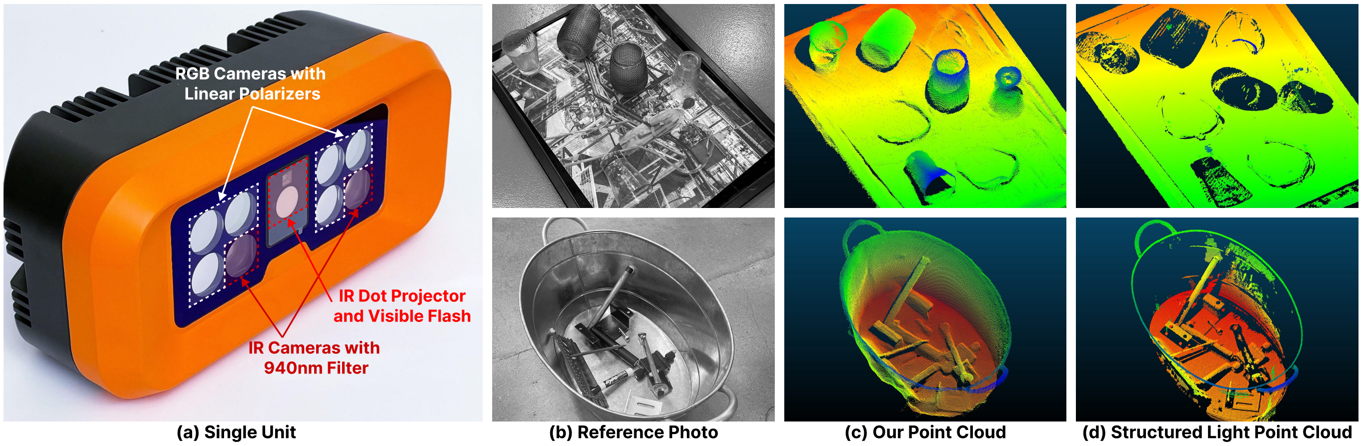
A Plenoptic 3D Vision System

AGASTYA KALRA, Intrinsic Innovation LLC, USA
VAGE TAAMAZYAN, Intrinsic Innovation LLC, USA
ALBERTO DALL'OLIO, Intrinsic Switzerland GmbH, Switzerland
RAGHAV KHANNA, Intrinsic Switzerland GmbH, Switzerland
TOMAS GERLICH, Intrinsic Innovation LLC, USA
GEORGIA GIANNOPOULOU, Intrinsic Switzerland GmbH, Switzerland
GUY STOPPI, Intrinsic Canada Corporation, Canada
DANIEL BAXTER, Intrinsic Innovation LLC, USA
ABHIJIT GHOSH, Intrinsic Innovation LLC, USA
RICK SZELISKI, Google DeepMind, USA
KARTIK VENKATARAMAN, Intrinsic Innovation LLC, USA


RGB Cameras with Linear Polarizers
IR Dot Projector and Visible Flash
IR Cameras with 940nm Filter

(a) Single Unit
(b) Reference Photo
(c) Our Point Cloud
(d) Structured Light Point Cloud

Fig. 1. **Our compact unit design enables the simultaneous capture of rich multi-modal data - including RGB, IR, and polarization - enabling robust and accurate 3D reconstruction.** (a) - The plenoptic stereo vision unit's component layout (see Section 3.1 for details). (b) - A reference photo captured by a cell phone camera of two challenging scenes: transparent objects on a textured background (row 1) and a metallic deep bin (row 2). (c) - Our system's point clouds, which successfully reconstructs the transparent objects and the bin walls. (d) - The baseline point clouds captured by an industrial structured light 3D sensor, which struggles to reconstruct transparent objects or the bin walls.

We present a novel multi-camera, multi-modal vision system designed for industrial robotics applications. The system generates high-quality 3D point clouds, with a focus on improving the completeness and reducing hallucinations for collision avoidance across various geometries, materials, and lighting conditions. Our system incorporates several key advancements: (1) a modular and scalable **Plenoptic Stereo Vision Unit** that captures high-resolution RGB, polarization, and infrared (IR) data for enhanced scene understanding; (2) an **Auto-Calibration Routine** that enables the seamless addition and automatic registration of multiple stereo units, expanding the system's capabilities; (3) a **Deep Fusion Stereo Architecture** - a state-of-the-art deep learning architecture trained fully on synthetic data that effectively fuses multi-baseline and multi-modal data for superior reconstruction accuracy. We demonstrate the impact of each design decision through rigorous testing, showing improved performance across varying lighting,

Authors' addresses: Agastya Kalra, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, agastyak@intrinsic.ai; Vage Taamazyan, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, vage@intrinsic.ai; Alberto Dall'olio, Intrinsic Switzerland GmbH, Schärenmoosstrasse 77, Zürich, 8052, Switzerland, dallolio@intrinsic.ai; Raghav Khanna, Intrinsic Switzerland GmbH, Schärenmoosstrasse 77, Zürich, 8052, Switzerland, raghavkhanna@intrinsic.ai; Tomas Gerlich, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, tgerlich@intrinsic.ai; Georgia Giannopoulou, Intrinsic Switzerland GmbH, Schärenmoosstrasse 77, Zürich, 8052, Switzerland, giannopolou@intrinsic.ai; Guy Stoppi, Intrinsic Canada Corporation, 111 Richmond St W, Toronto, M5H2G4, Canada, guystoppi@intrinsic.ai; Daniel Baxter, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, danielbaxter@intrinsic.ai; Abhijit Ghosh, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, ghoshabhijit@intrinsic.ai; Rick Szeliski, Google DeepMind, 100 Mayfield Ave., Mountain View, 94043, USA, szeliski@google.com; Kartik Venkataraman, Intrinsic Innovation LLC, 100 Mayfield Ave., Mountain View, 94043, USA, kartikvp@intrinsic.ai.

0730-0301/2024/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

geometry, and material challenges. To benchmark our system, we create an extensive industrial-robotics inspired dataset featuring sub-millimeter accurate ground truth 3D reconstructions of scenes with challenging elements such as sunlight, deep bins, transparency, reflective surfaces, and thin objects. Our system surpasses the performance of state-of-the-art high-resolution structured light on this dataset. We also demonstrate generalization to non-robotics polarization datasets. Interactive visualizations and videos are available at https://www.intrinsic.ai/publications/siggraphasia2024.

CCS Concepts: • **Computing methodologies → 3D imaging**; **Camera calibration**; • **Hardware → Displays and imagers**.

Additional Key Words and Phrases: Autocalibration, Plenoptic Imaging, Point Clouds, 3D Reconstruction, Deep Stereo, Sim2Real, Robotics

ACM Reference Format:
Agastya Kalra, Vage Taamazyan, Alberto Dall'olio, Raghav Khanna, Tomas Gerlich, Georgia Giannopoulou, Guy Stoppi, Daniel Baxter, Abhijit Ghosh, Rick Szeliski, and Kartik Venkataraman. 2024. A Plenoptic 3D Vision System. *ACM Trans. Graph.* 1, 1 (September 2024), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Industrial robots operate in demanding environments where reliable perception is crucial for safety and success. Consider a scenario where a vision system incorrectly interprets a thin wire, leading to a robot collision that damages expensive equipment and disrupts production – a costly setback. To address these challenges, we must design vision systems that can handle the complexities of industrial settings. The goal is to achieve accurate and robust 3D reconstruction across complex geometries, environments, and materials in industrial settings for the purposes of robotic manipulation and collision avoidance.

No vision system currently solves this challenging problem due to the following complex requirements:

- **Geometric Complexity:** The system must reconstruct typical objects in a robotic work cell, which can range from thin wires that are hard to see through to tall bins that create occlusions and cause inter-reflections.
- **Working Volume:** Industrial robots require large clearances and so the system must be capable of accurate 3D reconstructions at distances of up to 3m.
- **Environment:** A system must hold calibration in a factory that has constant mechanical and thermal perturbations.
- **Lighting:** The system must continue to work even under direct sunlight (ex: a workcell directly under a skylights) and in low-light (ex: *lights-out* manufacturing), making reliance on active illumination alone impractical.
- **Materials:** The system must handle a variety of materials from diffuse metals to anisotropic and transparent materials.
- **Accuracy:** The system must be accurate enough for the robot's safe trajectory planning by avoiding collisions.

The current industry standard solution is high-resolution gray-code phase-shift structured light [Su and Zhang 2010; Zhonghe et al. 2022]. These systems project multiple patterns to establish precise correspondence between camera and projector pixels, achieving millimeter-level accuracy at distances of 2-2.5 meters. However, they excel primarily in ideal conditions with favorable geometry, lighting, and materials. In scenarios involving occlusions, anisotropic materials, or sunlight, these systems may fail to reconstruct objects.

Our system tackles this challenge with a stereo-based deep learning approach. Unlike structured light, stereo vision doesn't solely rely on active illumination, making it more robust to difficult lighting and material conditions. However, stereo vision typically lacks the precision needed for industrial robotics. To achieve the required accuracy in industrial working volumes, a large baseline between cameras is necessary. This poses significant challenges in both software (finding correspondences across vast disparity ranges) and hardware (maintaining calibration over large baselines).

To overcome the challenges of large-baseline stereo matching, we employ a traditional stereo-inspired multi-baseline approach [Okutomi and Kanade 1991], incorporating both small and large baselines to enhance correspondence quality. However, adding more cameras exacerbates calibration difficulties. To address this we utilize another key insight: if cameras could continuously self-calibrate, we could freely add more units as needed. Therefore, we integrate controllable IR dots with an IR stereo pair and develop a novel IR-dot based auto-calibration system that rivals checkerboard pattern accuracy without an external target.

Additionally, we incorporate polarization into the stereo pair to expand the camera's ability to handle challenging materials and use high dynamic range (HDR) capture to address difficult lighting conditions. We tie this together in a state-of-the-art deep plenoptic stereo architecture (*P-Stereo*) that combines multiple modalities and baselines, trained exclusively synthetically to prevent over-fitting.

Our system demonstrates competitive performance with structured light on ideal objects and superior performance in challenging lighting, material, and geometry scenarios. In addition, we show generalization on outdoor polarization datasets whereas our *P-Stereo* is trained purely in the synthetic domain, thereby showcasing effective sim-2-real transfer capabilities.

The rest of our paper is structured as follows. Section 2 describes the background and the related work. Section 3 describes the hardware design of a single-unit, and the software processing that occurs on each unit, and auto-calibration of multiple units. Section 4 describes 3D reconstruction using *P-Stereo* and the synthetic data training pipeline. Section 5 describes the results. Finally Section 6 describes future work.

2 BACKGROUND

Metrics for Collision Avoidance. Several established point cloud evaluation metrics exist, including Chamfer distance [Barrow et al. 1977; Fan et al. 2017], Hausdorff distance [Aspert et al. 2002; Dubuisson and Jain 1994], and Completeness and Accuracy [Knapitsch et al. 2017]. However, these metrics do not consider the requirements of robotics (e.g. missing the rim of a bin in a point cloud is negligible for completeness and accuracy, but will have serious consequences for collision avoidance). Therefore we use the Collision Avoidance Metric (CAM) [Taamazyan et al. 2024]. CAM simulates approximate robot trajectories and estimates collisions using the predicted point cloud and compares against estimated collisions using the ground truth (GT) point cloud. This leads to the metrics we use here:

False Negative Rate (FNR, %): If the GT detects a collision earlier in the trajectory than the estimated point cloud, this creates a *false negative collision* (i.e. a missed collision). These impact safety as it can cause damage to the robot and its environment. FNR improvement is usually correlated with an improvement in completeness.

False Positive Rate (FPR, %): If the GT point cloud detects a collision later in the trajectory than the estimated point cloud, this creates a *false positive collision* (i.e. a ghost collision). These affect efficiency, not safety, and are therefore less important. An improvement in FPR is usually correlated with a reduction in hallucinations.

Time-of-Flight. Besides structured light and stereo, time-of-flight cameras also reconstruct depth. While they do not suffer from occlusions, they instead suffer from lower spatial resolution, ambient light and geometric interference. While promising research ideas exist to improve them [Bamji et al. 2022; Horaud et al. 2016], existing systems would face many challenges in a factory environment.

Deep Stereo Networks. Deep learning has become the dominant approach for stereo matching since the pioneering work of [Zbontar and LeCun 2015]. Recent methods can be broadly categorized into cost volume filtering and iterative refinement.

Cost Volume Filtering: These methods typically employ 3D convolutions to regularize the cost volume, as seen in successful architectures such as GA-Net [Zhang et al. 2019], GCNET [Kendall et al. 2017], PSMNet [Chang and Chen 2018], AANet [Xu and Zhang 2020], and others [Duggal et al. 2019; Guo et al. 2019; Mayer et al. 2016; Shen et al. 2021].

Iterative Refinement: This more recent approach achieves state-of-the-art results by employing recurrent units for learned optimization and iterative refinement of the disparity map. This concept, introduced in RAFT-Stereo [Lipson et al. 2021] as an extension of the RAFT optical flow method [Teed and Deng 2020], has been further developed in CREStereo [Li et al. 2022], IGEV-Stereo [Xu et al. 2023], and DLNR [Zhao et al. 2023]. Our deep plenoptic stereo method is also based on an iterative refinement approach, extending CREStereo [Li et al. 2022] and RAFT [Lipson et al. 2021] by adding support for very large disparities and multi-modal inputs.

Multi-View Stereo. These have also been dominated by deep learning based approaches recently, with notable approaches such as MVSNet [Yao et al. 2018], SurfaceNet [Ji et al. 2017], P-MVSNet [Luo et al. 2019], PatchmatchNet [Wang et al. 2021], IterMVS [Wang et al. 2022], and others [Gu et al. 2020; Xu and Tao 2020; Yang et al. 2020; Yao et al. 2019]. These methods typically involve cost volume computation and fusion with respect to a reference camera with all images of the same modality. Our approach distinguishes itself by enabling the fusion of images from different modalities. A multi-baseline fusion method is proposed in TriStereoNet[Shamsafar and Zell 2021], however, it is fixed with 2 small baselines.

Auto-calibrated camera systems Automatic calibration of camera systems [Faugeras et al. 1992] is defined as determining the transform between two camera coordinate systems without an explicit calibration target. It has been demonstrated successfully in SfM [Schönberger and Frahm 2016] and self-driving applications [Hogan et al. 2023]. There have been some attempts to use it for stereo calibration [Marko and Kubinger 2018]. However, here the scale factor was provided by manually measuring the scene with a tape-measure. While prior approaches have experimented with IR dots [Deetjen and Lentink 2018], these have required multiple shots with multiple cameras and projectors with a specific scene setup.

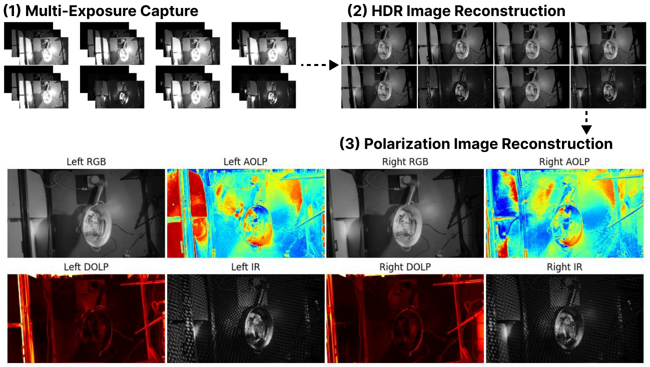

(1) Multi-Exposure Capture
(2) HDR Image Reconstruction
(3) Polarization Image Reconstruction
Left RGB
Left AOLP
Right RGB
Right AOLP
Left DOLP
Left IR
Right DOLP
Right IR

Fig. 2. **Our data capture and processing pipeline.** The processing pipeline (Section 3.2) transforms captured raw data into usable signals for advanced 3D reconstruction and analysis tasks.

Plenoptic Imaging The plenoptic function describes the complete space of light rays within a light field. In this work, we employ an eight-dimensional representation of this function:

$$I(x, y, \theta_x, \theta_y, \rho, \phi, \lambda, t), \tag{1}$$

Where I is the measured intensity of light, x, y as spatial resolution, θ_x, θ_y as multiple viewpoints, ρ, ϕ represent the degree of linear polarization (DOLP) and angle of linear polarization (AOLP), λ represents wavelength dependency, and t represents the time dimension. Our imaging system aims to fully sample this function to maximize point cloud quality.

Prior research has shown that polarization information can enhance performance on tasks involving transparent and highly reflective objects, including segmentation [Kalra et al. 2020; Mei et al. 2022], depth estimation and refinement [Ikemura et al. 2024; Kadambi et al. 2015; Zhu and Smith 2019]. Additionally, deep polarized stereo and multi-view stereo methods have been explored [Cui et al. 2017; Fukao et al. 2021; Huang et al. 2023; Tian et al. 2023]. However, these methods are typically based on physics-based heuristics or trained on real-world data similar to the test set, limiting their generalizability. In contrast, we demonstrate performance improvements using models trained entirely on synthetic data.

3 PLENOPTIC DATA CAPTURE

3.1 Plenoptic Stereo Vision Unit Hardware

Our hardware contains of the following imaging components as shown in Figure 1(a):

- **Multi-Aperture Polarization:** Each side of the stereo pair contains four 8MP cameras in a square, three RGB, one IR. Each RGB sensor is behind a polarizing filter rotated at three different angles: $0^o, 60^o, 120^o$. These three allow the capture of high-resolution polarization images simultaneously.

- **IR Active Stereo:** There is a dot projector in the middle of the camera and two 940nm IR sensors on each side. These form the active stereo pair used for increasing texture.
- **Flash:** There is also a visible spectrum flash in the middle of the camera to illuminate low-lighting conditions.

All of the in-unit intrinsics and extrinsics are calibrated using a checkerboard pattern and 75 images ahead of time.

3.1.1 Camera Architecture - Motivation. The rationale behind the hardware architecture comes down to the requirements of needing to capture linearly polarized images that will enable computation of the angle and degree of linear polarization (AOLP, DOLP). There exists significant prior work in exploiting polarization signals for many computer vision tasks. Our work aims to exploit this additional information in enabling more robust depth computation. Images captured with a standard imager fitted with a linear polarization filter show a sinusoidal relationship between the pixel intensity at a given location and the rotation of the linear polarization filter. The phase of this sinusoidal pattern encodes the azimuth angle of the surface normal at that point and the amplitude and offset encode the zenith angle. To recover the three parameters of this sinusoidal relationship, we need at least three observations of the scene captured with a linear polarizing filter at 3 different angles sufficiently apart from each other in spatial rotation [Atkinson and Hancock 2006]. Hence, we include three RGB imagers with the linear polarization filters rotated by 60° from each other (i.e. 0°, 60°, and 120°).

A fourth imager is included to sample the scene in Near-IR spectrum at 940 *nm*. We choose 940nm on account of it's absence in sunlight due to absorption by the water vapor in the atmosphere [Prieto-Blanco et al. 2006]. These four imagers are assembled in a 2x2 array module. A stereo framework of the array module gives us the final camera architecture shown in Figure 1. This architecture is coupled with the highest density near-IR pattern at 940*nm* to enable active stereo. All 8 imagers are set with fixed focus lenses and have standard auto-exposure algorithms when HDR is turned off.

3.1.2 Synchronization. The applications we considered for this version of the vision architecture are for static pick and place / assembly based robotic applications with HDR enabled. This did not require any necessity for microsecond level synchronization since the objects that the robot interacted with are static (i.e. not moving). The frame rates are relatively low with HDR (3 to 5 fps). The synchronization requirements were sufficient to within a few 10s of milliseconds. We implemented a USB-based call-and-response mechanism to estimate round-trip time between the camera and the Vision Computer. The Vision Computer then sets and periodically resets the time for each camera, ensuring synchronization. Images are captured at the top of each second and half-second.

In cases where tracking and pose estimation on moving objects must be supported, it is expected that faster synchronization between the camera, within a few tens of microseconds, will be necessary. It is anticipated that this can be achieved by transitioning to an Ethernet-based interface with support for the IEEE 1588 PTP protocol for sensor synchronization.

3.2 Image Processing

The images captured from the hardware (Figure 1(a)) are processed (Figure 2(b)) and converted into data usable for 3D reconstruction.

1. **Multi-Exposure Capture:** Images are captured from each of the 8 sensors at 3 separate exposures.
2. **ISP Processing:** We perform black level correction, vignetting correction, and demosaicing. We do not perform auto white-balance nor color correction.
3. **HDR Calculation:** HDR image is computed for each sensor.
4. **Polarization Calculation:** Using our *Plentopic Stereo*, a low-resolution depth map is quickly calculated on a single unit. This depth map is used to align the polarized RGB cameras and calculate the angle and degree of linear polarization through a least squares fit [Kadambi et al. 2015].
5. **Final Output:** RGB, AOLP, DOLP, and IR images are output as shown in Figure 2(b).

3.3 Multi-Unit IR-Dot Auto-Calibration

Stereo-based systems encounter accuracy challenges when measuring distances far exceeding than their baseline, a limitation evident in our single unit with a 10 cm baseline and 7.9 mm focal length, leading to notable depth estimation errors at longer ranges. To address this, we introduce a novel auto-calibration method that effectively creates a large 1 m virtual baseline by incorporating a second unit. This approach significantly improves triangulation accuracy. By estimating the transforms between units using infrared dot correspondences, we eliminate the need for calibration targets.

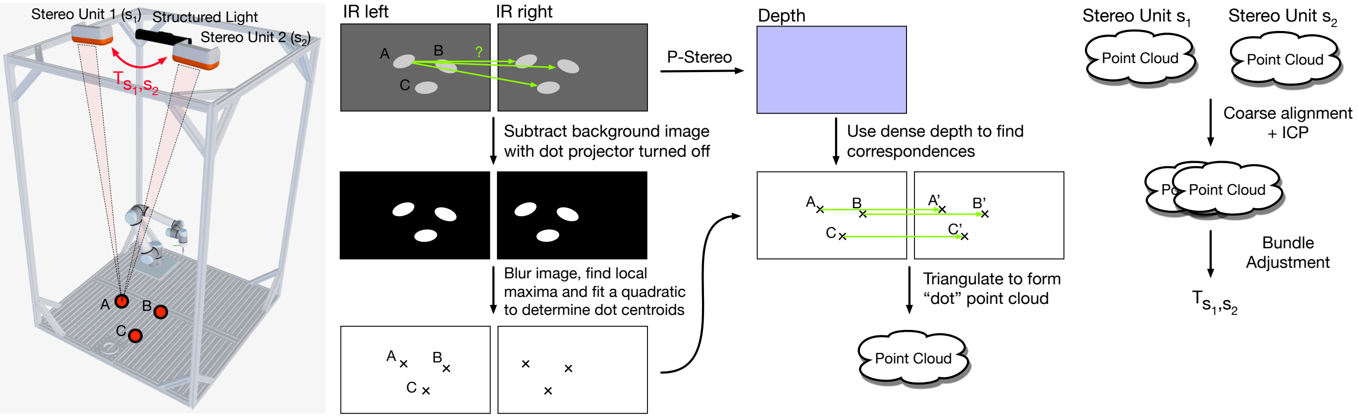
Our markerless solution not only enables automated in-situ calibration and drift detection during production but also enhances the overall system accuracy without disrupting the workflow. Traditional marker-based methods often prove unreliable in demanding production environments, as maintaining a calibration board's integrity over extended periods can be both challenging and costly. Below, we provide a high-level summary of our auto-calibration. A detailed step-by-step explanation is available in the supplement.

A pseudo-random dot pattern is projected at 940nm onto a surface visible to multiple infrared (IR) cameras. Each camera captures two images: one with the pattern and one without. By comparing these images, the precise 2D locations of the dots are determined.

These dot locations are then used to establish correspondences between different cameras, enabling the estimation of the cameras' positions and orientations relative to each other. This initial estimate is refined through a bundle-adjustment [Agarwal et al. 2023; Triggs et al. 2000] optimization process that minimizes the error with respect to the positional uniformity of dots across all camera views.

Optionally, the accuracy of the dot correspondences can be further enhanced using optical flow techniques [Lucas and Kanade 1981]. Most importantly, to ensure the ongoing calibration accuracy in a production environment, the above process can be periodically repeated and the results compared to previous calibrations to determine any potential drift without the need for manual intervention.

Our system's accuracy can therefore be tailored to any desired working distance by selecting a virtual baseline between cameras. This decoupling of accuracy from the physical hardware with customized baselines is a core aspect of our scalable and modular design.



(a) Workcell with two stereo units.

(b) Autocalibration steps that are performed for each stereo unit independently.

(c) Point cloud registration.

Fig. 3. **Our IR-dot based automatic calibration pipeline allows us to register multiple units to each other without requiring multiple images or calibration targets.** Details are available in Section 3.3 and the supplement.

This eliminates the need for multiple hardware versions with different physical baselines to support different working distances and accuracy requirements. Note that this approach is used only for estimating inter-unit transforms; intra-unit parameters are factory calibrated using standard techniques.

4 PLENTOPIC 3D RECONSTRUCTION

In this section, we introduce our novel *Deep Plenoptic Stereo (P-Stereo)* architecture and training pipeline that allows us to produce accurate and robust 3D reconstructions.

4.1 Base Architecture

Our approach (Figure 4) extends architectures such as *CREStereo* [Li et al. 2022] and *RAFT* [Lipson et al. 2021] by enabling two key features:

> *Multi-Modal Fusion:* This architecture allows the system to incorporate information from an arbitrary number of stereo units with RGB, IR, and polarized inputs trained fully on synthetic data - leading to superior completeness and generalization across lighting, material, and geometry.
>
> *Large Baseline Correspondence:* This architecture allows for accurate correspondence in disparity ranges of more than 1000px enabling stereo vision to show improved 3D accuracy at longer working distances.

The architecture (Figure 4) can be described by the following steps:

(a) **Input Stereo Pairs:** Our architecture supports RGB, Polarized, and IR stereo pairs. Since RGB and Polar images are already pixel-level aligned (see Section 3.2), we pass them as a single stereo pair, whereas IR is a seperate camera, therefore we pass it in as a seperate stereo pair. We select the largest available baseline as the reference pair because this leads to the highest resolution cost volume.

(b) **Feature Extraction:** Initially, all image pairs from all modalities are run independently through a shared feature extraction backbone that outputs features at (1/16, 1/8, 1/4) resolution. For aligned images (e.g. AOLP, DOLP, RGB), the feature maps are summed before computing the cost volume.

(c) **Cost Volume** Similar to CREStereo, cost volumes are computed at every image resolution (1/16, 1/8, 1/4) for each input stereo pair. However we compute it just once directly and at full resolution as done in RAFT rather than iteratively and locally like in CREStereo.

(d) **Cost Volume Fusion:** This is a crucial step in the process of multi-modal stereo fusion. After calculating all the cost volumes, each one is warped to align with the chosen reference cost volume. This is done by first determining the 3D location of every element within the reference camera's cost volume. Using the calibration data of the other stereo pairs, we find the corresponding 3D locations in the remaining cost volumes. With this mapping established, bilinear interpolation is used to warp all cost volumes onto the reference pair, where they are summed together. This enables the optimization to leverage information from every modality and baseline. For example, the large baseline cost volume tends to be noisy due to significant viewpoint shifts. This noise is mitigated by fusing the small baseline cost volume while retaining the increased accuracy of the estimated depth from the larger baseline cost volume.

(e) **Iterative Optimization:** Accurate disparity estimation in large-baseline scenarios (1000+ px disparity ranges) presents a significant challenge due to the wide search space required for finding correct correspondences. The multi-scale iterative disparity estimation employed by CREStereo encounters difficulties with local minima due to its reliance on an initial disparity and subsequent search within a confined window. Even with multiple scales, the largest window encompasses a mere 256 pixels in the original image. To overcome this limitation, we propose a novel approach that incorporates a disparity scale parameter, *d*, into the GRU's cost volume query. This parameter dynamically adjusts the sampling intervals within the cost volume, enabling a coarse-to-fine search

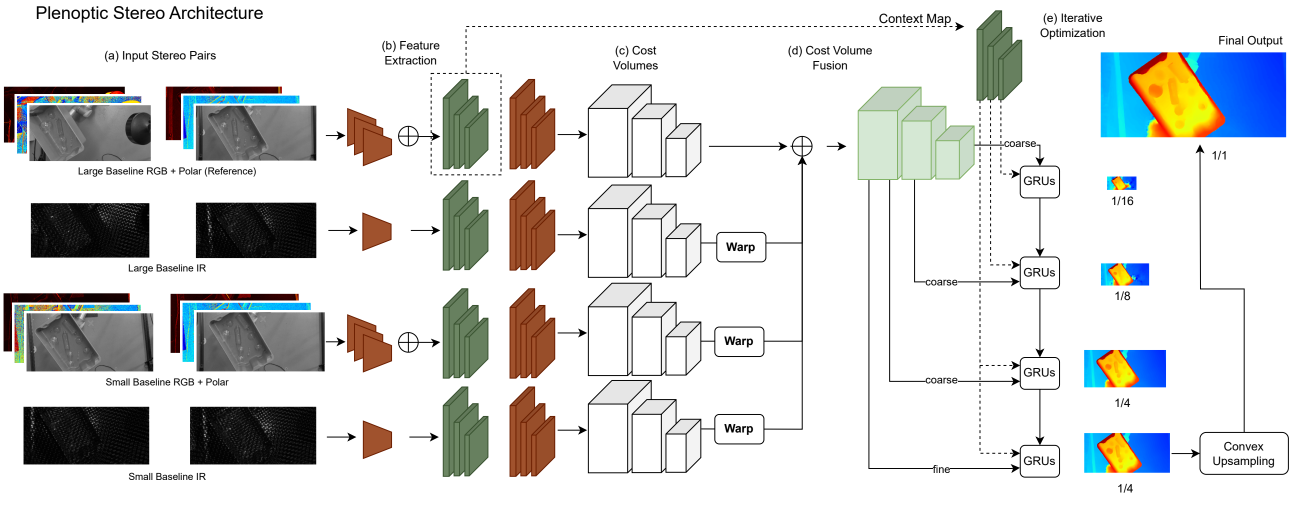


Fig. 4. **The proposed Plenoptic Stereo architecture can leverage information from multiple calibrated stereo pairs with different modalities and produce high quality reconstruction.** The detailed explanation is available in Section 4.1

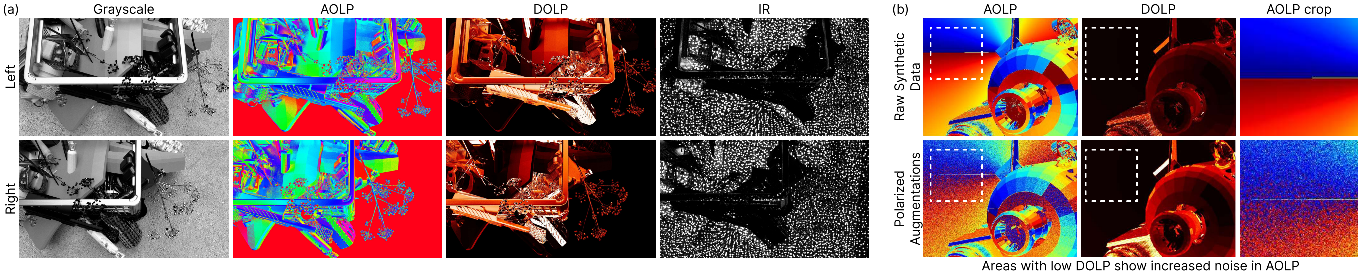


Fig. 5. **Our synthetic data generation pipeline.** (a) - Example scene generated for training our multi-baseline multimodal stereo system. The left and right images are rendered from left cameras of Unit 1 and Unit 2, respectively, simulating the large baseline. (b) - Our polarized data augmentations realistically model the correlation between low DOLP values and increased AOLP noise observed in real-world data, effectively closing the sim2real gap.

strategy. By initially exploring the cost volume at a coarse interval and then using a fine interval at the last scale, we effectively navigate the expansive disparity space. Without this coarse-to-fine traversal, the GRUs are unable to handle large disparity ranges.

At inference time, this architecture can run any combination of baselines and modalities and produce accurate pointclouds.

4.2 Synthetic Data Generation

To train our multi-baseline multimodal stereo system, we generated over 4,000 synthetic scenes. Each scene contains a random assortment of objects with diverse textures and materials, including shiny and transparent surfaces. These objects are positioned randomly within the scene, either through gravity-based simulation or direct placement in 3D space. Both cluttered and uncluttered scenarios are included to enhance the model's robustness. Two simulated stereo units with large, random baseline are incorporated into each scene. For every unit, we render grayscale and IR stereo images (with IR dots), along with a Stokes vector, which is then converted into AOLP (Angle of Linear Polarization) and DOLP (Degree of Linear Polarization) data. All rendered materials are physics-based, enabling full Fresnel equation calculations at the surface level. Mitsuba 3 [Jakob et al. 2022] was employed as the renderer for this synthetic data generation process. Figure 5 showcases examples of the generated grayscale, IR, and polarization-based synthetic data.

To bridge the gap between simulated and real-world polarization data, we introduce physically accurate data augmentations (Figure 5) for the AOLP and DOLP. Specifically, we randomly rotate AOLP to simulate viewpoint changes and randomly scale DOLP, instead of adjusting brightness or contrast. We also add noise to AOLP based on weak DOLP signal areas and to DOLP based on its values, replacing random noise addition. These polar augmentations enhance the realism of the polarization data. Step-by-step augmentations and additional details on synthetic data are available in the supplement.

To the best of our knowledge, this is the first work to utilize a fully synthetic, physics-based rendering pipeline for polarization-based multimodal stereo reconstruction tasks, as well as polarization-related tasks in general. To enhance the generalization of our models to real-world scenarios, we introduce random polarization states to some light sources within the scenes. This reflects the fact that real-world environments rarely adhere to the "unpolarized world" assumption, thus improving the quality and realism of our dataset. Training exclusively on synthetic data ensures that our model generalizes effectively to new scenes and avoids overfitting to specific train-test scenarios. The strong generalization capabilities of our model are demonstrated in Figure 9, where our model performs equally well on both our testing data and the testing data from DPS-Net [Tian et al. 2023], whereas the DPS-Net approach fails outside its training distribution.

5 EVALUATION

5.1 Evaluation Setup

Our evaluation setup is illustrated in Figure 3(a). Two cameras are mounted 2.5 meters above the floor to simulate the clearance for large industrial robots, with a baseline of approximately 1 meter. A UR5e robotic arm [Universal Robots 2024], positioned at the base of the cell, holds the evaluation scenes, while a structured light sensor [Photoneo 2024] mounted at the top serves as a baseline sensor. An additional visible light projector provides gray-code patterns for capturing the ground truth point cloud [Scharstein and Szeliski 2002, 2003].

We collected data from 115 scenes across various categories, including large car parts, a metal bin, cluttered metal parts in a plastic bin, cluttered textured objects, objects with reflective dark and light surfaces, thin objects, and transparent objects. For each scene, multiple robot poses were captured under three lighting conditions: room light, strong spotlight, and low light. Prior to data collection, each scene was spray-painted, and a structured light point cloud was captured to serve as ground truth data for Collision Avoidance Metric estimation, in conjunction with the robot poses.

5.2 Auto-Calibration

To assess our single-shot, marker-free calibration pipeline, we compare it to standard checkerboard target-based methods [Zhang 2000] using OpenCV [Bradski 2000] with both 1 and 15 unique checkerboard captures in Table 1. We use a large 800 mm x 600 mm checkerboard and during the 15 captures, we move it to maximize coverage of the working volume. We quantify the impact of each method by computing 2D reprojection error (average difference between detected corner locations and their projections) and 3D triangulation error (difference between estimated 3D positions of triangulated checkerboard corners and known positions of checkerboards). These metrics are calculated on a test set of 20 checkerboard images moved throughout the working area and provide a comprehensive evaluation of the accuracy of the estimated inter-unit extrinsic transforms and their impact on 3D reconstruction performance.

Table 1 demonstrates that incorporating a second stereo unit for IR Autocalibration significantly enhances 3D triangulation accuracy (1.703 mm to 0.488 mm). While multiple checkerboard patterns remain the most accurate method (0.172 mm), they require multiple captures and a calibrated board, presenting challenges in factory settings. Notably, these metrics are calculated from triangulating checkerboard corners, which allows for sub-pixel correspondence. In real-world environments, disparity errors can be a few pixels, or 4-8x higher than checkerboard triangulation errors, further underscoring the need for a second stereo unit. A key advantage of this approach compared with checkerboard-based methods is the ability to detect calibration drift without requiring human intervention.

Calibration Method	**# Captures**	**# Units**	**Calib. Target**	**2D Error (px)**	**3D Error (mm)**
N/A	0	1	N/A	N/A	1.703
OpenCV	1	2	Checker	0.887	0.253
OpenCV	15	2	Checker	**0.523**	**0.172**
IR Auto-calib.	1	2	N/A	0.772	0.488

Table 1. Comparison of average 2D re-projection and 3D triangulation errors for a single unit vs. checkerboard methods vs. our IR-dot auto-calibration.

5.3 3D Point Cloud Evaluation

We now discuss the ability of our work to create high-quality point clouds across different lighting, geometries, and materials. First we demonstrate that each modality, when added to our *P-Stereo* architecture leads to improved robustness as defined by our collision avoidance metrics FNR (%) and FPR (%) at a 10mm collision threshold (other thresholds in supplement). We then evaluate the full system compared to a high-resolution structured light system and demonstrate significant improvements on challenging categories and comparable results on less challenging categories. Finally, we show the generalization of our *P-Stereo* model by showing zero-shot generalization to the non-robotics RPS [Tian et al. 2023] polar dataset, outperforming the state-of-the-art DPS-Net system.

5.3.1 Multi-Modal Evaluation. In Table 2, we evaluate the impact of different modalities on FNR and FPR across different lighting conditions and categories. In this section, we use RGB stereo as a baseline because it represents the standard deep RGB stereo methods commonly used in the research community.

RGB + IR: By adding the IR dots as shown in Figure 2, we see a sharp improvement in the overall FPR & FNR, especially in *Cluttered Bin.* In Figure 6, we show the value of the IR stereo in separating foreground from background in textureless regions of an example *Cluttered Bin* scene. This feature is enabled by our *P-Stereo* architecture's ability to fuse stereo pairs through cost volume warping.

Multi-Baseline: The addition of the second unit allows the system to detect thin objects just mm off the floor as the effective baseline increases from 10cm to 90cm,. In Table 2 (*Thin Objects*), we see an FNR improvement from 7.7% → 4.0%. This improvement is reflected visually in Figure 7. Across the board, this capability shows the largest improvement in FNR (12.1% → 9.7%), and in the *Large Part* category, it brings the FNR from 3.33% → 1.0%, very close to that of structured light at 0.9. This improvement is only possible from our multi-baseline GRU optimization and our auto calibration.

Polarization: Polarization provides the system with additional information when RGB & IR fail. This occurs either in challenging lighting conditions or challenging materials. In Table 2, we see polarization reduce the FNR from 9.3% → 8.1% and 10.7% → 9.6% in the spotlight & dark scenarios, showing the value of the additional information. Furthermore in Table 2, we see that across the board, there is an improvement in FNR, especially with transparent objects where we go from 10.9% → 9.3% (see Figure 8). Even after this

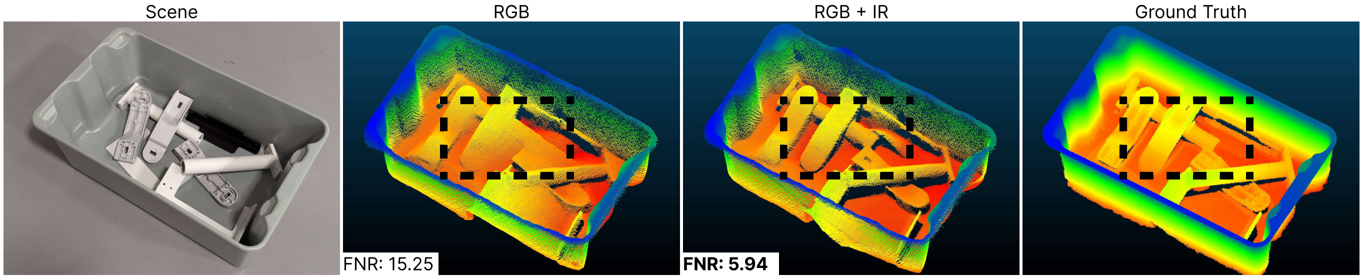


Fig. 6. **Fusing data from a separate IR stereo pair significantly improves point cloud reconstruction, particularly in areas with overlapping objects**. Our novel cost volume warping operation enables seamless integration of IR data, resulting in increased detail and accuracy compared to RGB alone. The RGB blurs the bottom of the bin and the parts while IR shows clean separation.

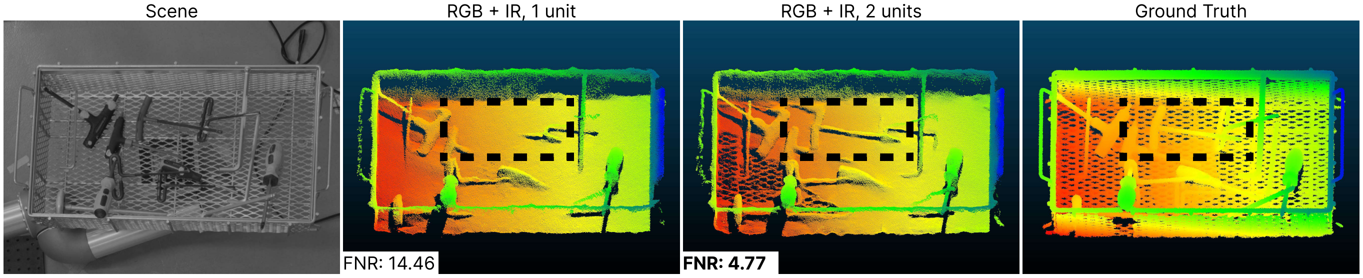


Fig. 7. **Adding multiple units improves the ability to reconstruct thin objects.** This figure shows the improvement obtained in the 3D reconstruction when using two camera units. By adding more units, we can leverage multiple views and multiple baselines giving the model the accuracy needed to reconstruct thin objects such as the screwdriver depicted in the scene.

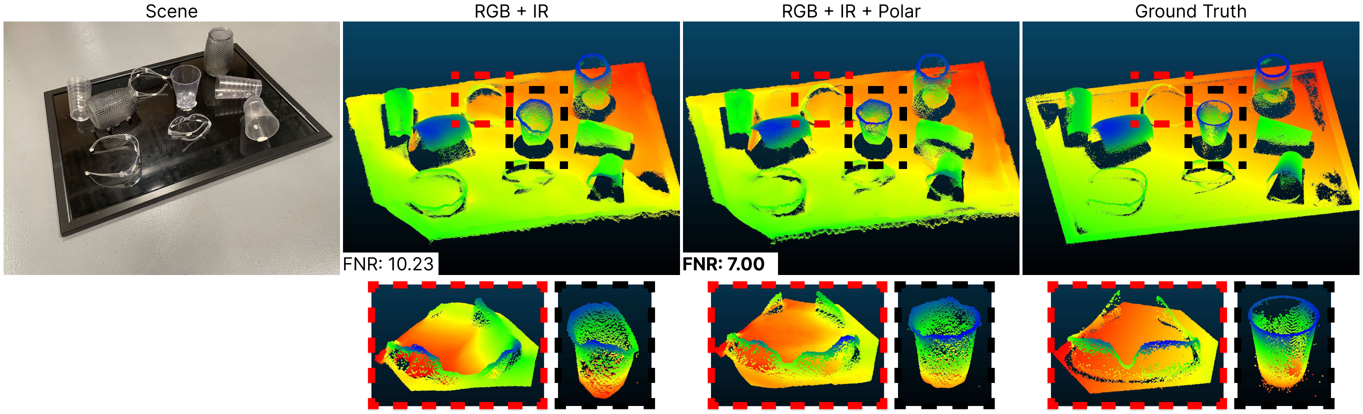


Fig. 8. **We see a substantial improvement in point cloud reconstruction achieved by incorporating polarization signals into our network.** In this challenging scene with low light, transparent surfaces, and complex objects (vases, cups), our architecture recovers details otherwise lost. The addition of polarization significantly enhances reconstruction, particularly of the safety glasses and plastic cup, showcasing the network's capability to handle real-world complexity.

improvement however, transparent objects are still challenging and a potential for future research. As a consequence of detecting more objects, polarization increases FPR. Reducing FNR is more important as preventing a collision trumps improving efficiency.

Generalization: Our system successfully generalizes polarization to new domains without training on real-world data. In Table 3 and Figure 9, we compare DPS-Net [Tian et al. 2023], the previous state-of-the-art for polarized stereo to our system on both their RPS dataset and our robotics dataset. Our system shows superior zero-shot generalization when both are trained only on synthetic data for both the RPS testing data (epe 3.4 ⟶ 2.0) and our testing data (FNR 27.52 ⟶ 14.21).

With real training data from the RPS dataset, DPSNet achieves improved epe (3.4 ⟶ 0.6), however this is clearly overfitting, as

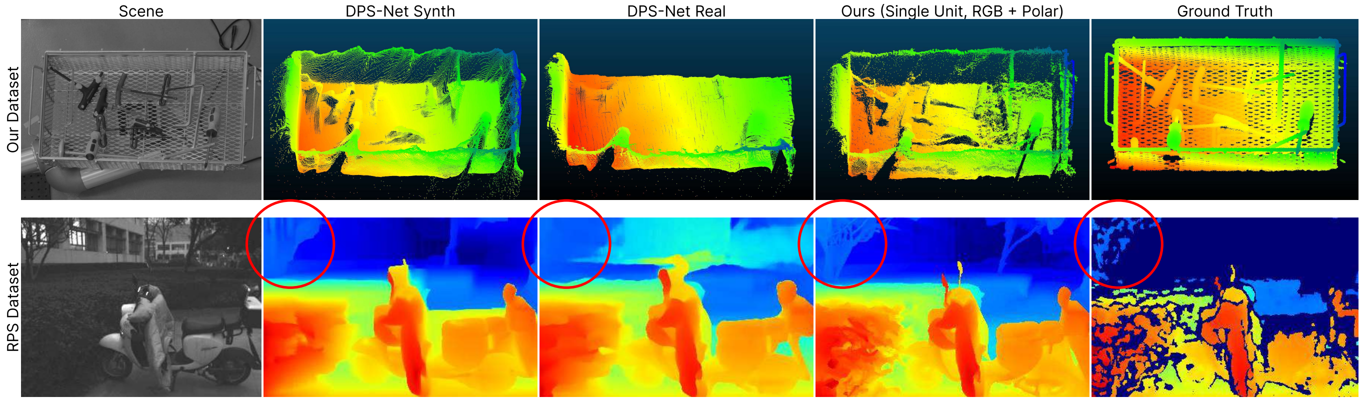


Fig. 9. **Our synthetically trained network generalizes well to non-robotics polarization data.** DPS-Net, despite real data training, suffers from hallucination, notably missing bin walls in our data and blurring trees in the RPS dataset. Note the limited quality of RPS ground truth.

Vision System			**Select Categories**								**Lighting Conditions**						**Full Dataset**	
			Large Part		Cluttered Bin		Thin Objects		Transparent		Roomlight		Spotlight		Dark			
Method	Modality	Units	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR
	Structured Light	1	**0.1**	0.9	0.4	13.5	1.0	8.7	0.4	18.8	**0.8**	22.3	**0.7**	23.3	**0.9**	19.3	**0.8**	21.6
Ours	RGB	1	2.5	3.1	2.1	13.3	1.6	8.6	2.2	13.6	2.1	13.2	2.1	13.3	3.2	17.4	2.5	14.6
Ours	RGB + IR	1	2.4	3.3	↓ 0.5	↓ 7.0	1.2	7.7	**1.4**	↓ 10.1	1.4	↓ 11.4	1.2	↓ 11.4	1.7	↓ 13.5	1.4	↓ 12.1
Ours	RGB + IR	2	↓ 1.5	↓ 1.0	**0.2**	↓ 5.1	**0.9**	↓ 4.0	3.2	10.9	0.9	↓ 9.2	1.1	↓ 9.3	1.3	↓ 10.7	1.1	↓ 9.7
Ours (Final)	RGB + IR + Polar	2	2.8	↓ **0.6**	0.4	↓ **4.6**	1.7	↓ **3.1**	3.7	↓ **9.3**	1.8	↓ **8.6**	2.1	↓ **8.1**	2.2	↓ **9.6**	2.0	↓ **8.8**

Table 2. **Evaluation on our dataset across modalities, lighting conditions, and categories.** Our final multi-modal approach shows a sharp reduction in FNR when compared to the structured light system. All of our approaches show highest error in *dark* scenes while the structured light system shows the best results in *dark* scenes. The largest improvement comes from adding a second unit. The improvement from polarization is most substantial in challenging lighting conditions such as spotlight & roomlight, and for transparent objects. We use a ↓ to denote a significant improvement FNR or FPR.

System	**Training Dataset**	**Training Domain**	**Testing Dataset**	
			RPS (epe)	**Ours (FNR)**
DPS-Net	IPS	Synth	3.6px	27.52
	IPS + RPS	Synth + Real	**0.6px**	55.77
Ours	**Ours**	Synth	2.0px	**14.21**

Table 3. **Comparison with DPS-Net polar stereo [Tian et al. 2023].** Trained purely on our synthetic data, our system demonstrates zero-shot generalization to the non-robotics RPS polar dataset [Tian et al. 2023], outperforming DPS-Net trained only on synthetic data. While adding real RPS data improves performance on the RPS testing data, it leads to overfitting and very poor results on our testing data (see Figure 9).

the generalization to our real test set is signficantly worse than the synthetic only model (FNR 27.52 ⟶ 55.77). This is due to the difficulty of producing accurate depth labels and the limited availability of real polarized data. For example, the RPS ground truth data was collected using a low-resolution sensor, resulting in blurry and sparse depth information (see Figure 9).

In contrast, our *P-Stereo* model combined with our realistic synthetic polarized training data consistently outperforms DPS-Net in generalization across both our dataset and the RPS dataset. This demonstrates our approach's robust polarized stereo vision.

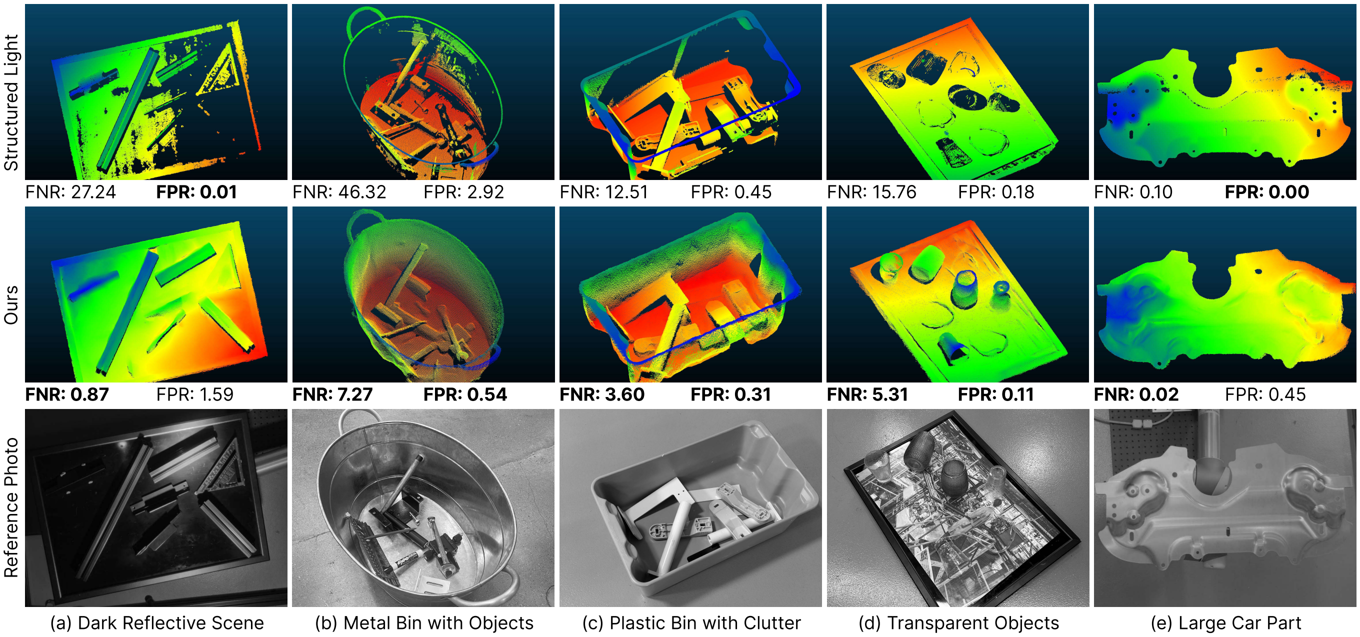
Ablations We investigated the impact of data augmentations and architectural choices through ablation studies. The results in Table 4 demonstrate that incorporating polarized data augmentations substantially improves sim2real transfer, reducing FNR by 2x. Notably, the effectiveness of these augmentations is amplified when polarization information is included in both the cost volume and the context map (the features used to initialize the GRU; see [Lipson et al. 2021], Figure 4). Excluding polarization from the context map leads to a degradation in performance.

5.3.2 *Structured Light Evaluation.*

The main motivation of our work was to answer whether one can build a stereo vision system that can produce competitive point clouds to high resolution structured light systems in the context of industrial robotics, specifically collision avoidance. For this we start by discussing the *ideal case* for structured light, then discuss key improvements in challenging cases such as occlusion, ambient light, and adversarial materials.

Ideal Case: The ideal case for structured light is a large, flat, diffuse object that allows for high-quality 3D scanning with minimal occlusions. This is shown in Figure 10 (e), and Table 2(Large Part). In this scenario, traditional stereo struggles to achieve good performance, however with the addition of RGB, IR dots, and multiple units, we show improved FNR compared to structured light (0.6% vs 0.9%) with worse, but still reasonable FPR.

Occlusions Structured light systems struggle to reconstruct occluded areas, like bin walls here, when they are visible only to either the camera or the projector, but not both. Our multi-modal stereo approach overcomes this limitation by requiring the information to



(a) Dark Reflective Scene (b) Metal Bin with Objects (c) Plastic Bin with Clutter (d) Transparent Objects (e) Large Car Part

Fig. 10. **Our multi-modal, multi-view vision system excels in challenging scenarios, providing more complete point clouds of occluded regions and materials that confound traditional structured light methods.** In (a) and (d), our system successfully reconstructs dark, shiny, and transparent objects, while (b) and (c) demonstrate superior reconstruction of bin walls. Even in ideal structured light conditions (e), our system remains competitive, producing a slightly less sharp but still accurate reconstruction. All scenes are around 2 meters away from the system. All FNR / FPR numbers shown above are for the specific scene, while Table 2 shows averages across categories.

be present in just one of the multiple cost volumes being combined. This results in a significant improvement in the reconstruction of these occluded areas, which is clearly seen in our results. In Table 2 (Cluttered Bin), we reduce FNR dramatically (13.5% → 4.6%) and in Figure 10 (b, c) we show superior reconstruction.

Challenging Materials Challenging materials prevent the projector pattern from being accurately registered in the captured images, leading to large false negative regions. From Figure 10 (a, d) and Table 2 (Transparent), we see that in these cases, the additional modalities and viewpoints available in our system allow us to achieve superior reconstructions (FNR 18.8% → 9.3%). In some cases, when transparent objects are not properly reconstructed, it can lead to potential collisions in operation.

Ambient Lighting Ambient lighting negatively impacts structured light systems, causing gaps in the resulting 3D pointcloud data, as demonstrated in Table 2(Dark vs Spotlight). Our system, however, is more robust to varying lighting conditions thanks to HDR capture (Figure 2) and its ability to function without relying solely on pattern projection.

Overall, our system shows a large improvement in FNRs (21.6% → 8.8%), with competitive FPRs when compared to structured light.

6 CONCLUSIONS

In this paper, we have presented a novel plenoptic multi-camera, multi-modal, multi-baseline stereo vision system designed to address the challenges of collision avoidance in demanding industrial robotic automation applications. By combining stereo vision with polarization, infrared sensing, and a unique self-calibration mechanism, our system demonstrates robustness to challenging lighting, materials, and geometries that often hinder traditional structured light methods. We acknowledge that some challenges remain for future work. For example, addressing adversarial materials (transparent, anisotropic) in varying lighting conditions poses challenges for robust collision avoidance.

Polar in Cost Volume	Polar in Context Map	Polar Augs	RGB Augs	FNR	FPR
✓				9.9	4.5
✓			✓	11.9	4.9
✓		✓		9.9	**2.0**
✓	✓	✓		**8.8**	**2.0**

Table 4. **Data augmentation choices are crucial for effective polarized sim2real transfer.** Incorporating physically accurate polarization augmentations into our synthetic data generation pipeline leads to a substantial 2x improvement in the FPR metric. We see further improvement from integrating polarization in the context map.

While our work primarily focuses on collision avoidance, our system's high-resolution multi-modal data opens promising avenues for future research in other robotic vision tasks, such as pose estimation, grasp estimation, reinforcement learning, and more. This potential for broader impact underscores the value of our approach as a step towards a *camera-only* system for industrial robotics.

REFERENCES

Sameer Agarwal, Keir Mierle, and The Ceres Solver Team. 2023. *Ceres Solver*. https://github.com/ceres-solver/ceres-solver

Nicolas Aspert, Diego Santa-Cruz, and Touradj Ebrahimi. 2002. Mesh: Measuring errors between surfaces using the hausdorff distance. In *Proceedings. IEEE international conference on multimedia and expo*, Vol. 1. IEEE, 705–708.

Gary A Atkinson and Edwin R Hancock. 2006. Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing* 15, 6 (2006), 1653–1664.

C Bamji, J Godbaz, M Oh, S Mehta, A Payne, S Ortiz, and S Nagaraja. 2022. A review of indirect time-of-flight technologies. *IEEE Transactions on Electron Devices* 69, 6 (2022), 2779–2793.

Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. 1977. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings: Image Understanding Workshop*. Science Applications, Inc, 21–27.

Gary Bradski. 2000. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* 25, 11 (2000), 120–123.

Jia-Ren Chang and Yong-Sheng Chen. 2018. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5410–5418.

Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. 2017. Polarimetric multi-view stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1558–1567.

Marc E Deetjen and David Lentink. 2018. Automated calibration of multi-camera-projector structured light systems for volumetric high-speed 3D surface reconstructions. *Optics express* 26, 25 (2018), 33278–33304.

M-P Dubuisson and Anil K Jain. 1994. A modified Hausdorff distance for object matching. In *Proceedings of 12th international conference on pattern recognition*, Vol. 1. IEEE, 566–568.

Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. 2019. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4384–4393.

Haoqiang Fan, Hao Su, and Leonidas J Guibas. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 605–613.

O.D. Faugeras, Q.-T. Luong, and S.J. Maybank. 1992. Camera self-calibration: Theory and experiments. In *Computer Vision — ECCV'92 (Lecture Notes in Computer Science)*, G. Sandini (Ed.), Vol. 588. Springer, Berlin, Heidelberg, 321–334. https://doi.org/10.1007/3-540-55426-2_37

Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino. 2021. Polarimetric normal stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 682–690.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2495–2504.

Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. 2019. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3273–3282.

Ciarán Hogan, Ganesh Sistu, and Ciarán Eising. 2023. Self-Supervised Online Camera Calibration for Automated Driving and Parking Applications. *arXiv preprint arXiv:2308.08495* (2023).

R Horaud, M Hansard, G Evangelidis, and C Ménier. 2016. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine Vision and Applications* 27, 7 (2016), 1005–1020.

Tianyu Huang, Haoang Li, Kejing He, Congying Sui, Bin Li, and Yun-Hui Liu. 2023. Learning accurate 3d shape based on stereo polarimetric imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17287–17296.

Kei Ikemura, Yiming Huang, Felix Heide, Zhaoxiang Zhang, Qifeng Chen, and Chenyang Lei. 2024. Robust Depth Enhancement via Polarization Prompt Fusion Tuning. *arXiv preprint arXiv:2404.04318* (2024).

Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. 2022. *Mitsuba 3 renderer*. https://mitsuba-renderer.org.

Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. 2017. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE international conference on computer vision*. 2307–2315.

Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. 2015. Polarized 3d: High-quality depth sensing with polarization cues. In *Proceedings of the IEEE International Conference on Computer Vision*. 3370–3378.

Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. 2020. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8602–8611.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. 2017. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*. 66–75.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. 2022. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. arXiv:cs.CV/2203.11483

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

Bruce D Lucas and Takeo Kanade. 1981. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, Vol. 2. 674–679.

Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. 2019. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10452–10461.

Thomas Marko and Wilfried Kubinger. 2018. Automatic Stereo Camera Calibration in Real-World Environments Without Defined Calibration Objects. In *Proceedings of the 29th DAAAM International Symposium*. 0102–0108.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4040–4048.

Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. 2022. Glass segmentation using intensity and spectral polarization cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12622–12631.

M Okutomi and T Kanade. 1991. A multiple-baseline stereo. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Photoneo. 2024. *Photoneo PhoXi 3D Scanner XL*.

Ana Prieto-Blanco, Peter RJ North, Nigel Fox, and Michael J Barnsley. 2006. Satellite estimation of surface/atmosphere parameters: a sensitivity study. In *Global Developments in Environmental Earth Orbservation from Space. Proceedings of the 25th EARSeL Symposium. Porto: Millpress Science Publishers*. 137–44.

Daniel Scharstein and Richard Szeliski. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision* 47 (2002), 7–42.

D Scharstein and R Szeliski. 2003. High-accuracy stereo depth maps using structured light. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1.

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Faranak Shamsafar and Andreas Zell. 2021. TriStereoNet: A Trinocular Framework for Multi-baseline Disparity Estimation. *arXiv preprint arXiv:2111.12502* (2021).

Zhelun Shen, Yuchao Dai, and Zhibo Rao. 2021. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13906–13915.

X Su and Q Zhang. 2010. Dynamic 3-d shape measurement method: a review. *Optics and Lasers in Engineering* 48, (2) (2010), 191–204.

Vage Taamazyan, Alberto Dall'olio, and Agastya Kalra. 2024. Collision Avoidance Metric for 3D Camera Evaluation. arXiv:cs.CV/2405.09755

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 402–419.

Chaoran Tian, Weihong Pan, Zimo Wang, Mao Mao, Guofeng Zhang, Hujun Bao, Ping Tan, and Zhaopeng Cui. 2023. DPS-Net: Deep polarimetric stereo depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3569–3579.

Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. 2000. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*. Springer, 298–372.

Universal Robots. 2024. *Universal Robots UR5e*.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. 2022. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8606–8615.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. 2021. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14194–14203.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. 2023. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21919–21928.

Haofei Xu and Juyong Zhang. 2020. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1959–1968.

Qingshan Xu and Wenbing Tao. 2020. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on*

Artificial Intelligence, Vol. 34. 12508–12515.

Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. 2020. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4877–4886.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*. 767–783.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5525–5534.

Jure Zbontar and Yann LeCun. 2015. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1592–1599.

Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. 2019. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 185–194.

Zhengyou Zhang. 2000. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence* 22, 11 (2000), 1330–1334.

Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. 2023. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1327–1336.

Ren Zhonghe, Fengzhou Fang, Ning Yan, and You Wu. 2022. State of the art in defect detection based on machine vision. *International Journal of Precision Engineering and Manufacturing-Green Technology* 9, 2 (2022), 661–691.

Dizhong Zhu and William AP Smith. 2019. Depth from a polarisation+ RGB stereo pair. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7586–7595.