Marigold: Affordable adaptation of diffusion-based image generators for image analysis," 2025. 2
- [33] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023. 2, 7
- [34] H. Lin, S. Peng, J. Chen, S. Peng, J. Sun, M. Liu, H. Bao, J. Feng, X. Zhou, and B. Kang, "Prompting depth anything for 4k resolution accurate metric depth estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 070–17 080. 2
- [35] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847. 3
- [36] Z. Wang, S. Chen, L. Yang, J. Wang, Z. Zhang, H. Zhao, and Z. Zhao, "Depth anything with any prior," *arXiv preprint arXiv:2505.10565*, 2025. 4
- [37] A. Bochkovskiy, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," in *The Thirteenth International Conference on Learning Representations*. 4
- [38] T. Guan, J. Guo, C. Wang, and Y.-H. Liu, "Bridgedepth: Bridging monocular and stereo reasoning with latent alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 27 681–27 691. 4, 5
- [39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017. 4
- [40] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048. 4, 5
- [41] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361. 5
- [42] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013. 5
- [43] 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 *German conference on pattern recognition*. Springer, 2014, pp. 31–42. 5
- [44] T. Schops, J. L. Schonberger, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3260–3269. 5
- [45] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets," *The International Journal of Robotics Research*, 2016. 5
- [46] Z. Rao, B. Xiong, M. He, Y. Dai, R. He, Z. Shen, and X. Li, "Masked representation learning for domain generalized stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5435–5444. 5, 6
- [47] Z. Shen, Y. Dai, X. Song, Z. Rao, D. Zhou, and L. Zhang, "Pcw-net: Pyramid combination and warping cost volume for stereo matching," in *European conference on computer vision*. Springer, 2022, pp. 280–297. 5, 6
- [48] G. Xu, X. Wang, Z. Zhang, J. Cheng, C. Liao, and X. Yang, "Igev++: Iterative multi-range geometry encoding volumes for stereo matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 5, 6
- [49] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3273–3282. 6
- [50] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," *Advances in neural information processing systems*, vol. 33, pp. 22 158–22 169, 2020. 6
- [51] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025. 7
- [52] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "Unidepth: Universal monocular metric depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 106–10 116. 7