BINOCULAR FOCUSED VISION: QUERY-CONDITIONED STEREO GEOMETRY AND SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern stereo matching achieves high-quality dense metric depth, yet its prevailing formulation remains task-agnostic: all pixels are treated as equally important. In many query-driven applications, however, the location of interest is known in advance, motivating geometry estimation that is explicitly conditioned on where accuracy is needed. We introduce Binocular Focused Vision and propose FocusStereo, a unified framework for query-conditioned stereo geometry and spatial reasoning. Given rectified stereo images, camera calibration, and a spatial query, FocusStereo conditions stereo correspondence formation on the query, reconstructs local matching distributions, decodes bounded metric geometry, selectively integrates reliable updates, and derives query-centered 3D relations. Under a unified query-centric protocol on VKITTI2 and HAMMER, FocusStereo reduces Focus AbsRel by 65.17% and 27.04% and Query-XYZ error by 61.13% and 15.37%, respectively, relative to the strongest competing result for each metric. Controlled ablations with fixed stereo initialization further isolate the contribution of query-conditioned reasoning while preserving global metric stability. These results show that spatial queries can effectively condition stereo correspondence, extending stereo perception from task-agnostic dense reconstruction to query-conditioned geometry and spatial reasoning.

1 INTRODUCTION

Stereo matching provides a fundamental means of recovering metric geometry from calibrated binocular images and supports a wide range of applications in 3D reconstruction, robotic perception, navigation, and interactive vision. Modern stereo networks have substantially improved dense disparity estimation through learned feature representations, cost-volume aggregation, recurrent refinement, large-scale synthetic training, and cross-domain learning Scharstein et al. (2001); Chang & Chen (2018); Lipson et al. (2021); Xu et al. (2023a); Wen et al. (2025). Despite these advances, the prevailing formulation remains largely unchanged: stereo systems aim to recover a task-agnostic dense disparity field over the entire image.

Many downstream tasks, however, are inherently spatially selective. In interactive measurement, manipulation, local navigation, or targeted 3D inspection, a user, planner, or upstream perception module often already specifies where geometric accuracy is needed. This raises a different question from conventional dense stereo estimation: can stereo correspondence reasoning itself be conditioned on an explicit spatial query? We refer to this problem as Binocular Focused Vision. Given rectified stereo images, camera calibration, and a spatial query q, the goal is to improve query-relevant metric geometry while preserving globally consistent depth. Importantly, *focused* does not mean cropping an ROI from an existing depth map or merely reweighting queried pixels. Instead, the query participates before correspondence is established. Unlike active perception, which changes viewpoints or sensing parameters to acquire new observations Bajcsy (1988), Binocular Focused Vision operates on a fixed binocular observation and changes how that observation is geometrically interpreted.

This formulation introduces two coupled requirements. First, query-conditioned reasoning should exploit additional local correspondence evidence without degrading already reliable global geom-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

etry, particularly around occlusions, depth discontinuities, and weakly textured regions. Second, downstream reasoning often requires more than the depth of a single queried pixel; local geometry should also be expressed relative to the query itself. These requirements motivate a unified treatment of correspondence, metric geometry, reliability, and query-centered spatial reasoning.

We therefore propose FocusStereo, a unified framework that introduces the spatial query directly into stereo correspondence formation. Starting from an initial metric state used only for scale and search initialization, FocusStereo reconstructs query-conditioned matching evidence, decodes bounded geometric proposals, selectively incorporates reliable updates, and derives query-centered 3D relations from the resulting depth. A key design principle is to separate update reliability from update magnitude, allowing unsupported proposals to preserve the initialized geometry rather than forcing local refinement.

FocusStereo is complementary to recent advances in task-agnostic and zero-shot stereo Xu et al. (2023a); Jiang et al. (2025); Bartolomei et al. (2025), prompted stereo reasoning Wang et al. (2026), and uncertainty-aware or selective prediction Lou et al. (2023); Geifman & El-Yaniv (2019). These approaches improve dense reconstruction, introduce internal geometric cues, or estimate prediction reliability, whereas our focus is on how an external spatial query conditions stereo correspondence and organizes the resulting metric geometry around the query. Thus, FocusStereo is not a post-processing module tied to a particular stereo estimator, but a query-conditioned stereo reasoning framework.

Experiments on VKITTI2 and HAMMER show consistent improvements in query-relevant geometry across substantially different scene scales. FocusStereo reduces Focus AbsRel by 65.17% and 27.04%, and Query-XYZ error by 61.13% and 15.37%, respectively, relative to the strongest competing result for each metric. Controlled ablations with fixed stereo initialization further show that query-relevant gains can be obtained while the global metric geometry remains largely stable.

Our main contributions are:

- **Binocular Focused Vision.** We formulate stereo perception as query-conditioned geometric reasoning, where an external spatial query conditions correspondence formation and query-centered 3D representation rather than merely selecting a region after depth prediction.
- **FocusStereo.** We introduce a unified framework coupling query-conditioned correspondence reconstruction, bounded metric geometry, reliability-aware fusion, and geometry-anchored spatial reasoning.
- **Query-centric evaluation.** We establish a unified protocol for evaluating global integrity, focused depth, query-point 3D localization, relative geometry, and selective reliability across VKITTI2 and HAMMER.

2 Related Work

2.1 Deep Stereo Matching and Generalization

Deep stereo matching has progressed from learned cost-volume reasoning to iterative and global correspondence modeling. PSMNet, GC-Net, GwcNet, and GA-Net established end-to-end volumetric matching and regularization Chang & Chen (2018); Kendall et al. (2017); Guo et al. (2019); Zhang et al. (2019), while DeepPruner, AANet, LEAStereo, HITNet, and CFNet improved efficiency and robustness through search-space reduction, adaptive aggregation, architecture design, multi-resolution propagation, and cascaded refinement Duggal et al. (2019); Xu & Zhang (2020); Cheng et al. (2020); Tankovich et al. (2023); Shen et al. (2021). More recent methods emphasize iterative or long-range reasoning: RAFT-Stereo performs recurrent refinement over correlation volumes Lipson et al. (2021), IGEV-Stereo incorporates geometry encoding Xu et al. (2023a), and STTR, CREStereo, GM-Stereo, GOAT, NMRF, and Selective-Stereo further explore Transformer matching, adaptive correlation, global aggregation, structured message passing, and frequency-aware refinement Li et al. (2021; 2022); Xu et al. (2023b); Liu et al. (2024); Guan et al. (2024); Wang et al. (2024).

Generalization has meanwhile been improved through large-scale synthetic data such as Scene Flow and synthetic-to-real training strategies Mayer et al. (2016); Zhang et al. (2024). These advances substantially improve dense disparity estimation across scenes and domains, but their primary objective remains recovery of a complete task-agnostic disparity field. Our problem differs in asking how stereo computation can be selectively conditioned on a spatial query while preserving global metric stability.

2.2 Foundation Stereo and Prompted Reasoning

Recent stereo systems increasingly leverage foundation-model priors and large-scale training to improve zero-shot generalization. FoundationStereo combines synthetic stereo data, monocular foundation features, and long-range cost-volume reasoning Wen et al. (2025); DEFOM-Stereo and Stereo Anywhere incorporate monocular depth priors into binocular estimation Jiang et al. (2025); Bartolomei et al. (2025); ZeroStereo expands training diversity through synthesized stereo data Wang et al. (2025); and Fast-FoundationStereo targets more efficient foundation-scale inference Wen et al. (2026).

PromptStereo introduces structural and motion cues into iterative stereo inference Wang et al. (2026), while other work combines binocular matching with additional optical cues Liu et al. (2025). In these approaches, the conditioning signal typically serves as an internal geometric prior or sensing cue. FocusStereo instead treats a spatial query as an external geometric request: it conditions correspondence formation itself and subsequently defines the reference frame for query-centered spatial reasoning.

2.3 Query-Driven Perception and Reliability

Active perception argues that sensing should be organized around task relevance rather than treating all observations uniformly Bajcsy (1988); Tsotsos (2020). Active stereo, adaptive monocular sampling, and studies of binocular vergence similarly adjust sensing or represent depth relative to a fixation point Klimmasch et al. (2021); Ahuja & Abbott (1993); Tilmon & Koppal (2021); Neri et al. (2004). These works motivate spatially selective geometry, but generally rely on viewpoint, optical, or sampling changes. FocusStereo instead operates on a fixed binocular observation, using an external spatial query to alter geometric reasoning rather than data acquisition.

Uncertainty estimation and selective prediction provide complementary mechanisms for controlling unreliable predictions. Kendall and Gal distinguish aleatoric and epistemic uncertainty Kendall & Gal (2017); Deep Evidential Regression and ELFNet learn predictive uncertainty through evidential formulations Amini et al. (2020); Lou et al. (2023); and SelectiveNet explicitly optimizes prediction with rejection under a risk–coverage criterion Geifman & El-Yaniv (2019). FocusStereo builds on this principle but couples reliability with query-conditioned geometric updating: it separates whether a local update should be accepted from how strongly it should modify the metric state, and expresses the resulting geometry relative to the query.

Together, these directions improve stereo accuracy, generalization, adaptive sensing, and reliability, but do not jointly use an external spatial query to condition stereo correspondence, metric updating, and query-centered spatial reasoning within a fixed binocular observation. This is the gap addressed by FocusStereo.

3 Method

3.1 Problem Formulation and Framework Overview

Given rectified stereo images $I_l, I_r \in \mathbb{R}^{H\times W\times 3}$, camera intrinsics K, stereo baseline B, and a spatial query $q = (u_q, v_q)$, we aim to recover query-relevant metric geometry and query-centered spatial relations from a fixed binocular observation. The query specifies only an image-space location and does not assume a semantic category.

To preserve metric scale and initialize local correspondence search, a generic stereo estimator provides an initial depth D_0 and confidence C_0,

$$(D_0, C_0) = \mathcal{S}_{\text{init}}(I_l, I_r). \tag{1}$$

FocusStereo then implements the query-conditioned mapping:

$$(D, P_q, \Delta P_q, R_q, U_q) = \mathcal{G}_\theta\left(I_l, I_r, q, D_0, C_0; K, B\right), \tag{2}$$

where D is the final metric depth, P_q the query-associated 3D position, ΔP_q the query-centered 3D displacement field, R_q the ordinal relation field, and U_q predictive uncertainty. Importantly, (D_0, C_0) serves only as a metric anchor, search initialization, and fallback state; query-conditioned correspondence evidence is reconstructed directly from the binocular features. When the available evidence does not support a reliable update, the initialized geometry is preserved.

As illustrated in Fig. 1, FocusStereo consists of four sequential components. Q-SER reconstructs query-conditioned local stereo correspondence evidence around the initialized disparity; FGD decodes this evidence into bounded geometric proposals; RGGF selectively incorporates reliable updates into the metric state; and GARD derives query-centered 3D displacements and spatial relations from the final fused depth. Thus, spatial reasoning is anchored to the final geometry rather than to an intermediate proposal.

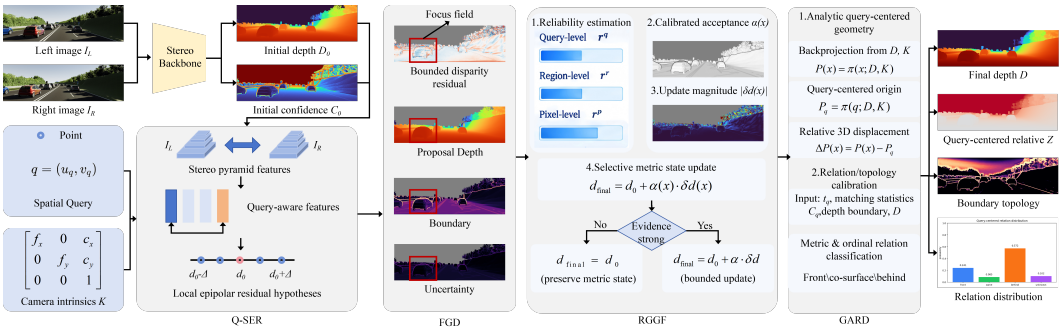


Figure 1: Overview of the proposed FocusStereo framework.

3.2 Query-Conditioned Stereo Evidence Reconstruction (Q-SER)

Q-SER reconstructs query-conditioned correspondence evidence directly from binocular features. The initial depth D_0 is used only to define the metric search center. We convert it to disparity as $d_0(x) = fB/D_0(x)$, and construct a soft query focus field that combines image-space proximity, compatibility with the initialized geometry, and a compact query kernel M_q:

$$W_q(x) = \mathrm{Norm}\left[\lambda_{xy} e^{-\frac{\|x-q\|_2^2}{2\sigma_{xy}^2}} + \lambda_z e^{-\frac{|D_0(x)-D_0(q)|}{\sigma_z}} + \lambda_m M_q(x)\right]. \tag{3}$$

Thus, W_q provides a query-dependent spatial support without requiring a semantic mask or ground-truth region.

A shared feature encoder extracts left/right representations F_v^s at scales $s \in \{4, 8\}$. We summarize the local query context into a token:

$$t_q = \phi\left([\mathrm{Pool}(F_l^4; M_q), D_0(q), r_x(q), r_y(q), C_0(q), \mathrm{Stat}(W_q)]\right), \tag{4}$$

which conditions both stereo branches through feature-wise affine modulation,

$$\widetilde{F}_v^s = F_v^s \odot [1 + \gamma_s(t_q)] + \beta_s(t_q), \quad v \in \{l, r\}. \tag{5}$$

The query therefore modifies the feature representation used for correspondence itself, rather than selecting a region only after dense stereo inference.

At each scale, Q-SER searches only a local residual range $k \in \{-K_s, \ldots, K_s\}$ around $d_0^s(x)$. For each hypothesis, the conditioned right feature is sampled at disparity $d_0^s(x) + k$, producing a matching response $c_s(x, k)$. Together with the resized focus field W_q^s, initial confidence C_0^s, query

response A_q^s, and normalized residual offset, these responses define a query-conditioned matching distribution:

$$p_s(k \mid x, q) = \mathrm{Softmax}_k \left[\mathcal{R}_s \left(c_s(x,k), W_q^s(x), C_0^s(x), A_q^s(x), \frac{k}{K_s} \right) \right]. \tag{6}$$

Rather than retaining only the most likely hypothesis, we summarize each distribution by its expected residual, normalized entropy, and peak probability,

$$\mu_s = \sum_k k p_s(k \mid x, q), \quad \mathcal{H}_s = \frac{-\sum_k p_s(k \mid x, q) \log p_s(k \mid x, q)}{\log |\mathcal{K}_s|}, \quad P_s^{\max} = \max_k p_s(k \mid x, q). \tag{7}$$

The fine scale preserves local disparity structure, while the coarse scale provides broader context and a larger effective search range. These multi-scale statistics jointly encode both the likely correspondence correction and the strength of its binocular support, and are passed to FGD and RGGF for geometric decoding and reliability-aware fusion.

3.3 Focus Geometry Decoding (FGD)

FGD converts the query-conditioned correspondence evidence from Q-SER into a bounded geometric proposal rather than directly regressing absolute depth. Let $\mathcal{Z}_q(x)$ denote the fused representation of multi-scale correspondence statistics, query-conditioned features, image features, the focus field, and initialized geometry. The decoder predicts:

$$(h_d(x), B_q(x), U_q(x)) = \mathcal{D}_{\mathrm{FGD}} \left(\mathcal{Z}_q(x) \right). \tag{8}$$

where $h_d(x)$ is a normalized disparity residual, $B_q(x)$ captures local depth discontinuities, and $U_q(x)$ estimates proposal uncertainty.

To keep the update within the correspondence range supported by the current observation, we bound the residual in disparity space:

$$\begin{aligned} L(x) &= \mathrm{clip} \left(\alpha_L d_0(x) + \beta_L, L_{\min}, L_{\max} \right), \\ \Delta d_q(x) &= L(x) h_d(x), \\ d_p(x) &= d_0(x) + W_q(x) \Delta d_q(x). \end{aligned} \tag{9}$$

The corresponding proposal depth is $D_p(x) = fB/d_p(x)$. Thus, W_q spatially localizes the update, while the bounded residual keeps the proposal anchored to the initialized stereo geometry. B_q and U_q provide structural and reliability cues for the subsequent fusion stage. Importantly, D_p is only an intermediate proposal; whether and how strongly it modifies the metric state is determined by RGGF.

3.4 Reliability-Guided Geometry Fusion (RGGF)

Local proposals from FGD are not uniformly reliable, especially around occlusions, weak textures, reflective surfaces, and depth discontinuities. RGGF therefore treats proposal fusion as a selective geometric update. It estimates query-level measurability p_q, region-level consistency $p_r(x \mid q)$, and pixel-level correspondence reliability $p_x(x \mid q)$, and combines them as:

$$s_q(x) = p_q p_r(x \mid q) p_x(x \mid q). \tag{10}$$

These estimates are conditioned on query-aware matching statistics, initial confidence, structural cues, proposal uncertainty, and the candidate update. Crucially, reliability determines whether an update should be accepted, rather than serving directly as its blending weight.

Given a reliability threshold τ calibrated on held-out data, RGGF separates binary acceptance from update magnitude:

$$\begin{aligned} g_q(x) &= \mathbb{I}[s_q(x) \geq \tau], \quad a_q(x) \in [0,1], \\ d(x) &= d_0(x) + g_q(x) a_q(x) \left[d_p(x) - d_0(x) \right] \\ &= d_0(x) + g_q(x) a_q(x) W_q(x) \Delta d_q(x) \end{aligned} \tag{11}$$

with final metric depth $D(x) = fB/d(x)$. When the evidence is insufficient, $g_q(x) = 0$ preserves the initialized geometry; otherwise, only a bounded fraction $a_q(x)$ of the proposed update is incorporated. The resulting D is the final metric geometry used by GARD.

3.5 GEOMETRY-ANCHORED RELATION DECODING (GARD)

Given the final metric depth D, GARD constructs query-centered 3D geometry analytically rather than regressing coordinates independently. Each pixel is back-projected as $P(x) = \pi^{-1}\big(x, D(x); K\big)$. To reduce sensitivity to single-pixel depth noise, we define the query origin from a compact core C_q and express the surrounding geometry relative to it:

$$\begin{aligned} P_q &= \frac{\sum_{x \in C_q} w_q(x) P(x)}{\sum_{x \in C_q} w_q(x) + \epsilon}, \\ \Delta P_q(x) = P(x) - P_q, &\quad \Delta Z_q(x) = P_z(x) - P_{q,z}, \end{aligned} \tag{12}$$

where w_q combines query-centered weighting with depth validity. Thus, ΔP_q preserves metric displacement while establishing the query as the local 3D reference.

To account for the increasing depth uncertainty with distance, we use a distance-adaptive tolerance:

$$m_q = \mathrm{clip}\left(\eta_z P_{q,z} + m_0, m_{\min}, m_{\max}\right), \quad R_q(x) = \begin{cases} \text{front}, & \Delta Z_q(x) < -m_q, \\ \text{co-surface}, & |\Delta Z_q(x)| \le m_q, \\ \text{behind}, & \Delta Z_q(x) > m_q. \end{cases} \tag{13}$$

A lightweight relation calibrator uses the query representation, matching statistics, and depth-boundary cues to calibrate ordinal confidence and local topology. Importantly, $P(x)$, P_q, and $\Delta P_q(x)$ remain analytically determined by the fused metric depth; learning is restricted to relation ambiguity rather than free 3D coordinate regression.

3.6 LEARNING OBJECTIVES

FocusStereo is trained following the dependency of its inference pipeline: geometry proposal formation, reliability-guided fusion, and relation calibration. We therefore adopt a three-stage optimization strategy, while executing the complete pipeline sequentially at inference time. The objectives are:

$$\begin{aligned} \mathcal{L}_{\text{geo}} &= \lambda_d \mathcal{L}_{\text{depth}} + \lambda_f \mathcal{L}_{\text{focus}} + \lambda_b \mathcal{L}_{\text{boundary}} + \lambda_c \mathcal{L}_{\text{cost}} + \lambda_u \mathcal{L}_{\text{uncertainty}} + \lambda_h \mathcal{L}_{\text{no-harm}}, \\ \mathcal{L}_{\text{fuse}} &= \lambda_{\text{cls}} \mathcal{L}_{\text{focal}}(s_q, y_{\text{gain}}) + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}} + \lambda_{\text{mag}} \mathcal{L}_{\text{smoothL1}}(a_q, a_q^*) \\ &\quad + \lambda_{\text{cal}} \mathcal{L}_{\text{calibration}} + \lambda_{\text{safe}} \mathcal{L}_{\text{no-harm}}, \\ \mathcal{L}_{\text{rel}} &= \lambda_{\text{ord}} \mathcal{L}_{\text{ordinal}} + \lambda_{\text{topo}} \mathcal{L}_{\text{topology}} + \lambda_{\text{id}} \mathcal{L}_{\text{identity}}. \end{aligned} \tag{14}$$

In the first stage, Q-SER and FGD learn query-conditioned correspondence distributions and bounded geometry proposals; the losses jointly supervise metric geometry, query-focused accuracy, local boundaries, matching evidence, uncertainty, and preservation of reliable initial geometry. In the second stage, the proposal pathway is fixed and RGGF learns update acceptance and magnitude, where $y_{\text{gain}} \in \{0, 1\}$ indicates whether the proposal improves the initialized geometry and a_q^* denotes the supervised target magnitude along the proposal direction. Finally, GARD is calibrated on the fused geometry using ordinal, topology, and identity objectives. No independent 3D coordinate regression is introduced; metric coordinates and relative displacements remain analytically determined by the final depth and camera geometry.

3.7 EVALUATION PROTOCOL FOR BINOCULAR FOCUSED VISION

Conventional stereo metrics average over all valid pixels and therefore cannot distinguish query-relevant improvements from changes in global reconstruction. We define four fixed evaluation regions before testing: the global valid domain Ω, query-focused region Ω_q, compact query core C_q, and query context Ω_q^{ctx}. Predicted and ground-truth disparities are obtained from metric depth as $d = fB/D$ and $d^* = fB/D^*$.

We evaluate global integrity using the standard Global AbsRel and D1-all over Ω, and focused accuracy using AbsRel and RMSE restricted to Ω_q. For robust query-level disparity measurement, we use the median disparity within the query core,

$$d_q = \mathrm{median}_{x \in C_q} d(x), \quad d_q^* = \mathrm{median}_{x \in C_q} d^*(x), \quad \mathrm{FDE} = |d_q - d_q^*|. \tag{15}$$

For query-centered 3D geometry, we recover $D_q = fB/d_q$ and $D_q^* = fB/d_q^*$, back-project them to $\widehat{P}_q = \pi^{-1}(q, D_q; K)$ and $P_q^* = \pi^{-1}(q, D_q^*; K)$, and define:

$$E_{\text{QueryXYZ}} = \|\widehat{P}_q - P_q^*\|_2,$$
$$E_{\text{RelXYZ}} = \frac{1}{|\Omega_q^{\text{ctx}}|} \sum_{x \in \Omega_q^{\text{ctx}}} \left\| \left(P(x) - \widehat{P}_q\right) - \left(P^*(x) - P_q^*\right) \right\|_2. \tag{16}$$

Query-XYZ measures absolute localization of the queried position, whereas Relative-XYZ evaluates surrounding 3D structure in the query-centered reference frame.

Finally, we evaluate selective reliability using the standard risk–coverage formulation,

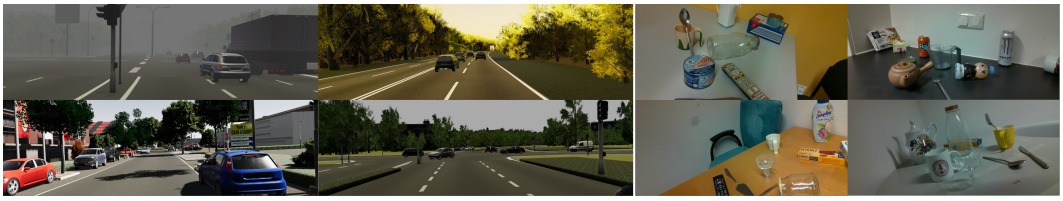
$$R(c) = \frac{1}{|\Omega_c|} \sum_{x \in \Omega_c} \frac{|D(x) - D^*(x)|}{D^*(x)}, \quad \text{AURC} = \int_0^1 R(c)dc, \tag{17}$$

where Ω_c contains predictions retained at coverage c after confidence ranking. Lower AURC indicates better identification of unreliable predictions. Our eight primary lower-is-better metrics are Global AbsRel, D1-all, Focus AbsRel, Focus RMSE, FDE, Query-XYZ, Relative-XYZ, and AURC, covering global integrity, focused accuracy, query-centered 3D geometry, and selective reliability.

4 Experiments

4.1 Experimental Setup and Evaluation Protocol

We evaluate FocusStereo on Virtual KITTI 2 (VKITTI2) Cabon et al. (2020) and HAMMER Jung et al. (2023), which provide complementary geometric regimes: VKITTI2 contains large-scale outdoor driving scenes with substantial depth variation, whereas HAMMER emphasizes close-range indoor manipulation with fine-grained geometry and frequent occlusions. Representative scenes from both datasets are shown in Fig. 2. Both datasets are evaluated using rectified stereo images, camera calibration, and fixed spatial queries under the common protocol defined in Sec. 3.7. Test samples, query locations, evaluation regions, and reliability thresholds are fixed before evaluation.



(a) (b)

Figure 2: Representative scenes from the evaluation datasets. (a) VKITTI2: large-scale outdoor driving scenes; (b) HAMMER: close-range indoor manipulation scenes. The two datasets complement each other in scene scale and local geometric structure.

We compare with Depth Anything 3 (DA3) Lin et al. (2025), RAFT-Stereo Lipson et al. (2021), IGEV-Stereo Xu et al. (2023a), GMStereo Xu et al. (2023b) and Selective-Stereo Wang et al. (2024). Stereo baselines use their SceneFlow-trained weights Mayer et al. (2016) in the cross-domain setting, and all methods share the same valid-depth masks, query cores, focus regions, and evaluation scripts. FocusStereo initializes (D_0, C_0) with FoundationStereo Wen et al. (2025); as described in Sec. 3, this state provides metric scale and correspondence-search initialization rather than the query-conditioned solution. The component analysis in Sec. 4.4 further controls for this initializer by using the same initial stereo state across variants.

Following Sec. 3.7, we report Global AbsRel, D1-all, Focus AbsRel, Focus RMSE, FDE, Query-XYZ, Relative-XYZ, and AURC as the eight primary lower-is-better metrics.

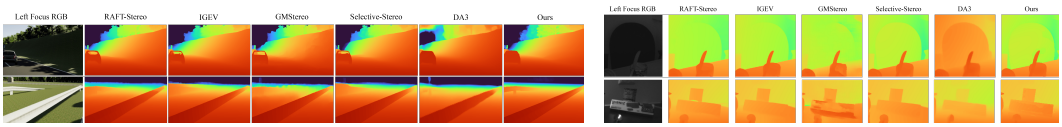
Table 1: Quantitative comparison on (a) VKITTI2 and (b) HAMMER. Lower is better for all metrics; best results are shown in bold.

(a) VKITTI2

Metric	DA3	RAFT	IGEV	GM	Selective	FocusStereo
Global AbsRel	0.14528	0.04978	0.09101	0.05635	0.06270	**0.01037**
D1-all	0.57355	0.04364	0.04817	0.05219	0.04337	**0.00788**
Focus AbsRel	0.17500	0.06014	0.06541	0.06254	0.07038	**0.02095**
Focus RMSE (m)	5.40475	2.77528	3.13750	2.92260	3.39508	**1.74246**
FDE (px)	4.95361	1.11162	1.18130	1.10375	1.23617	**0.21137**
Query-XYZ (m)	4.44549	1.25001	1.27070	1.20330	1.54763	**0.46776**
Relative-XYZ (m)	3.89242	1.87821	1.97140	1.90413	2.31000	**0.80786**
AURC	0.12322	0.04398	0.08887	0.05016	0.05037	**0.01777**

(b) HAMMER

Metric	DA3	RAFT	IGEV	GM	Selective	FocusStereo
Global AbsRel	0.13626	0.05830	0.06877	0.14370	0.06062	**0.03684**
D1-all	0.54694	0.12841	0.15337	0.43599	0.13779	**0.02730**
Focus AbsRel	0.13727	0.06327	0.05934	0.09122	0.06131	**0.04330**
Focus RMSE (m)	0.10250	0.05417	0.05274	0.08353	0.05488	**0.04445**
FDE (px)	4.02804	1.94256	1.80407	2.48862	1.77838	**1.41703**
Query-XYZ (m)	0.09730	0.04585	0.04183	0.06185	0.04269	**0.03540**
Relative-XYZ (m)	0.06981	0.04222	0.04193	0.08727	0.04142	**0.02889**
AURC	0.11509	0.05972	0.07208	0.15799	0.06068	**0.03571**



(a) VKITTI2 (b) HAMMER

Figure 3: Qualitative comparisons on VKITTI2 and HAMMER. The first column shows the queried region, followed by depth predictions from the compared methods over the same spatial region.

4.2 Quantitative Comparison with Existing Methods

Table 1 reports the quantitative results on VKITTI2 and HAMMER. FocusStereo achieves the lowest error across all eight primary metrics on both benchmarks. On VKITTI2, relative to the strongest competing result for each metric, it reduces Focus AbsRel by 65.17%, Query-XYZ by 61.13%, and Relative-XYZ by 56.99%. The corresponding reductions on HAMMER are 27.04%, 15.37%, and 30.23%, respectively. These consistent gains across large-scale outdoor and close-range indoor scenes indicate that the benefit of query-conditioned reasoning extends across substantially different geometric regimes.

4.3 Qualitative Comparison

Fig. 3 presents qualitative comparisons on VKITTI2 and HAMMER using the same input image, spatial query, and focus region for all methods. Across both datasets, competing methods exhibit varying degrees of boundary diffusion, over-smoothing, or unstable local geometry around thin structures and occlusion boundaries. In contrast, FocusStereo produces more coherent within-surface geometry while preserving clearer separation across local depth discontinuities.

These observations are consistent with the quantitative results in Table 1: FocusStereo concentrates its improvements on query-relevant local geometry, reducing depth leakage across discontinuities and better preserving fine-grained front-to-back structure rather than globally rewriting the depth field.

4.4 Component Analysis and Controlled Ablations

Table 2 reports controlled component ablations on (a) VKITTI2 and (b) HAMMER. The ablation results are obtained from a separate controlled run and are therefore used only for within-table comparisons across model variants; their absolute values are not directly compared with the main benchmark results in Table 1. Within this ablation run, all variants share the same stereo initialization and evaluation setting, allowing the effect of individual components to be isolated from the initializer.Here, Harm Rate measures the incidence of harmful updates relative to the initialized geometry, while Topology F1 evaluates query-centered relation topology. All variants share the same stereo initialization and evaluation setting, allowing the effect of query-conditioned reasoning to be isolated from the initializer. Adding Q-SER + FGD improves query-relevant geometry while leaving global depth largely unchanged. On VKITTI2, Query-XYZ decreases from 0.370475 to 0.358760 and Relative-XYZ from 0.527808 to 0.520790. On HAMMER, Focus AbsRel decreases from 0.052623 to 0.037934, while Query-XYZ decreases from 0.032473 to 0.015995. These results show that the proposal pathway primarily improves local and query-centered geometry rather than globally rewriting the initialized metric state.

Table 2: Controlled component ablations on (a) VKITTI2 and (b) HAMMER. ↓ indicates lower is better and ↑ indicates higher is better.

(a) VKITTI2

Variant	G-Abs↓	F-Abs↓	F-RMSE↓	Q-XYZ↓	R-XYZ↓	Topo.↑	Harm↓
Stereo init.	0.010580	0.007591	1.3751	0.370475	0.527808	0.82735	0.0000%
Q-SER + FGD	0.010575	0.007582	1.3650	0.358760	0.520790	0.82735	0.0039%
+ RGGF	0.010577	0.007585	1.3657	0.358668	0.520969	0.82735	0.0037%
+ GARD	0.010577	0.007585	1.3657	0.358675	0.520974	0.82937	0.0037%

(b) HAMMER

Variant	G-Abs↓	F-Abs↓	F-RMSE↓	Q-XYZ↓	R-XYZ↓	Topo.↑	Harm↓
Stereo init.	0.037741	0.052623	0.0625	0.032473	0.021208	0.71248	0.0000%
Q-SER + FGD	0.036381	0.037934	0.0554	0.015995	0.023732	0.71248	0.3256%
+ RGGF	0.036788	0.045398	0.0580	0.031453	0.023593	0.71248	0.1263%
+ GARD	0.036788	0.045398	0.0580	0.031453	0.023593	0.71829	0.1263%

RGGF primarily controls update risk: Harm Rate decreases from 0.0039% to 0.0037% on VKITTI2 and from 0.3256% to 0.1263% on HAMMER, although some mean geometric errors partially regress on HAMMER, reflecting the intended accuracy–risk trade-off. GARD leaves continuous geometry essentially unchanged while improving Topology F1 from 0.82735 to 0.82937 on VKITTI2 and from 0.71248 to 0.71829 on HAMMER. Together, the ablations support the intended decomposition of FocusStereo: Q-SER/FGD generate query-conditioned geometry, RGGF performs risk-controlled fusion, and GARD calibrates query-centered spatial relations.

5 Conclusion

We introduced Binocular Focused Vision, a query-conditioned formulation of stereo perception that shifts the objective from uniform dense reconstruction toward spatially selective geometric reasoning. Based on this formulation, FocusStereo conditions stereo correspondence on an explicit spatial query, performs reliability-aware metric updating, and derives query-centered 3D relations from the resulting geometry. Experiments on VKITTI2 and HAMMER demonstrate consistent improvements in focused depth, query-point localization, relative 3D geometry, and selective reliability, supporting spatial queries as an effective condition for stereo geometric reasoning.

The current formulation is limited to point queries, staged optimization, and evaluation on controlled benchmark settings. Future work will extend FocusStereo to richer query representations, more unified end-to-end training, and broader real-world interactive and robotic scenarios.

AI use statement

In this work, generative AI tools were used to assist with language translation and editing, manuscript organization and drafting, literature search and summarization, methodological discussion, interpretation and presentation of experimental results, and the preparation and refinement of scientific figures. AI-assisted outputs were treated as drafts or suggestions and were reviewed by the authors. All technical content, equations, experimental results, citations, and figures were checked against the underlying implementation, data, and source literature. The authors made the final research and writing decisions and take full responsibility for the content of this work.

References

N. Ahuja and A.L. Abbott. Active stereo: integrating disparity, vergence, focus, aperture and calibration for surface estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1007–1029, 1993. doi: 10.1109/34.254059.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression, 2020. URL https://arxiv.org/abs/1910.02600.

R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. doi: 10.1109/5.5968.

Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano 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)*, pp. 1013–1027, 2025. doi: 10.1109/CVPR52734.2025.00103.

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

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018. doi: 10.1109/CVPR.2018.00567.

Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4383–4392, 2019. doi: 10.1109/ICCV.2019.00448.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5459–5469, 2024. doi: 10.1109/CVPR52733.2024.00522.

Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3268–3277, 2019. doi: 10.1109/CVPR.2019.00339.

Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching, 2025. URL https://arxiv.org/abs/2501.09466.

HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, Ales Leonardis, Nassir Navab, and Benjamin Busam. On the importance of accurate geometry data for dense 3d vision tasks, 2023. URL https://arxiv.org/abs/2303.14840.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision?, 2017. URL https://arxiv.org/abs/1703.04977.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression, 2017. URL https://arxiv.org/abs/1703.04309.

Lukas Klimmasch, Johann Schneider, Alexander Lelais, Maria Fronius, Bertram Emil Shi, and Jochen Triesch. The development of active binocular vision under normal and alternate rearing conditions. *eLife*, 10:e56212, aug 2021. ISSN 2050-084X. doi: 10.7554/eLife.56212. URL https://doi.org/10.7554/eLife.56212.

Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation, 2022. URL https://arxiv.org/abs/2203.11483.

Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6177–6186, 2021. doi: 10.1109/ICCV48922.2021.00614.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views, 2025. URL https://arxiv.org/abs/2511.10647.

Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pp. 218–227, 2021. doi: 10.1109/3DV53792.2021.00032.

Yuhui Liu, Liangxun Ou, Qiang Fu, Hadi Amata, Wolfgang Heidrich, and Yifan Peng. Learned binocular-encoding optics for rgbd imaging using joint stereo and focus cues. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15833–15842, 2025. doi: 10.1109/CVPR52734.2025.01476.

Zihua Liu, Yizhou Li, and Masatoshi Okutomi. Global occlusion-aware transformer for robust stereo matching. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3523–3532, 2024. doi: 10.1109/WACV57701.2024.00350.

Jieming Lou, Weide Liu, Zhuo Chen, Fayao Liu, and Jun Cheng. Elfnet: Evidential local-global fusion for stereo matching. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17738–17747, 2023. doi: 10.1109/ICCV51070.2023.01630.

Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4040–4048, 2016. doi: 10.1109/CVPR.2016.438.

Peter Neri, Holly Bridge, and David J Heeger. Stereoscopic processing of absolute and relative disparity in human visual cortex. *Journal of neurophysiology*, 92(3):1880–1891, 2004.

D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pp. 131–140, 2001. doi: 10.1109/SMBV.2001.988771.

Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13901–13910, 2021. doi: 10.1109/CVPR46437.2021.01369.

Vladimir Tankovich, Christian Häne, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching, 2023. URL https://arxiv.org/abs/2007.12140.

Brevin Tilmon and Sanjeev J. Koppal. Saccadecam: Adaptive visual attention for monocular depth sensing, 2021. URL https://arxiv.org/abs/2103.12981.

John K. Tsotsos. *Active Vision*, pp. 1–9. Springer Berlin Heidelberg, Berlin, Heidelberg, 2020. ISBN 978-3-642-41610-1. doi: 10.1007/978-3-642-41610-1_105-1. URL https://doi.org/10.1007/978-3-642-41610-1_105-1.

Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching, 2024. URL https://arxiv.org/abs/2403.00486.

Xianqi Wang, Hao Yang, Gangwei Xu, Junda Cheng, Min Lin, Yong Deng, Jinliang Zang, Yurui Chen, and Xin Yang. Zerostereo: Zero-shot stereo matching from single images, 2025. URL https://arxiv.org/abs/2501.08654.

Xianqi Wang, Hao Yang, Hangtian Wang, Junda Cheng, Gangwei Xu, Min Lin, and Xin Yang. Promptstereo: Zero-shot stereo matching via structure and motion prompts, 2026. URL https://arxiv.org/abs/2603.01650.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching, 2025. URL https://arxiv.org/abs/2501.09898.

Bowen Wen, Shaurya Dewan, and Stan Birchfield. Fast-foundationstereo: Real-time zero-shot stereo matching, 2026. URL https://arxiv.org/abs/2512.11130.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching, 2023a. URL https://arxiv.org/abs/2303.06615.

Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1956–1965, 2020. doi: 10.1109/CVPR42600.2020.00203.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation, 2023b. URL https://arxiv.org/abs/2211.05783.

Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H. S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching, 2019. URL https://arxiv.org/abs/1904.06587.

Jiawei Zhang, Jiahe Li, Lei Huang, Xiaohan Yu, Lin Gu, Jin Zheng, and Xiao Bai. Robust synthetic-to-real transfer for stereo matching, 2024. URL https://arxiv.org/abs/2403.07705.