FoveaCam Duo: Foveated Stereo for Standoff Depth Sensing

Yuxuan Zhang*, Jacob Carter*, Hannah Kirkland, Michael Tomodakis, Noah Ralph, and Sanjeev J. Koppal†

Abstract—Robust depth perception is fundamental to autonomous navigation, manipulation, and scene understanding. Here we apply the benefits of foveation to depth perception with *FoveaCam Duo*: a passive stereo system inspired by biological foveation and crypsis. Our system uses a pair of high-resolution telephoto cameras steered by fast MEMS mirrors to produce high quality, dense depth maps at any designated region of interest with minimal processing. However, the MEMS-steered optics introduce angle-dependent distortion that cannot be accounted for with traditional stereo calibration, which assumes fixed camera geometry. To account for this, we introduce a three-stage calibration pipeline that maps MEMS mirror voltages to camera pointing angles and rectifies the foveated images via affine regression. We analyze how extending detection range quadratically expands the feasible operating region for standoff observation, and our simulation confirms that multi-fovea agents achieve substantially higher target coverage than single-fovea counterparts. In our evaluation, *FoveaCam Duo* maintains accurate depth estimation at distances where conventional stereo fails entirely. We also release a raster-scanned foveated stereo dataset with multi-viewpoint captures along a linear rail to compare to traditional stereo. In addition, we present a self supervised stereo algorithm to systematically compare to out-of-the-box and foveation-adapted state of the art depth prediction methods, demonstrating the integrity of stereo correspondence in our captures.

Index Terms—Embodied Sensing, Depth Sensing, Foveated Imaging, Stereo Vision, Computer Vision, Robotics

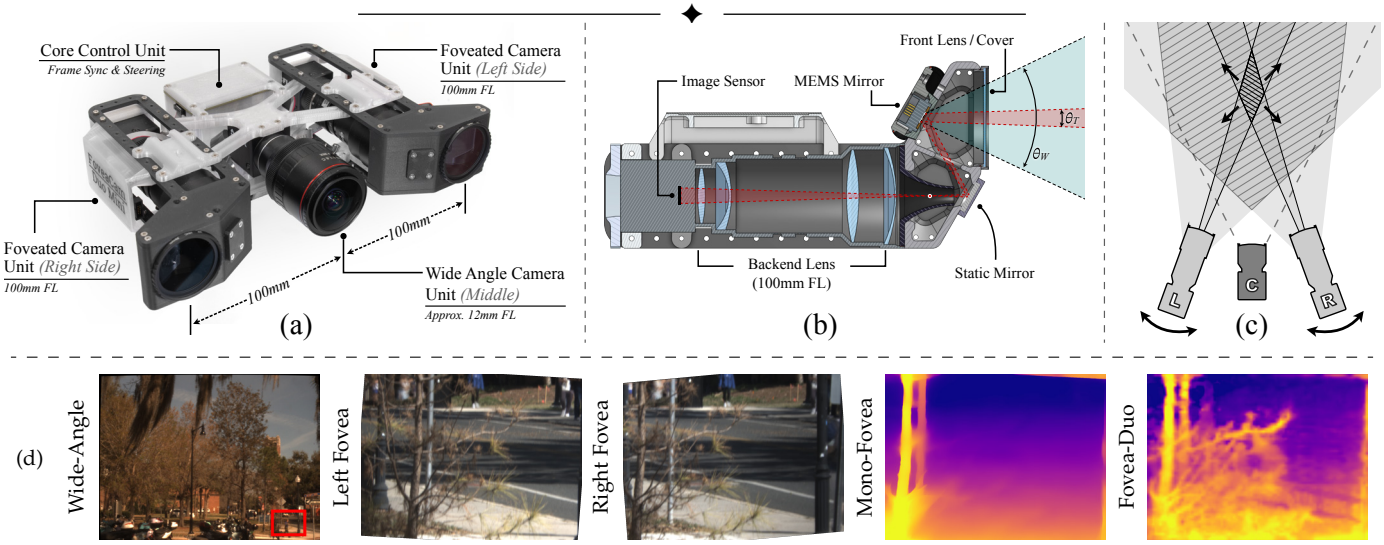


Fig. 1: *FoveaCam Duo* system overview and algorithm visualization. **(a)** Front view of the *FoveaCam Duo* assembly, with the left and right foveated cameras spaced 100 mm to either side of the central wide-angle camera. **(b)** Cross-section of a foveated telephoto camera, showing the MEMS mirror that rapidly steers the narrow telephoto field of view across regions of interest. **(c)** *FoveaCam Duo* utilizes two MEMS-mirror-steered telephoto cameras **L**, **R**, dynamically redirecting the foveated field of view within an extended region to deliver high-resolution depth at any point of attention. **(d)** Visualizations of custom foveated depth algorithm in comparison to monocular foveated depth estimation.

1 Introduction

Robust depth sensing is fundamental to autonomous navigation, manipulation, and scene understanding. Although LiDAR provides accurate range measurements, its sparsity and hardware constraints limit its applicability in dense perception tasks. Stereo vision offers a scalable and passive alternative for dense depth estimation. However, achieving accurate depth across both near and far ranges remains a fundamental challenge.

In this work, we describe and demonstrate *FoveaCam Duo*, a foveated stereo camera system inspired by biological *foveation* and *crypsis* [10], [11]. In biological vision, sensing resources are non-uniformly distributed, with high acuity concentrated in a small foveal region and lower resolution in the periphery, enabling efficient allocation of limited sensing capacity. This non-uniform allocation gives rise to biological *crypsis*, the evolutionary pressure to detect

- * *Equal Contribution.*
- *All authors are affiliated with the FOCUS Lab, Dept. of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA. E-mail: zhangyuxuan@ufl.edu (corresponding author).*
- † *Dr. Sanjeev J. Koppal holds concurrent appointments as an Associate Professor of ECE at the University of Florida and as an Amazon Scholar at Amazon Robotics. This project was performed at the University of Florida and is not associated with Amazon Robotics.*

TABLE 1: Comparison of depth-sensing hardware across *active illumination* (emitter-limited range, often sparse), *motorized vergence stereo heads* (bulky, slow), *multi-resolution camera arrays* (expensive, bulky, no depth), and ***FoveaCam Duo***. A "−" denotes that the information was either unreported or not applicable.

Category	System	Passive	Dense	Foveation	High Ang.	Compact	Deep Learning	Mult. Fovea
Active Illumination	Active IR Stereo [1]	✗	✓	✗	✗	✓	✗	✗
	Flash LiDAR (ToF) [2]	✗	✓	✗	✗	✓	✗	✗
	Scanning LiDAR [3]	✗	✗	✗	✗	✓	✗	✗
Motorized Vergence Stereo	Humanoid Robots [4], [5], [6], [7], [8]	✓	✓	✗	✗	✓	✗	✗
Multi-Res. / Hybrid	AWARE-2 [9]	✓	−	✗	✓	✗	✗	✗
★ Ours	**FoveaCam Duo**	✓	✓	✓	✓	✓	✓	✓

without being detected, where predators and prey direct high-acuity foveal gaze to mutually observe targets [12]. We analyze the resulting reciprocal visibility trade-off and show that enabling multi-fovea pursuit with *FoveaCam Duo* increases both target coverage and its own chance of survival in hostile environments. The *FoveaCam Duo* system consists of a central wide-angle camera coupled with two telephoto cameras on each side steered by a MEMS (Micro-Electro-Mechanical System) mirror. The wide-angle camera provides a broad contextual view while the two telephoto cameras dynamically point to any region of interest within the wide field of view, delivering high angular resolution stereo at point of attention.

***FoveaCam Duo* provides a better tradeoff between coverage and resolution**: Conventional equiangular stereo assigns uniform angular sampling across the field of view, a strategy Marr characterized as "luxurious memory capacity" due to its inefficient allocation of sensing resources [13]. Rather than increasing the total pixel budget, *FoveaCam Duo* redistributes sensing resolution through foveation, allocating stereo pixels where disparity sensitivity degrades with distance. This targeted reallocation compensates for the quadratic growth of depth error, enabling recovery of distant structure while imaging nearby regions at lower but sufficient resolution.

In summary, our contributions are:

- **§3.1** The hardware implementation of a novel foveated stereo camera system combining a wide-angle reference camera with a stereo pair of MEMS-mirror-steered telephoto cameras;
- **§3.2** A three-stage calibration pipeline rectifying foveated images to a virtual pinhole projection model and a performance evaluation by measuring its target localization accuracy following a 3D trajectory;
- **§4.4** An analysis of the reciprocal visibility-stealth trade-off governing standoff observation and simulated validation that multi-fovea pursuit increases target coverage and survivability; and
- **§5** A self-supervised, foveation-aware stereo depth algorithm demonstrating the integrity of disparity signal and improved quality in comparison to wide-angle stereo.

2 Related Work

Classic Computational Stereo Matching. Passive stereo vision estimates depth by triangulating corresponding points across views. Semi-Global Matching (SGM) [14], [15] remains one of the most widely used dense two-frame stereo methods, approximating a global 2-D smoothness prior via 1-D paths to produce high-quality disparity maps. A fundamental limitation of classical stereo is the quadratic growth of depth error with distance [16], such that even sub-pixel matching precision yields large errors at long range when angular resolution is low.

Learning-Based Stereo Matching. Deep learning has replaced handcrafted costs with learned features and end-to-end disparity regression. DispNet [17] introduced correlation-based matching, while PSMNet [18] leveraged 3D cost volumes and multi-scale context for robustness. Recent methods adopt iterative refinement strategies inspired by optical flow, with RAFT-Stereo [19] and IGEV-style models [20], [21] achieving state of the art accuracy via recurrent updates over dense correlation volumes. Hybrid approaches incorporate monocular priors (Depth-Anything-V2 [22]) such as MonSter [23], while foundation-style stereo models [24] improve zero-shot generalization through large-scale synthetic training. Despite these advances, learned stereo remains limited by angular sampling: when distant targets project to only a few pixels, disparity information is irrecoverable, motivating hardware approaches that increase angular resolution in regions of interest.

Foveated and Active Vision. Biological vision systems address the resolution-bandwidth trade-off through foveation, concentrating photoreceptors in a high-acuity fovea with low-resolution periphery [25], [26], coupled with rapid saccadic eye movements to redirect attention [27]. This principle inspired active perception in robotics, where sensors are dynamically controlled to improve task performance [5], [7], [28], [29], [30]. Conventional active stereo systems use pan-tilt mechanisms [4], [6], [8], [31], [32], while MEMS mirrors provide a compact, low-latency alternative via fast two-axis beam steering [33], [34]. Recent MEMS-based foveated cameras [35], [36], [37], [38] demonstrate this for monocular depth and attention. Mao *et al.* [39] instead focus on aberration correction via foveation, optimizing image quality rather than increasing information through view steering.

Multi-Resolution and Hybrid Camera Systems. Hybrid systems combine cameras to overcome single-sensor trade-offs. Large arrays [40] and gigapixel systems [9] achieve high-resolution wide-field imaging but are impractical due to scale. Prior work extends MEMS foveation to multi-object tracking [41] and adaptive depth sensing [42] using wide-angle guidance. Our approach instead realizes *stereo* foveated cameras with a wide-angle guide, enabling rapid

TABLE 2: Optical performance and electrical actuation specifications of the *FoveaCam Duo* system.

Specification		Measurements
Optical Field of View	Wide	$25.6° \times 19.4°$
	Fovea	$2.86° \times 2.15°$
Angular Resolution	Wide	$\sim$55 px/°
	Fovea	$\sim$502 px/°
Eff. Aperture	Fovea	$\approx f/21.5$
Steering	Actuation Range	$20° \times 20°$
	Extended FoV	$22.9° \times 22.2°$
MEMS Actuation	Response Time	<1 ms
	Settle Time	(Worst Case) <2 ms
	Step Precision	$\sim$0.08 px/step

redirection of a telephoto pair to provide on-demand high angular resolution for depth estimation.

Robotic Motion Planning for Active Sensing. Swingler and Ferrari [43] unify sensing and motion planning geometrically, while Lu *et al.* [44] and Wei *et al.* [45] jointly optimize motion and sensing for visibility-aware tracking. *FoveaCam Duo* introduces MEMS-steered foveae as a low-latency sensing degree of freedom, potentially reducing reliance on platform motion to maintain visibility.

3 *FoveaCam Duo*: Design and Calibration

3.1 System Design

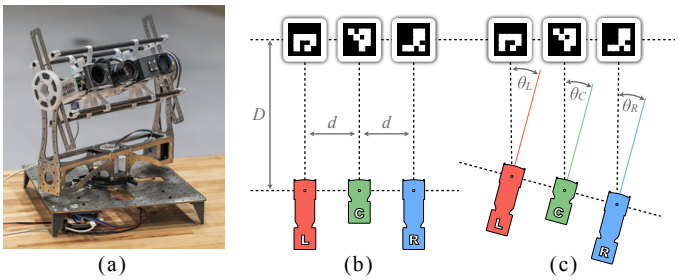
The *FoveaCam Duo* system (Fig. 1) consists of three cameras arranged on a horizontal baseline: a central wide-angle camera and two telephoto cameras positioned symmetrically on left and right side of the wide-angle camera, each equipped with a MEMS mirror that steers the telephoto optical axis. The telephoto component (Fig. 1) provides high angular resolution within a narrow field of view θ_T that can be dynamically directed to any region within the extended field of view θ_W. Prior MEMS mirror imaging systems suffered from ghost images caused by cover glass reflections [35], [36], [41]. Proposed mitigations, including post-processing and tilted cover glass, introduced artifacts of their own. *FoveaCam Duo* eliminates the problem entirely by removing the cover glass, using its semi-air-tight optical chamber depicted in Fig. 1 (b) for equivalent protection.

As listed in Table 2, the telephoto cameras achieve approximately 502 px/°, which is magnitudes higher than any known stereo setup built for depth sensing. The values are measured through a standard OpenCV checkerboard camera calibration pipeline [46]. Additionally, Table 2 reports the actuation performance specifications based on our measurements on optical performance and product specifications reported by the vendor of the MEMS mirrors.Derivation of reported metrics and model numbers of the off-the-shelf components used in *FoveaCam Duo* are listed in the supplementary material.

3.2 Dual Fovea Calibration

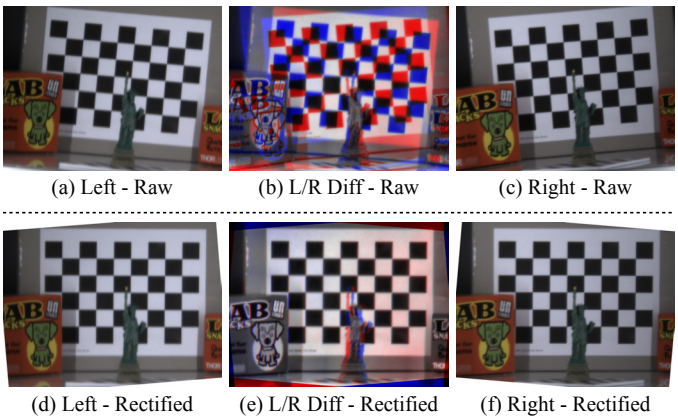
Calibration of the *FoveaCam Duo* system proceeds in three stages:

1) Intrinsic Calibration of the Wide-Angle Camera: The central wide-angle camera serves as the angular reference



(a) (b) (c)

Fig. 2: Extrinsic calibration setup. **(a)** The *FoveaCam Duo* system is mounted on a pan-tilt platform that sweeps the assembly through a grid of x and y angles. **(b)** Three ArUco markers, one per camera, are placed at distance D and spaced by the stereo baseline d. **(c)** Geometric diagram showing the angular error $\delta\theta = |\theta_C - \theta_L| \approx |\theta_C - \theta_R|$ introduced by geometry.



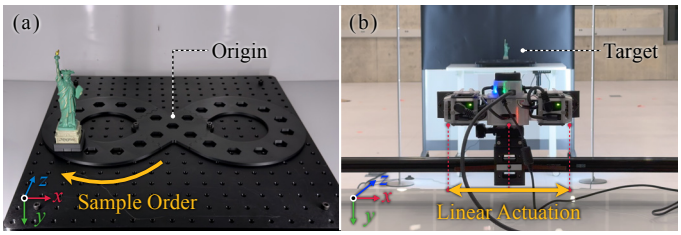
(a) Left - Raw (b) L/R Diff - Raw (c) Right - Raw

(d) Left - Rectified (e) L/R Diff - Rectified (f) Right - Rectified

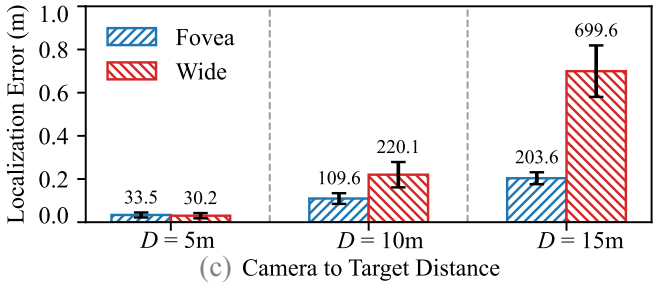
Fig. 3: Rectification of the foveated stereo pair. **Top:** raw images from the left and right foveated cameras and their red-blue anaglyph difference. The anaglyph (b) reveals severe misalignment. **Bottom:** after applying the angle-dependent correction $\mathbf{H}(\alpha_x, \alpha_y)$, the rectified anaglyph (e) shows that the left and right images are now well-aligned, with residual red-blue separation visible only on the foreground Statue of Liberty model due to its true stereo disparity.

for the entire system. We perform a standard intrinsic calibration [46] using a checkerboard to obtain the camera matrix and distortion coefficients, from which we derive a pixel-to-angle mapping. Specifically, for any pixel coordinate (p_x, p_y) in the wide-angle image, we compute the corresponding ray direction (α_x, α_y) relative to the camera's optical axis. This mapping is used in the subsequent extrinsic calibration stage to establish ground-truth poses for all three cameras.

2) Extrinsic Calibration with MEMS Voltage to Pointing Angle: Extrinsic calibration maps mirror control voltages (v_x, v_y) to pointing angles (α_x, α_y) of each foveated telephoto camera. We mount the system on a custom pan–tilt platform (Fig. 2 (a)) and sweep predefined poses while each of the three cameras tracks a dedicated ArUco marker. Markers are placed as in Fig. 2 (b) (spaced by the stereo baseline and aligned to the physical layout) so the angular subtense to each camera remains approximately constant.



(a-b) Target Localization Experiment Setup



(c) Camera to Target Distance

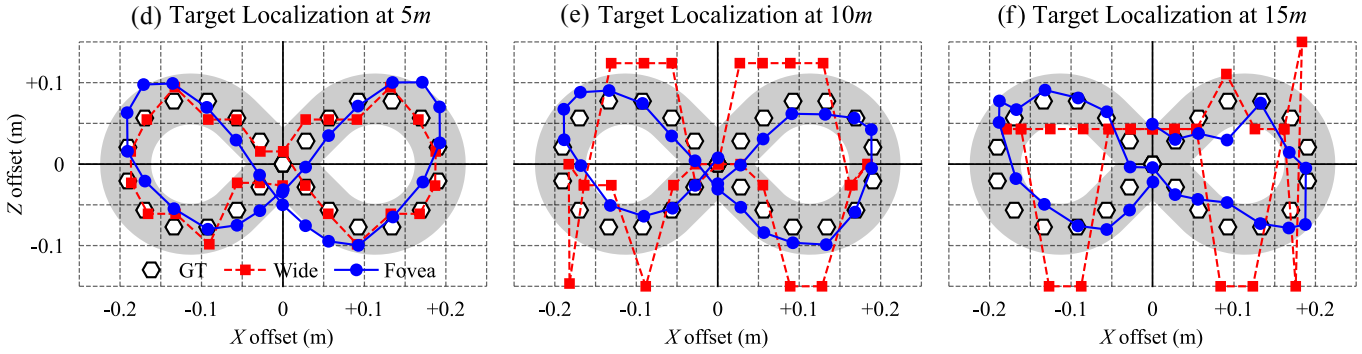


Fig. 4: Target localization experiment and results. **(a,b)** Setup: a Statue of Liberty model is placed at slots along a 3D-printed ∞-rail; a linear stage translates the center wide camera to form a virtual wide-angle stereo pair at identical baseline. **(c)** Localization error at 5/10/15 m (lower is better), with wide-angle stereo degrading sharply beyond 10 m. **(d–f)** Estimated trajectories for wide-angle "■" and foveated "●" stereo; *FoveaCam Duo* preserves the ∞ shape throughout while wide-angle fails past 10 m. Trajectories are centroid-aligned for display; mean drift is reported in (c). This confirms the triangulation advantage of *FoveaCam Duo* through high resolution fovea to compensate for quadratic depth error.

The induced geometric error is quantified by

$$\delta\theta = \arctan \frac{d - d\cos(\theta)}{D \pm d\sin(\theta)} \tag{1}$$

and is negligible for $D \gg d$ (Fig. 2 (b, c)). We sweep a 9×9 grid of (v_x, v_y), so that at each position the wide-angle camera observes its own marker and, using the intrinsic pixel-to-angle mapping (Sec. 3.2), provides ground-truth (α_x, α_y), while each foveated camera uses PID control to keep its marker centered. We fit a 3rd-order polynomial regression from stabilized driver voltages between (v_x, v_y) and (α_x, α_y) and find its inverse:

$$\begin{bmatrix} v_x \\ v_y \end{bmatrix} = K \cdot \begin{bmatrix} \alpha_x^3 & \alpha_y^3 & \alpha_x^2\alpha_y & \alpha_x\alpha_y^2 & \cdots & 1.0 \\ \alpha_x^3 & \alpha_y^3 & \alpha_x^2\alpha_y & \alpha_x\alpha_y^2 & \cdots & 1.0 \end{bmatrix} \tag{2}$$

$$\begin{bmatrix} \alpha_x \\ \alpha_y \end{bmatrix} = T \cdot \begin{bmatrix} v_x^3 & v_y^3 & v_x^2 v_y & v_x v_y^2 & \cdots & 1.0 \\ v_x^3 & v_y^3 & v_x^2 v_y & v_x v_y^2 & \cdots & 1.0 \end{bmatrix} \tag{3}$$

Reprojection results align well with ground truth, with regression plots in both voltage and angular space in Sec. 3 of the supplementary material.

3) Foveated Image Rectification: Images from the MEMS-steered telephoto optics exhibit angle-dependent affine distortion. For standard parallel-stereo processing, we rectify each foveated image into a virtual pinhole camera whose focal length matches the telephoto lens and whose optical axis is parallel to the wide-angle camera (classical rectified geometry). During calibration (Sec. 3.2), ArUco markers provide correspondences to estimate an affine projection $\mathbf{H}(\alpha_x, \alpha_y)$ that maps distorted foveated pixels to the virtual pinhole frame at sampled mirror angles (α_x, α_y). Marker geometry plus the wide-angle pixel-to-angle mapping yield target corner locations. Assuming negligible nonlinear lens distortion, a projective transform suffices (Fig. 3). Because $\mathbf{H}$ varies with steering, we fit each entry H_{ij} of the 3×3 matrix with a polynomial of (α_x, α_y) as

$$H_{ij}(\alpha_x, \alpha_y) = \sum_{m,n} c_{mn}^{(ij)}\, \alpha_x^m\, \alpha_y^n \tag{4}$$

including cross terms to capture axis coupling. At runtime we evaluate Eq. (4) to assemble $\mathbf{H}(\alpha_x, \alpha_y)$ and warp each image, with the rectified left/right pair then forming a standard parallel stereo input for disparity and triangulation. Detailed metrics are shown in the supplementary material.

3.3 Field Calibration Drift

While the MEMS mirror is highly repeatable with no measurable hysteresis in operation, the voltage-to-angle mapping drifts over time, predominantly as translational drift from thermal structural deformation, which is most noticeabe upon relocation of the system. We apply two mitigations: we perform one-shot in-field incremental calibration that corrects the linear term of the regression between voltage and pointing angle (Sec. 3.2), cancelling most translational drift without repeating the full pan-tilt sweep; we use metal structural reinforcement parts to reduce rotational drift. Residual drift is absorbed by downstream preprocessing and stereo model (Sec. 5).

3.4 Target Vergence

Directing both foveae onto a target is a 3D problem: the two telephoto axes must converge at the target depth, not merely

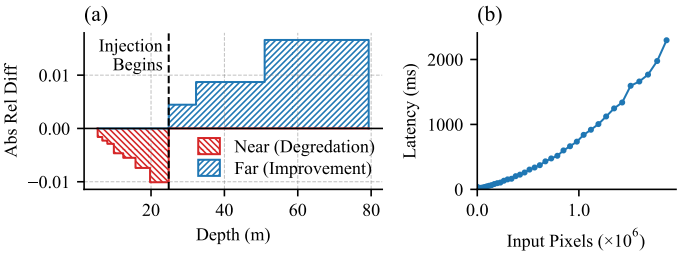


Fig. 5: **(a)** Depth-binned change in AbsRel ($\Delta = \text{FOV} - \text{MR}$) on V-Kitti2 [47]. Far regions (blue), where high-resolution patches are injected, show improved depth accuracy ($\Delta > 0$), while near regions show slight degradation ($\Delta < 0$) (red). **(b)** Demonstration of quadratically increasing latency as a function of input pixels in Depth-Anything-V2 [22]. While a monocular model, it is integrated as an intermediate step in many foundation stereo models [23], [24]. The quadratic relationship motivates splitting depth computations into multiple streams for transformer architectures.

point in a shared 2D direction. Given a region of interest on the wide-angle feed, we steer both foveae parallel to the central ray, then iteratively verge them via normalized cross-correlation template matching until the foveal views overlap on the target. This closed-loop scheme requires no explicit depth estimate and converges whenever the ROI contains matchable texture.

4 Theory and Analysis

Basic triangulation, $Z = \frac{fB}{d}$, shows that increased resolution improves disparity accuracy in foveated regions (Sec. 4.1). With this factor in mind, our research is driven by four intertwined observations:

1) Foveation can compensate for stereo depth uncertainty that grows quadratically with distance (Sec. 4.1, Fig. 4).
2) Choosing where to foveate allows reallocation of pixels from geometrically insensitive near regions to depth-critical far regions, yielding large gains in far-field accuracy with only minor degradation nearby (Sec. 4.2).
3) Distributing resolution across foveated regions reduces the number of globally interacting tokens, reducing quadratic cost of transformer stereo networks (Sec. 4.3).
4) Improved long-range perception expands the operational envelope for downstream robotics tasks such as target localization, navigation, and planning (Sec. 4.4).

4.1 Disparity advantage in triangulation

Depth error from stereo triangulation is well known to grow quadratically with distance: $|\delta Z| = \frac{Z^2}{Bf}|\delta d|$ [48]. *FoveaCam Duo* compensates for this by dynamically increasing the effective focal length at the region of interest, which proportionally increases the observed disparity and reduces depth uncertainty at range. We demonstrate such advantage with a real-world target localization experiment.

To evaluate the depth estimation accuracy of *FoveaCam Duo*, we designed a controlled target localization experiment using a small Statue of Liberty model mounted on a 3D-printed "∞"-shaped rail. The rail contains slots at predefined positions, providing ground-truth target locations relative to its geometric center, and is placed at standoff distances of 5 m, 10 m, and 15 m. As a baseline, we compare against a conventional wide-angle stereo pair with the same baseline and total field of view as the foveated system's steerable field of regard. For each stereo pair, the target is localized via normalized cross-correlation template matching, followed by sub-pixel refinement (up to $\frac{1}{256}$ px) to estimate disparity at the target center. Given rectified stereo geometry, depth is recovered via standard triangulation.

Fig. 4(c) summarizes the resulting depth errors. At 10 m, the wide-angle system degrades significantly as the target occupies fewer pixels and disparity compresses toward quantized values. In contrast, the foveated system maintains larger disparities due to its $\sim 9\times$ higher angular resolution, resulting in substantially lower error. At 15 m, wide-angle stereo fails entirely, with disparity approaching the noise floor and the reconstructed trajectory no longer resembling the ground-truth "∞" pattern. The foveated system continues to produce coherent depth estimates and successfully recovers the characteristic trajectory shape.

4.2 Depth-Dependent Tradeoff Under Equal Bandwidth

For reallocation analysis, we construct two configurations of FoundationStereo [24] on VKITTI2 [47] with equal total pixel count: uniform medium resolution (MR) and a low-resolution backdrop with 25% of pixels as high-resolution foveal patches (FOV). To quantify the tradeoff between near-field degradation and far-field improvement, we partition the depth range into percentile bins ordered by increasing ground-truth depth and compute metric differences between foveated (FOV) and medium-resolution (MR) configurations. For lower-is-better metrics, signs are flipped so that positive values consistently indicate improvement.

We summarize this behavior using a Near-Far Tradeoff Ratio $\mathcal{T}$, the relative magnitude of far-field improvement compared to near-field degradation (see supplementary Sec.2 for full definition). Values $\mathcal{T} > 1$ indicate that improvements in far regions outweigh losses in near regions. Across all evaluated metrics, we observe $\mathcal{T} > 1$ (AbsRel: 2.02, RMSE: 7.41, δ_1: 3.28), demonstrating that modest degradation in near regions corresponds to substantially larger gains in far regions. Per bin AbsRel is visualized in Fig. 5 (a). This supports reallocating resolution from geometrically insensitive near-field areas to depth-critical far-field regions, which *FoveaCam Duo* uniquely enables.

4.3 Token Reduction for *FoveaCam Duo*

Even under equal bandwidth, transformer-based stereo networks incur quadratic cost in global attention (Fig. 5 (b)): for number of tokens $N = \frac{HW}{P^2}$, the cost of a global self-attention layer scales as $\mathcal{C}(N) \propto N^2$, where P, H, and W are the patch size, image height, and image width in pixels. *FoveaCam Duo* splits computation into a low-resolution backdrop N_{low} and multiple high-resolution foveal streams N_i:

$$N_{\text{total}} = N_{\text{low}} + \sum_i N_i, \;\; \mathcal{C}_{\text{FOV}} \propto N_{\text{low}}^2 + \sum_i N_i^2 < \mathcal{C}_{\text{MR}}. \quad (5)$$

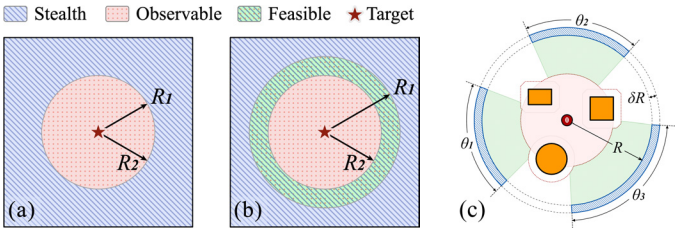


Fig. 6: Duality between visibility and stealth constraints. **(a)** Equivalent observation capabilities. the observer must either break stealthiness to observe the target, or fail to observe the target to retain stealth. **(b)** Larger observer detection range. An annular band exists where stealth and visibility constraints are both satisfied. **(c)** δR-based interpretation of feasible area with obstacles. The visible angular spans $\theta_1, \theta_2, \theta_3$ determine the retained sectors, yielding the same quadratic scaling in $D_{\max}^2 - D_{\min}^2$ after accumulation.

Quadratic convexity ensures independent streams reduce unnecessary global interactions, lowering attention complexity while reallocating resolution to depth-sensitive regions. This principle applies to transformer-based stereo architectures such as STTR [49], CREStereo [21], and IGEV-Stereo [20], and models that integrate transformer-based monocular depth estimators [23], [24].

4.4 Analysis of Reciprocal Visibility Constraints

FoveaCam Duo is inspired by the recurring strategy of crypsis in natural vision: survive by seeing first, seeing clearly, and seeing without being seen. Consider a predator in a mature ecosystem. Its behavior can be interpreted as an implicit optimization process balancing perception and exposure. These ecological pressures map directly to several geometric constraints. An observer (predator) that wishes to distinguish a feature of size S_t on some target i (prey). Given a peak angular resolution θ_T, the distance between the observer location A and the target location T_i must satisfy:

$$\begin{aligned} D_i &\triangleq \|A - T_i\|_2 < D_{\max}, \\ D_{\max} &\triangleq \frac{S_t}{2\tan(\frac{\theta_T}{2})} \approx \frac{S_t}{\theta_T}. \end{aligned} \tag{6}$$

In order to minimize vigilance of the target, the observer must keep a minimal distance $D_i > D_{\min}$. Additional environmental constraints on the observer's location given occluding obstacles and line-of-sight to multiple targets are formalized in Sec. 4 of the supplementary material and used in our simulation study.

4.4.1 Stealth v.s. Visibility Reciprocity

In the case when the target has the same or better detection capability than the observer, *i.e.*, the stand-off distance $D_{\min}$ is equal to or greater than the observer's max detection distance $D_{\max}$, there exists no feasible location for the observer to maintain stealth:

$$\begin{aligned} &D_{\max} \leq D_{\min} \;\Rightarrow \\ &\{A \mid D_{\min} < \|A - T\| < D_{\max}\} = \emptyset. \end{aligned} \tag{7}$$

This relation reveals a geometric duality between visibility capability and stealth: the visibility constraint imposes an *upper* bound on standoff distance ($D < D_{\max}$) to preserve resolvability, while stealth imposes a *lower* bound ($D > D_{\min}$) to avoid detection. Improving one objective tightens the other, and when $D_{\max} \leq D_{\min}$ the feasible interval collapses to the empty set.

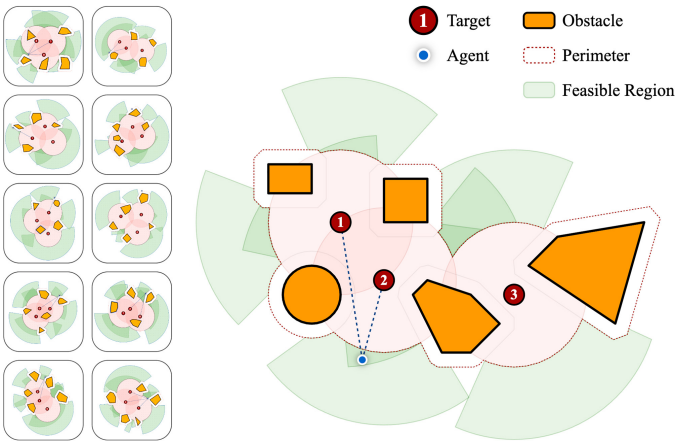


Fig. 7: A multi-target scene with a robotic observer located within the obstacle-aware feasible region of two targets. Line-of-sight rays from the agent to targets ❶ and ❷ visualize feasible observations, while interacting collision, occlusion, and stealth v.s. visibility constraints carve annular bands into disconnected sectors. The first map is full-size for detail, with additional maps as thumbnails.

Conversely, when the observer has a positive detection-range margin ($D_{\max} > D_{\min}$), a non-empty feasible region emerges:

$$\{A \mid D_{\min} < \|A - T\| < D_{\max}\} \neq \emptyset, \tag{8}$$

allowing for effective observation without violating stealthiness. As illustrated in Fig. 6, this region is the annular band between the two bounds $D_{\max} - D_{\min}$, and a larger margin directly expands the navigable feasible area. In a 2D context, navigable area $\mathcal{A}_{\text{nav}}$ grows quadratically with this detection-range margin: $\mathcal{A}_{\text{nav}} \propto (D_{\max}^2 - D_{\min}^2)$.

The above derivation treats $D_{\max}$ and $D_{\min}$ as hard boundaries, but in practice the observer's own depth estimate for D_i carries some uncertainty ΔZ_{near}, inflating the stealth bound to give the effective navigable area

$$\mathcal{A}_{\text{nav}} \propto D_{\max}^2 - (D_{\min} - \Delta Z_{\text{near}})^2. \tag{9}$$

By increasing angular resolution, *FoveaCam Duo* expands the annular band from *both* sides: $D_{\max}$ grows outward with increased range while a longer focal length shrinks $D_{\min}$ as ΔZ_{near} decreases. Traditional stereo is limited as it can only increase $D_{\max}$ at the expense of either field of view (telephoto lenses) or bandwidth (high resolution image sensors). A full derivation of ΔZ_{near} is given in Sec. 5 of the supplementary material.

4.4.2 Introduction of Obstacles

For a deterministic or known map, obstacles only attenuate the per-radius visible angular span and leave the quadratic scaling intact as shown in Fig. 6, with the proportionality constant absorbing obstacle complexity. The full derivation is given in Sec. 6 of the supplementary material.

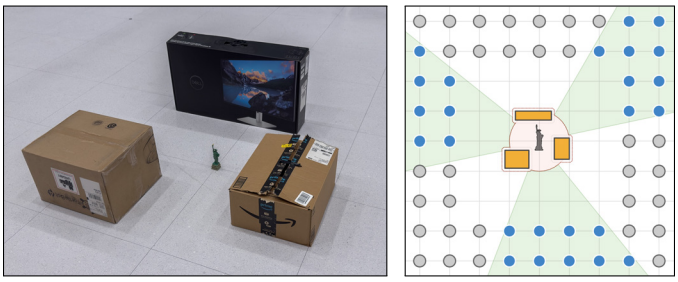


Fig. 8: Proof-of-concept real experiment using a toy scene. **Left:** physical scene with a target and obstacles on a tiled floor. **Right:** top-down schematic showing both simulation predictions (shaded green regions) and real-world observations (dots); floor-tile corners define a grid of agent positions verified in experiment. Blue dots denote successful observation, gray denotes failures. The validation results agree with the feasible area predicted by the simulator.

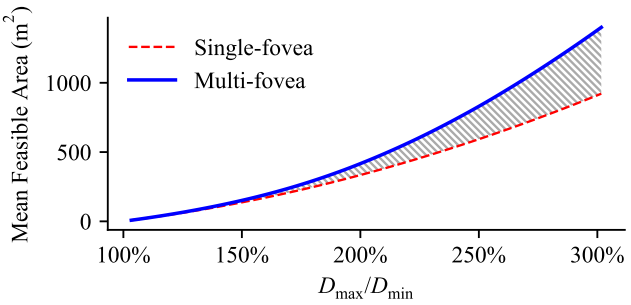


Fig. 9: Mean feasible navigable area as a function of detection-range ratio $D_{\max}/D_{\min}$ for multi-fovea and single-fovea configurations, averaged over 10 randomized maps. Both curves exhibit the predicted quadratic growth; the multi-fovea system's ability to track multiple targets yields a progressively larger feasible region (hatched gap).

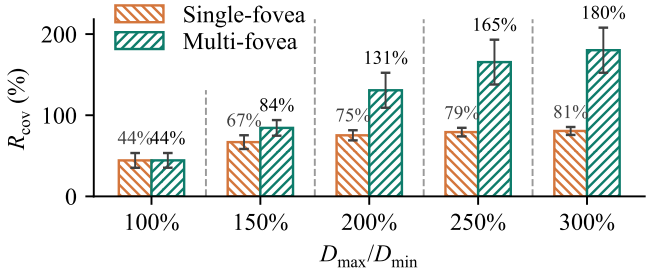


Fig. 10: Mean coverage rate R_{cov} over one circumnavigation loop for multi-fovea and single-fovea agents at varying detection-range ratios $D_{\max}/D_{\min}$, averaged over 10 randomized maps (error bars: $\pm 1\sigma$). The multi-fovea agent's ability to track multiple targets simultaneously yields substantially higher coverage, with the gap widening as the ratio increases.

To quantify how the reciprocal range margin translates into success rate in a stochastic or unknown environment, we compute the probability of a successful observation over a random agent location for a fixed target location. We model obstacle centers as a homogeneous 2D Poisson process [50], [51] with per-unit-area occupancy probability $p \in [0, 1]$. This results in an equivalent Poisson rate of λ is $\lambda = -\ln(1-p) \geq 0$. An obstacle of radius r_o blocks a sightline whenever its center falls within distance r_o of the segment $\overline{AT}$. For a sightline of length $d = \|A - T\|$, the swept blocking corridor has area

$$\mathcal{A}_{\text{blk}}(d) = 2r_o d + \pi r_o^2. \tag{10}$$

Under the Poisson assumption, the number of obstacle centers in this corridor is Poisson-distributed with mean $\lambda \mathcal{A}_{\text{blk}}(d)$, so the probability that no obstacle intersects $\overline{AT}$ is

$$\begin{aligned} P\left(\overline{AT} \text{ is clear}\right)\Big|_d &= P\left(\prod_0^d N(\mathcal{A}_{\text{blk}}(\delta d)) = 0\right) \\ &= e^{-\lambda \mathcal{A}_{\text{blk}}(d)} = (1-p)^{\mathcal{A}_{\text{blk}}(d)}. \end{aligned} \tag{11}$$

We now take A to be uniformly distributed over the map domain Ω with area $|\Omega|$ and define a successful observation as an agent placement that simultaneously satisfies both standoff bounds and has a clear line-of-sight to T. By averaging the joint event (in-range *and* unblocked) over Ω, converting to polar coordinates centered at T, we find the overall success probability as a function of fovea range $D_{\max}$:

$$P_{\text{succ}}(D_{\max}) = \frac{2\pi}{|\Omega|} \int_{D_{\min}}^{D_{\max}} (1-p)^{\mathcal{A}_{\text{blk}}(r)}\, r\, dr. \tag{12}$$

Finally, differentiating under the integral sign with respect to the upper limit $D_{\max}$ yields a strictly non-negative sensitivity, confirming that extending visibility range strictly improves the probability observation:

$$\frac{d\,P_{\text{succ}}}{dD_{\max}} = \frac{2\pi D_{\max}}{|\Omega|}(1-p)^{\mathcal{A}_{\text{blk}}(D_{\max})} \geq 0, \tag{13}$$

with the full derivation in Sec. 6 of the supplementary material. Even in the presence of randomly distributed obstacles, increasing the visibility range $D_{\max}$ monotonically increases the probability of successful observation. In the next section, we verify this property through simulation.

4.4.3 *Simulation-Based Quantitative Analysis*

As shown in Fig. 7, collision, occlusion, and line-of-sight constraints remove contiguous angular sectors, yielding disconnected and non-convex feasible regions. The robotic agent is coupled with explicit line-of-sight rays to targets 0 and 1, illustrating feasible observations under clutter while other directions are blocked by obstacles. To empirically validate this behavior, we run simulations across 10 randomly generated maps, each with independently sampled obstacle layouts and target positions, and report results averaged over all scene instances. The total navigable area where at least one target was reachable as a function of the maximum detection range $D_{\max}$ was averaged across the 10 randomized maps. As shown in Fig. 9, the relation consistently follows the quadratic dependence on detection-range margin derived in Sec. 4.4.1 and Sec. 4.4.2.

Next, we simulated a robotic agent as shown in Fig. 7. The agent circumnavigated the perimeter constrained by the resolution, stealthiness, and line-of-sight constraints defined in Sec. 4.4. The coverage rate R_{cov} is defined as the integral

of the number of reachable targets along one loop:

$$R_{\text{cov}} = \oint N\big(A(t)\big)\, dt \tag{14}$$

where $N\big(A(t)\big)$ is the number of targets satisfying all visibility constraints from the agent location $A(t)$ at normalized time $t \in [0, 1]$. For a multi-fovea agent N counts all simultaneously visible targets, while for a single-fovea agent $N \in \{0, 1\}$. We repeated the loop simulation on each of the 10 randomized maps and report the mean and standard deviation in Fig. 10.

To validate the feasibility analysis against physical reality, we reproduced a scaled-down scene with a target and a small set of obstacles on a tiled floor (Fig. 8), using floor-tile corners as a predefined grid of agent positions. At each grid point, *FoveaCam Duo* recorded whether the target was observable under the same constraints used in simulation. The successful (blue) and failed (gray) observations align with the simulator's predicted feasible area (shaded green), confirming that the simulation faithfully reflects real-world geometry.

5 Stereo Signal Validation via Self-Supervised Reconstruction

Convergent stereo geometry introduces a correspondence structure that differs fundamentally from the rectified, wide-baseline settings assumed by modern learned stereo systems. In particular, convergent telephoto imaging produces signed, view-dependent disparities arising from non-parallel optical axes, violating the epipolar consistency and monotonic disparity assumptions exploited by foundation stereo models [20], [23], [24]. In contrast, classical multiview stereo methods are typically designed for wide-baseline capture with strong viewpoint diversity, where sufficient parallax enables robust correspondence aggregation [52]. In telephoto convergent configurations, however, viewpoints become tightly clustered with weak effective parallax, yielding disparity signals that fall outside the regimes these methods are designed to handle.

To ensure the integrity of stereo disparity in a domain where existing methods are not applicable, we implement a self-supervised transformer-based stereo framework. We decouple correspondence learning from imaging geometry by operating on rectified virtual pinhole views, enabling consistent epipolar reasoning despite non-parallel cameras. Self-supervision in combination with a foveated dataset allows us to produce stable correspondence signals where standard stereo and multiview methods degrade.

5.1 FoveaCam Duo Dataset

To support research on foveated stereo and high-resolution depth estimation, we release a raster-scanned dataset captured with *FoveaCam Duo* containing approximately 2K foveated stereo image pairs. For each scene, we steer the foveated cameras through a dense grid of pointing angles, with the left and right foveas converged at the approximate average scene depth. At each grid position, we capture synchronized wide-angle and foveated image triplets, resulting in dense angular coverage of each scene at both wide-angle and telephoto resolutions.

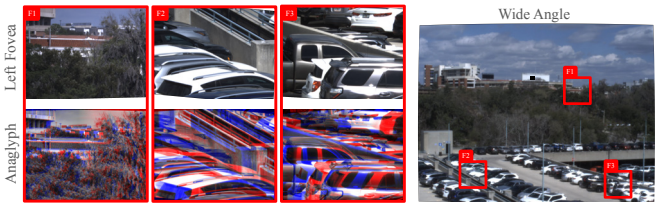


Fig. 11: Visualization of one collected scene of the rasterized dataset. For each scene, a wide-angle shot is captured alongside 81 pairs of foveated images. Here we show the left fovea image capture, the anaglyph between the left and right fovea image, and the wide-angle capture.

A subset of scenes includes multi-viewpoint captures. In these sequences, we mount *FoveaCam Duo* on a custom portable linear rail and translate the entire camera system along the x axis to five positions (-300 mm, -100 mm, center, $+100$ mm, $+300$ mm). At each rail position, the full raster scan is repeated, yielding five complete scans per scene. This dataset is structured to enable controlled evaluation of foveated stereo geometry and may serve as a basis for deriving depth estimates via wide-baseline triangulation. Example scenes are shown in Fig. 11.

5.2 Architecture

We adopt a transformer-based stereo architecture adapted to the rectified foveated setting. The left and right images are first rectified and roughly aligned to reduce large global offsets. Features are extracted using a pretrained encoder, and a monocular depth prior is computed from the left image. These representations are fused using a transformer with bidirectional attention between left and right features, and gated conditioning on the monocular prior and camera geometry. The final representation is decoded into a disparity prediction. Further implementation details are provided in the supplementary.

5.3 Training Objective

The model is trained using a combination of photometric reprojection, monocular prior, and smoothness losses:

$$\mathcal{L} = \lambda_{\text{photo}}\, \mathcal{L}_{\text{photo}} + \lambda_{\text{prior}}(t)\, \mathcal{L}_{\text{prior}} + \lambda_{\text{smooth}}\, \mathcal{L}_{\text{smooth}}. \tag{15}$$

The photometric term enforces consistency under 3D reprojection, while the prior term provides scale and shift-invariant supervision from a frozen monocular depth model. The prior weight is annealed to zero during training, encouraging the model to rely on stereo cues. This formulation is intended to promote geometrically consistent correspondences rather than optimize for absolute metric accuracy.

5.4 Training and Evaluation

The model is trained in a self-supervised manner using the objective above. We evaluate performance using photometric reprojection error, which serves as a proxy for correspondence quality. We compare against a foveated monocular baseline (Depth-Anything-V2) and fine-tuned stereo models. Monocular predictions are aligned per scene

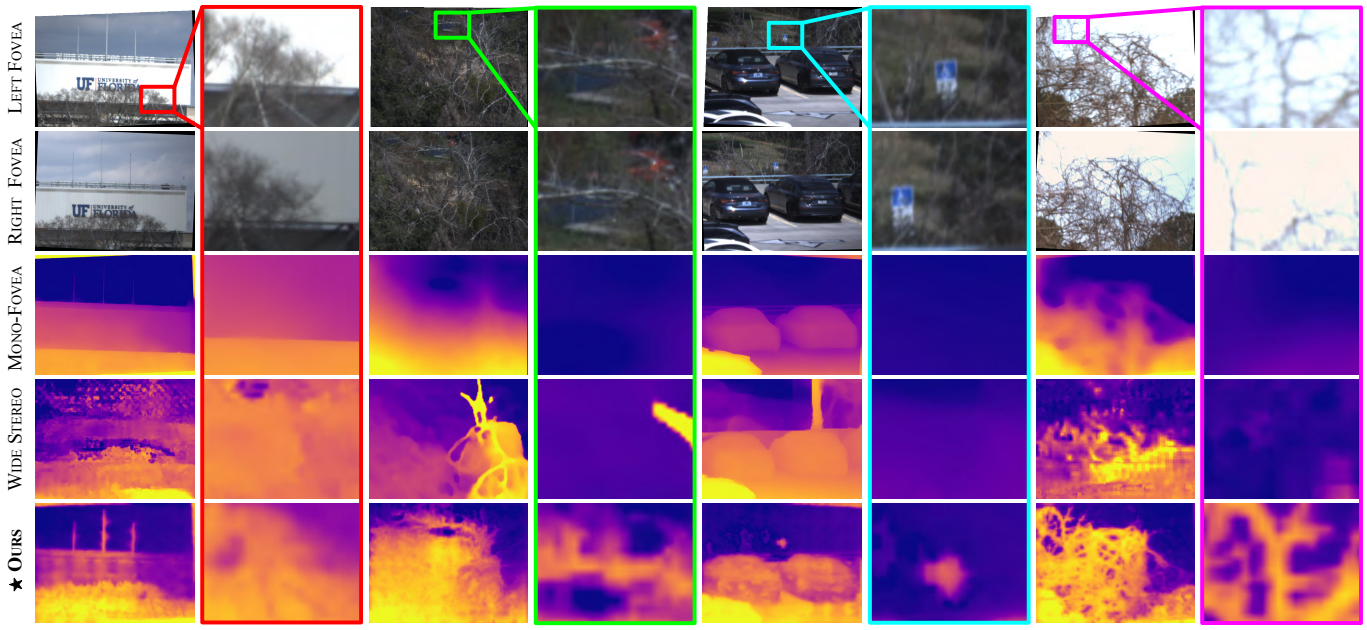


Fig. 12: Comparison between monocular prior [22], wide-angle traditional stereo stereo [24], and our self-supervised stereo method. Top two rows show left and right foveal crops, the third row shows the aligned monocular depth prior [22], the fourth row wide-angle foundation stereo [24], and the bottom row shows our stereo predictions. Despite having no access to full monocular training data, our method leverages stereo cues to recover high-frequency scene details, sharper boundaries, and more accurate depth variation, particularly in regions with thin structures, occlusions, and low texture, where the monocular prior tends to oversmooth. Although noisier than comparisons due to a lack of priors, our unsupervised method locks on to real stereo features in the world demonstrating the integrity of stereo correspondence.

using scale and shift. As shown in Table 3, our method reduces reprojection error relative to both monocular and adapted stereo baselines. We also ablate against different implementations of the monocular loss, showing the value of annealing over training and the model's lack of reliance on the prior alone.

Notably, fine-tuned IGEV fails to produce stable outputs, indicating that the limitation arises from geometric mismatch rather than model capacity. Qualitative results (Fig. 12) show that our method recovers fine-scale structure that is oversmoothed or missing in monocular predictions, despite having no access to large-scale supervised data.

These results suggest that meaningful stereo cues are preserved in the foveated convergent setting when the learning formulation accounts for its geometry.

While photometric reprojection error is used as our primary metric, this evaluation serves as a proxy for correspondence quality rather than a definitive measure of metric depth accuracy. To provide additional validation, we perform a limited metric comparison on a subset of scenes using wide-baseline stereo reconstruction as reference, observing consistent depth alignment (see supplementary). These results suggest that the recovered disparities are metrically meaningful, though comprehensive ground-truth evaluation is left for future work.

TABLE 3: Quantitative comparison using mean photometric reprojection error.

Method	Mean Photo. Error ↓
Mono-Fovea (aligned)	0.0410
Fine Tuned IGEV	0.0415
Ours (full strength prior loss)	0.0856
Ours (no prior loss)	0.0480
Ours	**0.0354**

6 Conclusion

We presented *FoveaCam Duo*, a foveated stereo camera system that overcomes the fundamental trade-off between angular resolution and field of view in conventional stereo. Our three-stage calibration pipeline ensures parallel epipolar lines, facilitating the detection of stereo cues. Localization experiments demonstrate that *FoveaCam Duo* maintains accurate at distances which wide-angle stereo completely fails. We also produce a novel architecture for foveated stereo trained using self-supervision on real world data, improving depth quality over wide-angle stereo. We showed that conventional stereo pipelines are fundamentally mismatched to foveated imagery, and developed a dedicated foveation-aware approach to achieve accurate long-range depth.

Acknowledgments

We thank the National Science Foundation and the Office of Naval Research for their support and the authors were partially supported by NSF grants 1942444 and 2330416 as well as ONR grants N000142312429 and N000142312363.

REFERENCES

[1] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel RealSense stereoscopic depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1267–1276.

[2] M. Hansard, S. Lee, O. Choi, and R. Horaud, *Time-of-Flight Cameras: Principles, Methods and Applications*. Springer, 2013.

[3] R. Halterman and M. Bruch, "Velodyne HDL-64E LiDAR for intelligent vehicle applications," in *SPIE Defense, Security, and Sensing*, vol. 7707, 2010, pp. 103–112.

[4] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 119–155, 2003.

[5] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: An open platform for research in embodied cognition," *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems (PerMIS)*, pp. 50–56, 2008.

[6] M. Björkman and D. Kragic, "Active 3D scene exploration with foveated vision," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2002.

[7] D. Coombs and C. Brown, "Real-time binocular smooth pursuit," *International Journal of Computer Vision*, vol. 11, no. 2, pp. 147–164, 1993.

[8] M. Chi, "Convergent active stereo: Investigating depth perception with a robotic binocular camera system," Master's thesis, York University, Toronto, Ontario, Canada, 2025.

[9] D. J. Brady, M. E. Gehm, R. A. Stack, D. L. Marks, D. S. Kittle, D. R. Golish, E. M. Vera, and S. D. Feller, "Multiscale gigapixel photography," *Nature*, vol. 486, pp. 386–389, 2012.

[10] M. Stevens and S. Merilaita, "Animal camouflage: Current issues and new perspectives," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, pp. 423–427, 2009.

[11] J. Endler, "An overview of the relationships between mimicry and crypsis," *Biological Journal of the Linnean Society*, vol. 16, pp. 25 – 31, 06 2008.

[12] M. F. Land and D.-E. Nilsson, *Animal Eyes*. Oxford University Press, 2012.

[13] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[14] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 2005, pp. 807–814 vol. 2.

[15] ——, "Stereo processing by semiglobal matching and mutual information," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'08)*, vol. 30, no. 2, 2008, pp. 328–341.

[16] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1–3, p. 7==42, 2002.

[17] N. Mayer, E. Ilg, P. Häusser, 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.

[18] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

[19] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in *International Conference on 3D Vision (3DV)*, 2021, pp. 218–227.

[20] 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.

[21] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," 2022.

[22] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," 2024. [Online]. Available: https://arxiv.org/abs/2406.09414

[23] J. Cheng, W. Liao, Z. Cai, L. Liu, G. Xu, X. Wang, Y. Wang, Z. Yuan, Y. Deng, J. Zang, Y. Shi, J. Tang, and X. Yang, "Monster++: Unified stereo matching, multi-view stereo, and real-time stereo with monodepth priors," 2025.

[24] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," 2025.

[25] E. L. Schwartz, "Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception," *Biological Cybernetics*, vol. 25, no. 4, pp. 181–194, 1977.

[26] H. Xiao, J. Ackermann, B. Deng, and G. Wetzstein, "Policy-based foveated imaging and perception," 2026. [Online]. Available: https://arxiv.org/abs/2606.02565

[27] A. L. Yarbus, *Eye Movements and Vision*. Springer New York, NY, 1967.

[28] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[29] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," in *International Journal of Computer Vision*, vol. 1, no. 4, 1988, pp. 333–356.

[30] S. R. Chettri, M. Keefe, and J. R. Zimmerman, "Stereo pair design for cameras with a fovea," in *Intelligent Robots and Computer Vision X: Neural, Biological, and 3-D Methods*, D. P. Casasent, Ed., vol. 1608, International Society for Optics and Photonics. SPIE, 1992, pp. 358 – 365. [Online]. Available: https://doi.org/10.1117/12.135102

[31] D. Wan and J. Zhou, "Stereo vision using two ptz cameras," *Computer Vision and Image Understanding*, vol. 112, no. 2, pp. 184–194, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314208000313

[32] K. Okumura, H. Oku, and M. Ishikawa, "High-speed gaze controller for millisecond-order pan/tilt camera," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 6186–6191.

[33] V. Milanović, G. A. Matus, and D. T. McCormick, "Gimbal-less monolithic silicon actuators for tip–tilt–piston micromirror applications," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 10, no. 3, pp. 462–471, 2004.

[34] A. D. Yalcinkaya, H. Urey, D. Brown, T. Montague, and R. Sprague, "Two-axis electromagnetic microscanner for high resolution displays," *IEEE Journal of Microelectromechanical Systems*, vol. 15, no. 4, pp. 786–794, 2006.

[35] B. Tilmon, E. Jain, S. Ferrari, and S. Koppal, "Foveacam: A mems mirror-enabled foveating camera," 04 2020, pp. 1–11.

[36] B. Tilmon and S. J. Koppal, "Saccadecam: Adaptive visual attention for monocular depth sensing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6009–6018.

[37] S. Liu, C. Pansing, and H. Hua, "Design of a foveated imaging system using a two-axis mems mirror," in *International Optical Design*. Optica Publishing Group, 2006, p. ME25. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=IODC-2006-ME25

[38] H. Hua and S. Liu, "Dual-sensor foveated imaging system," *Appl. Opt.*, vol. 47, no. 3, pp. 317–327, Jan 2008. [Online]. Available: https://opg.optica.org/ao/abstract.cfm?URI=ao-47-3-317

[39] S. Mao, Y. N. Mishra, and W. Heidrich, "Fovea stacking: Imaging with dynamic localized aberration correction," *ACM Trans. Graph.*, vol. 44, no. 6, Dec. 2025. [Online]. Available: https://doi.org/10.1145/3763278

[40] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *ACM SIGGRAPH*, 2005, pp. 765–776.

[41] Y. Zhang and S. J. Koppal, "Foveacam++: Systems-level advances for long range multi-object high-resolution tracking," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 11 594–11 601.

[42] F. Pittaluga, Z. Tasneem, J. Folden, B. Tilmon, A. Chakrabarti, and S. J. Koppal, "Towards a mems-based adaptive lidar," *arXiv preprint arXiv:2003.09545*, 2020.

[43] A. Swingler and S. Ferrari, "On the duality of robot and sensor path planning," in *52nd IEEE Conference on Decision and Control*, 2013, pp. 984–989.

[44] W. Lu, G. Zhang, and S. Ferrari, "A randomized hybrid system approach to coordinated robotic sensor planning," in *49th IEEE Conference on Decision and Control (CDC)*, 2010, pp. 3857–3864.

[45] H. Wei, W. Lu, P. Zhu, G. Huang, J. Leonard, and S. Ferrari, "Optimized visibility motion planning for target tracking and localization," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 76–82.

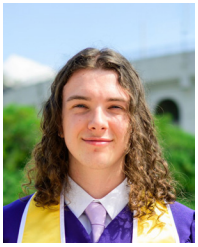
[46] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[47] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," 2020.

[48] B. Horn, *Robot vision*. MIT press, 1986.

[49] 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 (ICCV)*, October 2021, pp. 6197–6206.
[50] J. F. C. Kingman, *Poisson Processes*, ser. Oxford Studies in Probability. Oxford University Press, 1993, vol. 3.
[51] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. Wiley, 1995.
[52] W. Jang, P. Weinzaepfel, V. Leroy, L. Agapito, and J. Revaud, "Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors," 2025. [Online]. Available: https://arxiv.org/abs/2503.17316

Yuxuan Zhang is pursuing his Ph.D. in Electrical and Computer Engineering at the University of Florida, after earning his Bachelor's in Physics from Xi'an Jiao Tong University. He builds end-to-end intelligent robotic systems spanning Machine Learning, Software, Electrical, and Mechanical Engineering. His current research focuses on foveated vision systems, vergence stereo algorithms, and their applications in robotics.



Jacob Carter graduated from Louisiana State University in 2024 with a B.S. in Computer Science. He is currently pursuing his Ph.D. in Electrical and Computer Engineering at the University of Florida. He is interested in AI in computer vision, specifically depth estimation and rolling shutter correction.

Hannah Kirkland is a graduate research assistant in the FOCUS Lab. They graduated with a B.S in Electrical Engineering from UF in 2021. Their interests include computational photography, unconventional computing, and computer vision.

Michael Tomadakis is a 2024 Computer Science graduate from the University of Florida. He's been passionate about graphics since discovering shaders in Unreal Engine 3 twelve years ago, but his interests have matured into Computer Vision and Neural Rendering. He is working on NeRF stitching and adjacent techniques.

Noah Ralph is an undergraduate student at the University of Florida majoring in Electrical Engineering. His interests include 3D printing, Biomechanics, Robotics, and Medical Devices.

Sanjeev J. Koppal is an Associate Professor at the University of Florida's Electrical and Computer Engineering Department, where he leads the FOCUS Lab. He obtained his B.S. degree from the University of Souther California in 2003 as a Trustee Scholar, and his Masters and Ph.D. degrees from the Robotics Institute at Carnegie Mellon University. His interests span computer vision, computational photography and optics, novel cameras and sensors, 3D reconstruction, physics-based vision, and active illumination